Statistical Policy
Working Paper 23

Seminar on New Directions in Statistical Methodology

Part 1 of 3

Federal Committee on Statistical Methodology

Statistical Policy Office
Office of Information and Regulatory Affairs
Office of Management and Budget

June 1995

MEMBERS OF THE FEDERAL COMMITTEE ON
STATISTICAL METHODOLOGY

(June 1995)

Maria E. Gonzalez, Chair
Office of Management and Budget

M. Denice McCormick Myers, Secretary
National Agricultural Statistics Service

Susan W. Ahmed
National Center for
Education Statistics

Yvonne M. Bishop
Energy Information
Administration

Cynthia Z.F. Clark
National Agricultural
Statistics Service

Steven Cohen
Administration for Health
Policy Research

Lawrence H. Cox
Environmental Protection
Agency

Zahava D. Doering
Smithsonian Institution

Daniel Kasprzyk
National Center for
Education Statistics

Nancy Kirkendall
Energy Information
Administration

Daniel Melnick
Substance Abuse and Mental
Health Services Administration

Robert P. Parker
Bureau of Economic Analysis

Charles P. Pautler, Jr.
Bureau of the Census

David A. Pierce
Federal Reseerve Board

Thomas J. Plewes
Bureau of Labor Statistics

Wesley L. Schaible
Bureau of Labor Statistics

Rolf R. Schmitt
Bureau of Transportation
Statistics

Monroe G. Sirken
National Center for
Health Statistics

Robert D. Tortora
Bureau of the Census

Alan R. Tupek
National Science Foundation

Denton R. Vaughan
Social Security
Administration

Robert Warren
Immigration and Naturalization
Service

G. David Williamson
Centers for Disease Control
and Prevention

PREFACE

The Federal Committee on Statistical Methodology was organized by the Office of Management and Budget (OMB) in 1975 to investigate issues of data quality affecting Federal statistics. Members of the committee, selected by OMB on the basis of their individual expertise and interest in statistical methods, serve in a personal capacity rather than as agency representatives. The committee conducts its work through subcommittees that are organized to study particular issues and prepare working papers presenting their findings. The subcommittees are open by invitation to Federal employees who wish to participate. This is the 23rd Statistical Policy Working Paper published under the auspices of the committee since its founding.

On May 25-26, 1994, the Council of Professional Associations on Federal Statistics (COPAFS) hosted a "Seminar on New Directions in Statistical Methodology." Developed to capitalize on work undertaken during the past fifteen years by the Federal Committee on Statistical Methodology and its subcommittees, the seminar focused on a variety of topics that have been explored thus far in the Statistical Policy Working Paper series and on work on statistical standards undertaken by the Statistical Policy Office at OMB. The subjects covered at the seminar included:

- Economic Classification Revisions
- Disclosure Limitation Methodology
- Customer Surveys
- Advances in Data Editing
- Time Series Revision Policies
- Incentives in Surveys
- Computer Assisted Survey Information Collection
- Longitudinal Surveys
- Cognitive Testing and Self-Administered Questionnaires
- Statistical Uses of Administrative Records
- Small Area Estimation
- Nonresponse in Surveys

Each of these topics was presented in a two-hour session that featured formal papers and discussion, followed by informal dialogue among all speakers and attendees.

Statistical Policy Working Paper 23, published in three parts, presents the proceedings of the "Seminar on New Directions in Statistical Methodology." In addition to providing the papers and formal discussions from each of the twelve sessions, the working paper includes Graham Kalton's keynote address, "Improving the Quality of Federal Statistics," and comments by Norman M. Bradburn, Robert M. Groves, and Katherine K. Wallman at the closing session, "Toward an Agenda for the Future."

We are indebted to all of our colleagues who assisted in organizing the seminar, and to the many individuals who not only presented papers but also prepared these materials for publication.

Table of Contents

Wednesday, May 25, 1994

Part 1

KEYNOTE ADDRESS

IMPROVING THE QUALITY OF FEDERAL STATISTICS 1
Graham Kalton, Westat, Inc.

Session 1 - ECONOMIC CLASSIFICATION REVISIONS

ECONOMIC CLASSIFICATIONS IN THE NEW NORTH AMERICAN INDUSTRY CLASSIFICATION SYSTEM (NAICS). 17
Jack E. Triplett, Bureau of Economic Analysis

REVISING THE UNITED STATES STANDARD OCCUPATIONAL CLASSIFICATION (SOC) SYSTEM. 29
Thomas J. Plewes, U.S. Bureau of Labor Statistics

COMMENTS ON THE REVISIONS OF THE STANDARD INDUSTRIAL AND OCCUPATIONAL CLASSIFICATIONS. 34
Joel Popkin, Joel Popkin & Company

COMMENTS ON ECONOMIC CLASSIFICATION REVISIONS. 37
Joe Mattey, Federal Reserve Board

Session 2 - DISCLOSURE LIMITATION METHODOLOGY

RESTRICTED DATA VERSUS RESTRICTED ACCESS: A PERSPECTIVE FROM "PRIVATE LIVES AND PUBLIC POLICIES". 43
George T. Duncan, Carnegie Mellon University

STATISTICAL DISCLOSURE LIMITATION METHODOLOGY. 57
Nancy J. Kirkendall, Energy Information Administration

DISCUSSION OF PRESENTATIONS ON STATISTICAL DISCLOSURE METHODOLOGY. 68
Stephen E. Fienberg, Carnegie Mellon University

SELECTED ASPECTS OF RESTRICTED DATA. 81
Tore Dalenius, Brown University

Session 3 - CUSTOMER SURVEYS

QUALITY MANAGEMENT FOR CUSTOMER SATISFACTION SURVEYS. 87
Richard M. Devens, Jr., U.S. Bureau of Labor Statistics

COMPARABILITY IN CUSTOMER SATISFACTION SURVEYS: PRODUCTS, SERVICES AND GOVERNMENT AGENCIES. 99
Michael D. Johnson, University of Michigan

CUSTOMER SURVEYS: DISCUSSION. 121
Robert M. Groves, University of Michigan

DISCUSSION. 126
Elizabeth Martin, U.S. Bureau of the Census

Session 4 - ADVANCES IN DATA EDITING

IMPROVING OUTLIER DETECTION IN TWO ESTABLISHMENT SURVEYS. 137
Julia L. Bienias, David M. Lassman, Scott A. Scheleur, and Howard Hogan, U.S. Bureau of the Census

TIME SERIES AND CROSS SECTION EDITS WITH APPLICATIONS TO FEDERAL RESERVE DEPOSIT REPORTS 152
David A. Pierce and Laura Bauer Gillis, Federal Reserve Board

DISCUSSION. 172
Sandra A. West, U.S. Bureau of Labor Statistics

DISCUSSION. 177
Brian V. Greenberg, U.S. Bureau of the Census

Session 5 - TIME SERIES REVISION POLICIES

TIME SERIES REVISIONS: THE EFFECTS ON GROSS DOMESTIC PRODUCT. 183
Robert P. Parker and Teresa L. Weadock, U.S. Bureau of Economic Analysis

RAISING THE NATION'S UNEMPLOYMENT RATE. 209
John E. Bregger, U.S. Bureau of Labor Statistics

COMMENTS ON PARKER AND WEADOCK, TIME SERIES REVISIONS: THE EFFECTS ON GROSS DOMESTIC PRODUCT. 220
Murray F. Foss, American Enterprise Institute

Part 2

Session 6 - INCENTIVES IN SURVEYS

TIME, DOLLARS, AND DATA: SUCCEEDING WITH REMUNERATION IN HEALTH SURVEYS. 225
Trena M. Ezzati-Rice, Andrew A. White, William R. Mosher, National Center for Health Statistics, Maria Elena Sanchez, Agency for Health Care Policy and Research

THE USE OF INCENTIVES TO SURVEY "HARD-TO-REACH" RESPONDENTS: A BRIEF REVIEW OF EMPIRICAL RESEARCH AND CURRENT RESEARCH PRACTICE. 256
Richard A. Kulka, National Opinion Research Center

DISCUSSION. 290
Diane K. Willimack, National Agricultural Statistics Service

DISCUSSION. 295
W. Sherman Edwards, Westat, Inc.

Thursday, May 26

Session 7 - COMPUTER ASSISTED SURVEY INFORMATION COLLECTION

REDESIGNING A QUESTIONNAIRE FOR COMPUTER-ASSISTED DATA COLLECTION: THE CURRENT POPULATION SURVEY EXPERIENCE. . . 301
Cathryn Dippo and Anne Polivka, U.S. Bureau of Labor Statistics, Kathleen Creighton, Donna Kostanich, and Jennifer Rothgeb, U.S. Bureau of the Census

AN OVERVIEW OF THE 1992 CENSUS OF AGRICULTURE CATI SYSTEM. 311
Jeanette K. Mon, U.S. Bureau of the Census

DISCUSSION: WHAT CAN CAI LEARN FROM HCI?. 363
Mick. P. Couper, Joint Program in Survey Methodology

DISCUSSION OF TWO PAPERS ABOUT CASIC. 378
Sandra Sperry, Westat, Inc.

Session 8 - LONGITUDINAL SURVEYS

PANEL DESIGN AND ESTIMATION STRATEGIES IN THE NATIONAL MEDICAL EXPENDITURE SURVEY. 381
Steven B. Cohen, Agency for Health Care Policy and Research

METHODOLOGICAL ISSUES ENCOUNTERED IN FOLLOWING A COHORT OF EIGHTH GRADERS. 428
Steven J. Ingels, National Opinion Research Center, Jeffrey A. Owings, National Center for Education Statistics

DISCUSSION. 446
Gary M. Shapiro, Abt Associates

DISCUSSION. 449
James M. Lepkowski, University of Michigan

Part 3

Session 9 - COGNITIVE TESTING AND SELF-ADMINISTERED QUESTIONNAIRES

LITERACY LIMITATIONS AND SOLUTIONS FOR SELF-ADMINISTERED QUESTIONNAIRES. 453
Judith T. Lessler and James O'Reilly, Battelle Memorial Institute

THE LANGUAGE OF SELF-ADMINISTERED QUESTIONNAIRES AS SEEN THROUGH THE EYES OF RESPONDENTS. 470
Cleo R. Jenkins, U.S. Bureau of the Census and Don Dillman, U.S. Bureau of the Census and Washington State University

DISCUSSION. 517
Jared B. Jobe, National Center for Health Statistics

DISCUSSION. 524
Roger Tourangeau, National Opinion Research Center

Session 10 - STATISTICAL USES OF ADMINISTRATIVE RECORDS

IMPROVING DATA QUALITY THROUGH INCREASED DATA SHARING. . .531
Edward A. Trott, Bureau of Economic Analysis

HEALTH REFORM INFORMATION SYSTEMS: GREAT EXPECTATIONS, UNCERTAIN PROSPECTS. 540
Edward L. Hunter, National Center for Health Statistics

DISCUSSION. 549
Miron L. Straf, Committee on National Statistics, National Academy of Sciences

Session 11 - SMALL AREA ESTIMATION

SMALL AREA ESTIMATION FOR THE NATIONAL HEALTH INTERVIEW SURVEY USING HIERARCHICAL MODELS. 555
Donald Malec, National Center for Health Statistics, J. Sedransk, State University of New York, Albany

THE ROLE OF DESIGN BASED VARIANCES AND COVARIANCES IN SMALL DOMAIN ESTIMATION. 569
Robert E. Fay, U.S. Bureau of the Census

DISCUSSION. 595
Phillip S. Kott, National Agricultural Statistics Service

DISCUSSION. 599
David A. Marker, Westat, Inc.

Session 12 - NONRESPONSE IN SURVEYS

EXPLORING NONRESPONSE IN U.S. FEDERAL SURVEYS. 603
Maria Gonzalez, Office of Management and Budget, Dan Kasprzyk, National Center for Education Statistics, Fritz Scheuren, Internal Revenue Service

MODEL-BASED REWEIGHTING FOR NONRESPONSE ADJUSTMENT. . . . 625
David A. Binder, Sylvie Michaud and Claude Poirier, Statistics Canada

DISCUSSION. 646
J. Michael Brick, Westat, Inc.

DISCUSSION. 649
Joseph L. Schafer, Pennsylvania State University

TOWARD AN AGENDA FOR THE FUTURE

Norman M. Bradburn. 657
National Opinion Research Center

Robert M. Groves. 660
The Joint Program in Survey Methodology

Katherine Wallman. 665
Office of Management and Budget

May 25, 1995

Roger was so special. If you thought you had a good idea, it was smart to first pass it by Roger. He would think about your idea for a while, and then in his ever nice way tell you if he thought it was indeed a good idea. Then he would think about it some more, and come up with an even better idea.

This symposium is dedicated to Roger Herriot, a unique and wonderful person. We will all miss him.

Edward J. Spar
Executive Director
COPAFS

Part 1

KEYNOTE ADDRESS

IMPROVING THE QUALITY OF FEDERAL STATISTICS

IMPROVING THE QUALITY OF FEDERAL STATISTICS

Graham Kalton
Westat, Inc.

1. Introduction

This is the second seminar hosted by the Council of Professional Associations on Federal Statistics (COPAFS) related to the Statistical Policy Working Paper Series of the Federal Committee on Statistical Methodology (FCSM). In deciding on a theme for this talk, I reviewed the working paper for the previous seminar, entitled *Seminar on Quality of Federal Data* (U.S. Office of Management and Budget, 1991). As my title indicates, I have chosen a similar theme. Within the broad subject of "Improving the Quality of Federal Statistics", my main focus is on approaches to improving quality across the federal statistical system as a whole rather than in specific programs. I shall also pay particular attention to the role that the FCSM can best play in achieving the objective of quality improvement.

At the outset I should make it clear that my choice of topic is not to be taken to imply any criticism of the current quality of federal statistics. Indeed, I have a high regard for the federal statistical programs and the professionalism of the federal statistical workforce. Rather, my comments are made from the perspective that, however good the current situation, improvements are always possible.

Equally, my discussion of the role of the FCSM should be interpreted in the same light. Like Bob Groves, who gave the keynote address at the previous seminar, I believe that the FCSM and its working paper series perform a valuable service towards the goal of improving the quality of federal statistics. The working papers also make an important contribution to the general survey and statistical literature. For example, like Bob, I have employed the working papers for teaching purposes. Last fall, I used the excellent recent working paper on indirect estimators (U.S. Office of Management and Budget, 1993a) in my sampling course in the Joint Program in Survey Methodology. The suggestions made below for the Statistical Working Paper Series are offered in the spirit of seeking improvements in a series that has established itself as an extremely useful product.

2. Factors Involved in Quality Improvement

In discussing improvements in the quality of federal statistics, I am interpreting the term "quality" to include not only the production of accurate estimates, but also the production of relevant and timely statistics in a cost-efficient manner, and

the ready accessibility of statistics and data to users. The components of quality thus include:

- *Accuracy*. The estimates produced should have low bias and variance for the parameters being estimated.

- *Relevance*. The parameters being estimated should be the ones that are relevant for users. To ensure relevance, statisticians need to maintain regular and close contacts with users.

- *Timeliness*. The estimates should be up-to-date. The more out-of-date the estimates, the less relevant they are. Equally, other statistical products, such as public use tapes, need to be made available to users in a timely manner.

- *Accessibility*. Statistical products need to be accessible to users through such mechanisms as publications, public use tapes, CD-ROMs, and diskettes. Although restrictions on access may be needed to protect the confidentiality of survey respondents (using such techniques as cell suppression in tables and top-coding and suppression of variables in public use tapes), these restrictions need to be implemented in ways that minimize their consequences for the utility of the data. Good documentation of statistical products is needed to make them readily accessible.

- *Cost-efficiency*. The procedures used to collect and analyze statistical data should be ones that are most cost efficient, taking into account the need to satisfy the other components of quality.

Improving quality thus encompasses: using improved methods of data collection and processing to produce more accurate data; refining definitions of statistical concepts to better meet policy needs; instituting procedures to enable statistics to be produced more rapidly; developing ways to improve access to statistical data; and introducing more cost-efficient methods of data collection, processing and analysis. The broad definition of quality that I am using seems the appropriate one, and one that is consistent with the breadth of the Statistical Policy Working Paper Series. For instance Working Paper 11 deals with industry coding schemes, Working Paper 17 deals with survey coverage, Working Paper 19 deals with computer assisted survey information collection, Working Paper 21 deals with indirect estimates, and Working Paper 22 - an update of Working Paper 2 - deals with statistical disclosure limitation methodology (U.S. Office of Management and Budget, 1984, 1990a, 1990b, 1993a, 1994, and U.S. Department of Commerce, 1978, respectively). Gonzalez's (1994) description of the activities of the FCSM contains a useful review of the broad scope of the Statistical Policy Working Paper Series.

Improvements in the quality of federal statistics can come about in several ways. One is by improving the flow of communication between the user and producer of statistics. As noted above, relevance is a key component of quality and relevance requires the producer to fully understand user needs. Equally, users need to appreciate the inherent limitations in the capabilities of the statistical system that produces the statistics they employ. Regular contacts between users and producers are essential to keep producers aware of changing user priorities and of changes in the structure of society that need to be taken into account in producing relevant statistics in a changing world. The importance of user/producer communication is mainly one that needs to be addressed at the individual statistical program level. Since I am focussing on improving quality in the statistical system at large, I will not expand on this important issue here.

A second means for improving the quality of federal statistics is by increasing the use of what my colleague David Morganstein terms Current Best Methods (CBMs). He prefers this terminology to the more usual Standard Operating Procedures (SOPs) because it conveys the principle that the best methods are evolving over time. In this respect, improvements are produced by increasing the awareness of CBMs among those involved in producing federal statistics and by facilitating their use.

A third means of quality improvement is through research on statistical methodology. Such research can serve both to identify problems with existing methods and to suggest improved methods. The results of this research then feed into the evolving CBMs.

In the last issue of the *Survey Statistician*, Morganstein (1993) describes the application of the process of continuous quality improvement in the survey statistics group at Westat. He identifies three primary elements of the program: employee development; documentation and the use of CBMs; and improved technology. These same elements seem equally applicable for improvements in federal statistics.

The challenges of achieving quality improvement across the federal statistics system are, of course, far greater than they are within a single survey statistics department. Indeed, the challenges are much greater in the decentralized U.S. statistical system than they are in centralized systems such as those at Statistics Canada, and the Australian and Netherlands Bureaus of Statistics. This is not the place to discuss the relative merits of centralized and decentralized statistical systems. All that needs to be noted here is that the problems of communication of current best methods in a decentralized system are severe. The large number of U.S. government agencies that are conducting statistical work provides an indicator of the communication challenge. According to OMB's Annual Report to Congress, in 1994 there are around 80 different agencies that receive direct funding

for major statistical programs. The eleven principal federal statistical agencies receive less that two-fifths of the total major statistical program funding.

As discussed later, I see a prime function of the FCSM as being one of encouraging the use of CBMs. The committee can serve this function by developing working papers detailing CBMs and disseminating them to those engaged in federal statistical activities. In the decentralized environment of the U.S. statistical system, dissemination is a major challenge, a point to which I shall return.

3. Contributors to Quality Improvement

In considering the range of contributors to quality improvement in federal statistics, it is useful to distinguish between employee development on the one hand and the development of CBMs and methodological research on the other. A highly-skilled work force is critical for the production of high quality statistics. The essential components of a highly-skilled statistical workforce are, first, the recruitment of well-trained statisticians, with training appropriate to their job requirements and, second, continuing education over the course of their careers to keep them up-to-date with the many advances that are being made.

In response to a shortage in the numbers of trained survey statisticians, at the end of the 1980's members of the federal statistical system pressed for the establishment of a "Center for Survey Methods" to provide instruction and research training at a Washington-based university. I should like to note here the important contributions of Hermann Habermann and the agency heads of the Bureau of Economic Analysis, the Bureau of the Census, the Economic Research Service of USDA, and the Bureau of Labor Statistics, who worked on the proposal for the Center as part of a 1990 legislative initiative under the leadership of Michael Boskin, then chair of the President's Council of Economic Advisers. These efforts were successful, leading to the establishment of the Joint Program in Survey Methodology (JPSM) at the University of Maryland, a joint program of the University of Maryland at College Park, the University of Michigan and Westat, Inc. The program is now underway, with the first year of the MS program in Survey Methodology completed, and with a proposal for a Ph.D. program in progress. I am pleased to be a faculty member of the Joint Program, a program which I believe holds great promise for improving the quality of federal statistics through training. I should like to recognize the strong support given to the program by Kathy Wallman and all in the OMB Statistical Policy Branch.

In addition to the JPSM at the University of Maryland, there have also been expansions to the programs of other universities in the Washington area that are of direct interest to those working

with federal statistics, such as the recently introduced Federal Statistics Certificate and Masters Degree programs at George Mason University and Masters Degree in Statistics for Policy Analysis at The American University. The numerous courses offered by the universities in the Washington area are generally made available to both degree seeking and non-degree seeking students, and they are often given at times chosen to fit in with full-time work schedules. They provide excellent opportunities for federal statisticians to obtain graduate training in a wide range of subjects. Many federal statisticians have, for instance, learned about such topics as variance estimation with complex samples and recent developments in survey methodology at evening courses at George Washington University and the USDA Graduate School. Through such offerings the universities in the Washington area make important contributions to the training of federal statisticians.

With the major advances taking place in all aspects of federal statistics, there is the need for continual updating and upgrading of the skills of the statistical workforce. One has only to reflect briefly on the advances in methods for questionnaire design, computerized data collection, variance estimation, handling nonresponse, small area estimation, and data disclosure limitation that have occurred within the past ten to fifteen years to realize that a substantial investment in continuing education is essential for keeping federal statisticians up-to-date on best current methods.

As well as through university courses taken on an ad-hoc basis, continuing education can be achieved through short courses, seminars and conferences. Perhaps in response to the recent methodological developments, there has been an impressive expansion of such offerings in recent years. An extensive array of continuing education opportunities is now available for federal statisticians. Moreover, many federal statisticians avail themselves of these opportunities, which I take to be a positive indication both of the desire of federal survey statisticians to upgrade and update their skills and of the strong support of the leadership of the statistical programs for continuing education.

For those in the Washington area, the Washington Statistical Society (WSS) has for many years been making major contributions to continuing education through its short courses and its extensive seminar series. The JPSM also now offers regular short courses at both introductory and advanced levels. In addition, continuing education short courses are regularly offered at the annual meetings of the American Statistical Association (ASA), the American Association for Public Opinion Research (AAPOR) and at the biennial sessions of the International Statistical Institute (ISI).

Conferences serve both as a form of continuing education and as a way to stimulate research work. The scientific programs of the ASA and AAPOR annual meetings and the ISI biennial sessions are

rich in contributions relevant to improving federal statistics. In addition, several series of more specialist conferences have been established in recent years, including the Bureau of the Census's Annual Research Conference, the Conferences on Health Survey Research Methods, the international conferences on survey methods, the Statistics Canada symposia, and now the COPAFS seminars.

Federal statistical programs have much to gain from the attendance and active participation of their staffs in such conferences. I would particularly single out the value of international conferences. We need to keep in touch with the statistical developments that are occurring throughout the world. Sometimes statisticians in other countries can benefit from research conducted in the U.S. and sometimes U.S. statisticians can benefit from research conducted elsewhere. In addition, the increasing interest in the production of comparable economic, social, and environmental statistics across countries points to the need for greater contact between, and collaboration of, government statisticians in different countries.

In the area of employee development, I should finally like to note the significant contribution made by the impressive program of research meetings run throughout the year by the Washington Statistical Society. In addition to three short courses, the WSS held as many as 57 meetings during the 1993-94 year, covering a wide range of topics of interest to federal statisticians. Many of the WSS presentations are made by federal statisticians, and the meetings are generally well attended.

Turning to quality improvement through promoting current best methods, there are again many contributors. Much of the work in this area is initiated and conducted by individual statistical programs, but there are important inputs from other bodies. For instance, many programs have advisory committees that provide expert advice on both substance and methods. In addition COPAFS provides advice, as does the Committee on National Statistics (CNSTAT) of the National Academy of Sciences. Panels of CNSTAT have conducted in-depth studies of specific programs and also of many aspects of federal statistical methodology. The latter include studies of missing data (Madow, Nisselson and Olkin, 1983; Madow, Olkin and Rubin, 1983; Madow and Olkin, 1983), surveying subjective phenomena (Turner and Martin, 1984), microsimulation modeling (Citro and Hanushek, 1991), and confidentiality and accessibility of government data (Duncan, Jabine and de Wolf, 1993).

Quality improvements also come about by improving current best methods. Improvements in CBMs arise out of methodological research, and once again there are many contributors. Much important methodological research is conducted by the federal statistical agencies. Much is also conducted in universities, in survey organizations and in other settings, in the U.S. and

elsewhere, and in the government statistical agencies in other countries. The challenge to maintaining CBMs as "current" and "best" is that of keeping abreast of the large volume of methodological research, and applying its results effectively in current practice. Networks of contacts are needed within the federal statistical system and between federal statisticians and those conducting methodological research elsewhere to keep CBMs up-to-date.

Given the many contributors to quality, what should be the role of the FCSM? Clearly, the FCSM is not well-positioned to contribute directly to quality improvements in programs on an individual basis. Rather, its prime role should be to provide a means for transfer of innovations across programs and for coordination of methodologies where called for.

In her contribution at the closing session of the 1991 symposium, Margaret Martin (1990, p.462) succinctly summarized four functions that the FCSM might perform:

"(1) exchange knowledge, techniques or experience among committee members to enhance the quality of the member agencies' own operations;

(2) provide "state of the art" reports to encourage best practice among a broader group;

(3) recommend areas for improvement and needed directions for research; and

(4) obtain consensus on such issues as - defining problems and the priorities among them, developing or changing classifications or other concepts, and setting statistical standards."

I think that these four functions provide a good agenda for the Committee.

4. Activities of the FCSM

This section considers each of the functions Margaret Martin lists for the FCSM in turn.

4.1 Exchange of Knowledge, Techniques and Experience Among Subcommittee Members.

In forming a subcommittee to produce a working paper on a particular subject, the FCSM draws upon the expertise on that subject that is available throughout the federal statistical workforce. Membership of a subcommittee then potentially provides the opportunity for an individual to engage in discussions with

others working in the subject, often with different perspectives and experiences. Such a dialogue has the important benefit that the exchange of knowledge, techniques and experience can lead to improvements in the methods applied in the statistical programs from which the subcommittee draws its members. This benefit is particularly important when the subject is one that involves only one or two persons in any one program, so that there is little opportunity for within-program dialogue on it.

In practice, Margaret Martin's comments suggest that subcommittees often have little time for such productive dialogue. Rather, much subcommittee work is report drafting and reviewing, activities that are performed in evenings and at weekends. If this is the case, it is unfortunate: a valuable function of the subcommittee is being lost.

I appreciate that this may not be a good time to ask for additional resources. Nevertheless, the leadership of the statistical programs should recognize the significant rewards that can accrue to their programs and to the federal statistical system more generally from subcommittee activities, and they should seek to ensure that adequate resources are provided to enable the subcommittees to carry out their work as effectively as possible. In part, this means allowing subcommittee members sufficient time to fully perform their roles and in part it means providing each subcommittee with appropriate support staff to work efficiently. The latter could include administrative staff to organize meetings and maintain schedules; editorial staff to help with the production of the working paper; and junior statisticians to serve as research assistants to help with literature reviews and bibliographies if needed (an activity that can provide a valuable learning experience for the junior statisticians).

4.2 Production of Working Papers.

The FCSM has stimulated the production of 22 papers in the Statistical Policy Working Paper Series to date. As I have already remarked, these papers make an important contribution to improving the quality of federal statistics, and to the survey statistics literature more generally. I am therefore somewhat concerned that there appear to be possible signs of some slackening in the pace of working paper production in the last few years. I hope that this is not a true loss of momentum, because I believe there is much more that could usefully be done.

Most of the working papers that I have seen contain valuable descriptions of the applications of the methodology under study across a range of statistical programs. They thus provide a useful review of the state of current practice and help to foster cross-fertilization among programs. To the extent that the programs reviewed are employing current best methods, they document what those methods are. My concern is that the focus may be too narrow.

I think that the working papers would sometimes be improved by a broader perspective on current best methods, examining both the methods used in the government statistical programs of other countries and those used outside government. I acknowledge that some subcommittees attempt to go in this direction, but I think that a more systematic approach along these lines would enhance the value of the working paper series.

In his address, Bob Groves suggested the possibility of including members from outside the federal statistical system in the subcommittees. I note that this suggestion was adopted for the disclosure limitation working paper, with Tom Jabine serving as a member of the subcommittee. I think that this suggestion merits more widespread application. Other possibilities include inviting outside experts to make presentations at subcommittee meetings, arranging small workshops for subcommittee members and outside experts to discuss the issues, and inviting outside experts to review draft working papers. Individuals from outside the federal statistical system may even be asked to draft one or more chapters for a working paper.

If a working paper is to be viewed as a document of best current methods, then it should do more than simply review current practices. It should include recommendations for what are the best current methods, recognizing the variety of different circumstances in which the methods may be applied. To reach agreement on such recommendations may often be difficult, and clearly requires much discussion among the subcommittee members. Lack of sufficient discussion time may well be the reason that the recommendations in the working papers are often not as developed as would be desirable.

Another consequence of viewing the working papers as a means of promoting current best methods is that they should be seen as evolving documents that need to be updated as improved methods are developed. An example here is the latest working paper on statistical disclosure that updates a 1978 working paper to take account of the major advances that have occurred in the intervening period. Progress in recent years in other areas suggests the need to update other working papers, for instance those on developing questionnaires, telephone data collection, the use of microcomputers, and even the fairly recent working paper on computer assisted survey information collection.

The working papers should be prepared to meet the needs of their primary readership, which I take to be those working on federal statistical programs. They should aim to address the questions to which these readers would like answers. In this regard, I should like to recall the wide range of statistical programs that I have outlined earlier, many of which are relatively small. It is in fact the smaller programs that are likely to benefit most from the working papers, since they necessarily lack

the range of expertise that is often internally available in the large statistical agencies. The needs of the smaller programs should be borne in mind in preparing a working paper. To ensure their needs are met, it would be advisable to secure adequate representation of the smaller programs on each subcommittee.

The working papers are valuable only to the extent that they are read. Many able statisticians devote a great deal of effort to the production of each working paper. However, it is my impression that less effort goes into the distribution of the product. The working papers need to reach the desks of those for whom they were written, and mechanisms are needed to ensure that this is achieved. It is also valuable to have a widespread distribution outside the federal statistical system both in the U.S. and abroad. The papers have a great deal to offer to those involved in statistical work in many organizations and countries, and their exposure to a wide spectrum of readers opens up greater likelihood of future improvements. To achieve greater circulation of the working papers it may be useful to publicize them more extensively in appropriate newsletters and journals in the U.S. and abroad and to build up an international network of contacts to aid in the distribution. The recent article describing the working paper series by Gonzalez (1994) is helpful in this regard.

I am not in a position to suggest the best distribution system for the working papers in the federal statistical programs. One possibility might be to identify an individual in each program to serve as a liaison to the FCSM, and send copies of the working papers to that individual. The individual might also be asked to provide suggestions of topics for the FCSM to study. Another possibility is to organize a well-publicized workshop on the topic of a working paper as it is released. Since the working papers have become substantial documents, the workshop could provide a useful primer for those interested in its contents. To some extent, this COPAFS seminar serves such a role, but it is more general in nature spanning the contents of many working papers. The WSS may be able to play an important role in helping to achieve a wide dissemination of the working papers to statisticians in the Washington area.

4.3 Areas for Improvement and Directions for Research.

A number of the working papers indicate areas for improvement and for research, but these issues are not as fully developed as might be desirable. I attribute this situation to the limited discussion time available to the subcommittees. To identify needed improvements goes beyond describing current methods to pinpointing their weaknesses and coming up with ways by which the weaknesses may be addressed. Developing an effective research agenda requires a great deal of deliberation by the subcommittee.

Subcommittees of the FCSM are appointed on the basis of their technical expertise in the given subject area. As such, they are well-positioned to determine incremental research agendas for the given subject. They are, however, less suited to making proposals for major restructuring. In the last seminar, Fritz Scheuren (1993) and Steve Fienberg (1993) talked about the possibility of paradigm shifts in federal statistics. It would be useful to consider setting up federal committees of a different type, composed of individuals with wide experience and broad vision, to examine the possibilities of major changes in the ways federal statistics are produced. Users of statistical data have an important role to play in such committees. As an example, the possibility of continuous measurement in place of the Census long form, which is currently under discussion at the Bureau of the Census, raises a number of possibilities for substantial changes in other data collection efforts. Such committees may be separate from the FCSM, but they should maintain close contacts with it.

4.4 Developing Consensus Across Statistical Programs.

Margaret Martin notes that the objective of obtaining consensus on definitional, conceptual and classification issues has not been well met by the activities of the FCSM. Such consensus building requires lengthy discussions, and shortage of discussion time may again be the root of the problem. Also, different programs will have vested interests in preserving their own definitions, and that will make the attainment of consensus difficult. With a decentralized statistical system, the risk of definitional differences occurring when several programs overlap in their subject matters is high. Consensus building on definitions and methods across programs holds promise of significant advances in fields that cut across different agencies (e.g., aging, children, disability).

5. Topics for Future Working Papers

In concluding, I shall take the opportunity to put forward some specific suggestions for future working papers. Before giving them, I should however like to make two general points. First, I think that the FCSM should have a mechanism for generating suggestions from the federal statistical community at large. At an earlier point, I suggested that liaison persons be appointed in each program. If that suggestion were adopted, one role of those appointed could be to seek suggestions from their colleagues and to forward them to the FCSM. Another possibility is for the FCSM to convene meetings from time to time, perhaps in conjunction with the WSS, to discuss possible subjects for working papers.

My second point concerns the form of the working papers. With the needs of the statisticians working in the smaller statistical programs in mind, I suggest that the FCSM could usefully commission

some of the working papers to be prepared in a manual-style format, reviewing the given methodology in a relatively nontechnical and applied way, and giving practical advice on the implementation of the methodology (e.g., the availability of software). Manuals of this type could be extremely helpful to those inexperienced in the use of the methodology. They need not be lengthy documents; indeed the shorter the document, the more useful it might be. Working Paper 9 on *Contracting for Surveys* (U.S. Office of Management and Budget, 1983) is along these lines. Other illustrations are provided by the manuals on sampling errors (Butcher and Elliot, 1986) and on weighting for nonresponse (Elliot, 1991) produced at the U.K. Office of Population Censuses and Surveys.

In addition to updating some of the existing working papers as discussed above, my specific suggestions for new working papers, undoubtedly blinkered by my own interests, are:

- *Quality profiles*. The error profile for the CPS (Brooks and Bailar, 1978), which was the third report in the working paper series, was an important advance in treating total survey error. Since then the SIPP Quality Profile (Jabine, King and Petroni, 1990) and the Schools and Staffing Survey Quality Profile (Jabine, 1994) have appeared, and other quality profiles are being developed. A subcommittee might usefully develop a blueprint of what such quality profiles should contain, and the methods that may be employed to produce the requisite data, based on the experience that has been gained to date.

- *Economic statistics*. At the previous seminar, Bob Groves commented that there is a distinct bias in the working paper series towards household surveys at the expense of economic statistics. I observe no change in that situation, and think that this should be remedied.

- *Customer surveys*. The requirement that government agencies conduct customer satisfaction surveys has brought many agencies with no prior experience of surveys in direct contact with survey research. In response to this situation, the U.S. Office of Management and Budget (1993b) has produced a resource manual on customer surveys and the JPSM has run a series of short courses to provide training in the conduct of such surveys. A detailed working paper on the subject would be extremely useful.

- *Evaluation research*. Large sums of money are spent by many agencies conducting experimental and quasi-experimental studies to evaluate and compare the effectiveness of various programs. A working paper on this subject could make a valuable contribution to this work.

- *Nonresponse adjustment methods.* Considerable advances have been made in methods of weighting adjustment for total nonresponse and imputation methods for item nonresponse since the late 1970s when the CNSTAT Panel on Incomplete Data studied the subject. Imputation methods are also being used more widely. A working paper on weighting and imputation could be particularly useful for those programs that have little prior experience in this area.

- *Variance estimation.* A working paper that examines the current methods and software for variance estimation, that considers the presentation of sampling errors in survey reports, and that deals with the use of generalized variance functions could be extremely useful, especially for those working in the smaller statistical programs.

6. Concluding Remarks

In concluding, let me restate that my suggestions for the Statistical Policy Working Paper Series are made in the spirit of continual quality improvements in what is a very successful activity. My particular plea is to the leadership of the statistical programs to make sure that this work is supported in the way it deserves. The quality of federal statistics derives considerable benefit from the Working Paper Series. The success of an endeavor such as this depends on the tireless support of those behind it. In this case, the FCSM is exceedingly fortunate to have Maria Gonzalez at the helm. Without her unstinting efforts over many years, it could not have succeeded as it has.

REFERENCES

Brooks, C. and Bailar, B. (1978), *An Error Profile: Employment as Measured by the Current Population Survey*, Statistical Policy Working Paper 3, Office of Federal Statistical Policy and Standards, U.S. Department of Commerce.

Butcher, B. and Elliot, D. (1986), *A Sampling Errors Manual*. London: HMSO.

Citro, C.F. and Hanushek, E.A. (eds.) (1991), *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling. Vol. I Review and Recommendations, Vol II Technical Papers*. Washington, D.C.: National Academy Press.

Duncan, G.T., Jabine, T.B. and de Wolf, V.A. (eds.) (1993), *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. Washington, D.C.: National Academy Press.

Elliot, D. (1991), *Weighting for Non-response: A Survey Researcher's Guide*. London: Office of Population Censuses and Surveys.

Fienberg, S.E. (1991), Towards an Agenda for the Future. In U.S. Office of Management and Budget *Seminar on Quality of Federal Data*, pp. 455-461. Statistical Policy Working Paper 20. Washington, D.C.: Statistical Policy Office.

Gonzalez, M.E. (1994), Improving Data Quality Awareness in the United States Federal Statistical Agencies, *American Statistician*, 48, 12-17.

Jabine, T.B. (1994), *Quality Profile for SASS: Aspects of the Quality of Data in the Schools and Staffing Survey (SASS)*, NCES 94-340. Office of Educational Research and Improvement. Washington, D.C.: U.S. Government Printing Office.

Jabine, T.B., King, K.E. and Petroni, R.J. (1990), *Survey of Income and Program Participation (SIPP): Quality Profile*. Washington, D.C.: U.S. Bureau of the Census.

Madow, W.G., Nisselson, H. and Olkin, I. (eds.) (1983), *Incomplete Data in Sample Surveys. Vol. 1: Report and Case Studies*. New York: Academic Press.

Madow, W.G., Olkin, I. and Rubin, D.B. (eds.) (1983), *Incomplete Data in Sample Surveys. Vol. 2: Theory and Bibliographies*. New York: Academic Press.

Madow, W.G., and Olkin, I. (eds.) (1983), *Incomplete Data in Sample Surveys. Vol. 3: Proceedings of the Symposium*. New York: Academic Press.

Martin, M.E. (1991), Towards an Agenda for the Future. In U.S. Office of Management and Budget *Seminar on Quality of Federal Data*, pp. 462-464. Statistical Policy Working Paper 20. Washington, D.C.: Statistical Policy Office.

Morganstein, D. (1993), The Application of Total Quality Management (TQM) Within a Survey Statistics Department. *Survey Statistician*, July 1993, 5-7.

Scheuren, F. (1991), Paradigm Shifts: Administrative Records and Census-taking. In U.S. Office of Management and Budget *Seminar on Quality of Federal Data*, pp. 53-65. Statistical Policy Working Paper 20. Washington, D.C.: Statistical Policy Office.

Turner, C.F. and Martin, E. (eds.) (1984), *Surveying Subjective Phenomena. Vols. 1 and 2*. New York: Russell Sage Foundation.

U.S. Department of Commerce (1978), *A Report on Statistical Disclosure and Disclosure-Avoidance Techniques*, Statistical Policy Working Paper 2. Washington, D.C.: Office of Federal Statistical Policy and Standards.

U.S. Office of Management and Budget (1983), *Contracting for Surveys*, Statistical Policy Working Paper 9. Washington, D.C.: Office of Information and Regulatory Affairs.

U.S. Office of Management and Budget (1984), *A Review of Industry Coding Systems*, Statistical Policy Working Paper 11. Washington, D.C.: Statistical Policy Office.

U.S. Office of Management and Budget (1990a), *Survey Coverage*, Statistical Policy Working Paper 17. Washington, D.C.: Statistical Policy Office.

U.S. Office of Management and Budget (1990b), *Computer Assisted Survey Information Collection*, Statistical Policy Working Paper 19. Washington, D.C.: Statistical Policy Office.

U.S. Office of Management and Budget (1993a), *Indirect Estimators in Federal Programs*, Statistical Policy Working Paper 21. Washington, D.C.: Statistical Policy Office.

U.S. Office of Management and Budget (1993b), *Resource Manual for Customer Surveys*. Washington, D.C.: Statistical Policy Office.

U.S. Office of Management and Budget (1994), *Report on Statistical Disclosure Limitation Methodology*, Statistical Policy Working Paper 22. Washington, D.C.: Statistical Policy Office.

Session 1

ECONOMIC CLASSIFICATION REVISIONS

Economic Classifications in the New North American Industry Classification System (NAICS)

Jack E. Triplett
Chief Economist, Bureau of Economic Analysis
Chairman, Economic Classification Policy Committee

Prepared for the Seminar on New Directions in Statistical Methodology[1]
Washington, D.C.
May 25, 1994

I. Why Has the United States Mounted this Effort?

The Economic Classification Policy Committee (ECPC), acting on behalf of the Office of Management and Budget (OMB), Statistics Canada, and Mexico's Instituto Nacional de Estadística, Geografía e Informática (INEGI), have agreed to create a new North American Industry Classification System (NAICS). Differences in the internal pressures on the three countries' respective statistical systems create corresponding differences in our motivations, and in the constraints we face. That the three countries have joined together in this effort suggests that the similarities among us may be more significant than the differences. Nevertheless, some of the pressures that have influenced classifications in the United States are not irrelevant to international as well as national discussions of classification systems.

I should first emphasize that the U.S. Standard Industrial Classification (SIC) system is quite old (it dates from the 1930's), it has been well tried in statistical agency collection programs, and it has often been revised (the last time in 1987) in an attempt to keep it up to date. Data users have had abundant experience with the industry data that this classification system provides, and they have had many years to learn of its strengths and shortcomings. Statistical agencies have had many opportunities to react to user experiences, <u>within the parameters</u> that have guided this classification system for the past 50 years.

Yet, increasing public dissatisfaction with the U.S. SIC system has been expressed through its last several revisions. Discussions of the adequacy of the SIC in the United States have occurred not merely or solely in professional exchanges between economists and statisticians, and have not occurred solely within the boundaries of narrowly technical dialogues. In the United States, focus on problems of the SIC has extended widely to the popular press and the business press.

---

[1] This paper was originally presented at a meeting of The Statistical Programme Committee (NACE), Statistical Office of the European Communities (Eurostat), Luxembourg, March 17, 1994.

Perhaps this public attention to industrial classifications is unique to the United States. To the extent that it is unique, we believe that, at least in part, it is a response to a mature industry classification system that has increasingly been viewed as inappropriate for generating the data that are needed for economic analysis.

The United States has experienced increasing private sector and government sector demands for data for the purpose of economic analysis. As one example, a major statistical initiative has in the last several years been launched [8], and a major priority within that initiative is to improve U.S. data for the measurement and analysis of productivity. We must, therefore, ask: Does the existing SIC system produce the industry data that are wanted for economic analysis, including the industry data needed for productivity measurement? These are exactly the questions that were posed for classification systems at the Williamsburg international conference on classifications [1].

II. The ECPC Examination of Classifications Starts from the Use of Economic Data

The ECPC was established by OMB in 1992 to conduct a "fresh slate" examination of economic classification systems. The ECPC began its investigation with an examination of the uses of economic data that are produced using classification systems. What is the purpose of classification systems? When economic data are produced from these systems, what are the uses of the data for which classifications are designed? The review of these issues is contained in ECPC Issues Paper No. 1, "Conceptual Issues" [3].

The ECPC's approach is a departure from the traditional approach to classifications, at least as it has developed in the United States. In the traditional view, there are many uses for data, and because there are many uses, it has been believed that the classification system must produce data for all the uses. This means, effectively, that the uses of data have relatively little ultimate role in the design of the classification system, because the requirements for different users tend to cancel each other. The traditional classification is a compromise between competing ends. The nature of these compromises is not dictated by the use of the data, nor do the designers of the classification system have a framework from which to examine the costs to the data user of the compromises incorporated into the system.

The ECPC, in common with the traditional view, also recognizes that there are multiple uses for data that are produced by classification systems. However, in contrast to the traditional view, the ECPC has concluded that the economic use of data must determine the design of the classification system; if it does so,

this will assure that the data produced by the classification system meet the intended use.

If there are alternative demands for data that must be grouped or classified, and if these alternative demands for data have different implications for the classification system, the ECPC concludes that different classification systems must be constructed. In the ECPC's view, when different requirements for classifications grow out of different data uses, this implies that different classification systems should be set up, each one designed to meet the intended need. In the traditional approach in the United States, a compromise has been sought to meet everyone's needs within a single classification system.

If tailoring classification systems to the needs for data implied a very large number of different classification systems, the ECPC's position might be impractical. We also have concluded, however, that the major analytic needs for classified data can themselves be grouped into two major classes of uses. This matter is discussed at some length in ECPC Issues Paper No. 1 [3].

Briefly, one class of uses requires that a classification system be erected on a production-oriented concept (which may also be called a supply-based concept). A second major class of uses requires that data be grouped according to a market-oriented concept (which may also be called a demand-based concept).

Thus, the major difference between the ECPC's position and the traditional one in the United States condenses to the following questions: Should there be two different classification systems, each designed to produce data for one of the two classes of uses? Or should there be only one compromise system for both classes of uses?

In international discussions of classifications, the situation is a little different from the U.S. tradition, because multiple classifications already exist. The United Nations systems include the International Standard Industrial Classification of all Economic Activities, Third Revision (ISIC), and the Central Product Classification (CPC), and Europe has Nomenclature des Activités économiques des Communautés Européennes (NACE), Classification des Produits associée aux Activités (CPA), and another system called PRODCOM.

Yet, the issues that have been debated are quite similar: What are the data uses for which different classifications are required? Should some of the classification systems (usually, the product classification system) be connected in some manner to the others (the industry classifications)? And whatever distinctions can be drawn in principle among the various systems (by reading their introductions and statements of principles, for

example), their implementations are marked by compromise. The connections between data use, classification concept, and construction of the classification system have been influenced by compromise among competing demands for data, rather than by the determination to tailor each classification system to a major economic use of grouped data.

Though seldom stated explicitly in this language, almost all of the literature discussing classification principles, whether in the United States or in international forums, can be understood as a conflict between designing a production-oriented, or supply-based, classification system and designing a market-oriented, or demand-based, classification system. That is, much discussion of classification problems concerns, at the root of the matter, the conflict between providing data for production-side economic analysis, on the one hand, and for market- or demand-oriented analysis on the other. The ECPC's issues papers, particularly ECPC Issues Paper No. 1, attempt to make more clear and explicit what has unfortunately remained implicit in much of the past discussion of economic classifications.

The distinction made in ECPC Issues Paper No. 1, however, is actually quite old. After this project was well under way, David Wharton of Statistics Canada called my attention to a very enlightening article on economic classifications published by R.H. Coats in 1925. Coats' pertinent observations for our time include the following:

> "...the basic principle in classification is that mutually exclusive concepts may not be united on an equality in the same category.... It is precisely this elementary rule that statisticians too often ignore. Called upon for statistics of aggregates from many and diverse standpoints, they attempt to meet the demand within the limits of a single classification. This leads inevitably to confusion as between principles....
>
> "...To state as was stated in the resolution originally tabled at the Geneva Conference on Labour Statisticians that a combination of principles <u>must</u> be adopted is surely to abandon the issue prematurely...." [2, emphasis in original].

Allowing for changes in economic language over the past 70 years, a subsequent passage of Coats' article can be interpreted as discussing production-oriented and market-oriented principles as alternative concepts for classification systems. Coats proposed also a third principle--distinguishing the stage of process in the hierarchy of the classification system.

III. Why Was the Production-Oriented Concept Chosen for NAICS?

A statement adopted by the ECPC, Statistics Canada, and Mexico's INEGI reads, in part:

> "The uses of industrial statistics which include measuring productivity, unit labor costs, and the capital intensity of production require that information on outputs and inputs be used together. Moreover, statistical agencies in the three countries expect to be called upon to produce information on inputs and outputs, industrial performance, productivity, unit labor costs, employment, and other statistics in order to analyze the effects of the North American Free Trade Agreement. An industry classification system erected on a production-oriented, or supply-based, conceptual framework will assure maximum usefulness of industrial statistics for these and similar purposes. Therefore, the three countries agree that the new North American Industry Classification System should conform to a production-oriented economic concept" [5].

The reasoning behind the three countries' decision may be summarized as follows. An <u>industry</u> is grouping of economic activities. Though it inevitably groups the products of the economic activities that are included in the industry definition, it is not solely a grouping of products.

Put another way, an industry groups producing units. Accordingly, an industry classification system provides a framework for collecting the variables that describe production--inputs and outputs--together on a consistent basis. The industry system thus groups data for analyses for which it is important that inputs and outputs be used together.

What uses of economic data require that inputs and outputs be used together, and be collected on the same basis? Such uses include production analyses, productivity measurement, studying input usage and input intensities, and so forth. For these uses, producing units should be grouped together by similarities in their production processes, which is exactly the production-oriented concept discussed in ECPC Issues Paper No. 1.

Thus, the North American countries have chosen the production-oriented concept as the framework for industry statistics (a) because important production analysis uses of data require groupings of producing units, and (b) because these uses are the ones that require that inputs and outputs be collected together on a comparable basis. The production-oriented concept for classification systems is discussed at greater length in ECPC Issues Paper No. 1

IV. Planning for the Alternative, Product Grouping System

A classification system that groups or aggregates products is a very different system from an industry classification system. An example of a product grouping system is the Central Product Classification (CPC) prepared by the United Nations Statistical Office.

A product grouping system[2] satisfies a different need--a different use of data--from the one served by an industry classification system. A product grouping system is used for analyses from the demand side--to define markets to study market power or to conduct marketing studies, for demand estimation, for determining the extent of substitution among commodities, and so forth. One does not want a product grouping system for studying productivity; an industry classification system produces the data for productivity analysis.

A product grouping system has the following characteristics:

(a) It should incorporate, and facilitate the analysis of, relationships among products--demand relations, substitution relations, marketing relationships, uses by consumers or by other ultimate purchasers.

(b) For demand and market analyses, the inputs to production generally do not matter for the intended data use. As a

---

[2] The term "product system" has been used to encompass at least three different ideas, only one of which is the product grouping system. The first, which might be termed the "product enumeration system," provides a list of all the products (goods and services) that exist. For example, the Harmonized System (HS) of the Customs Cooperation Council provides in principle a listing of all the products that move in international trade. A product enumeration system can also contain a grouping system, and frequently does so for organizational reasons, though the enumeration system's grouping system is not necessarily constructed to facilitate economic analysis, and is usually not suitable for analytic purposes. The listings in the product enumeration system provide building blocks for the product grouping system, which is described in the text. Finally, one often needs to list the outputs in each industry of the industry classification system. Such a listing has sometimes been given its own name; other times it has been referred to simply as the "index items" or "indexes"--for example, in the United States, the products in the "alphabetical index" of the *Standard Industrial Classification Manual, 1987* [9]. This "index system" also uses, with certain exceptions, the listings in the product enumeration system.

consequence, only the outputs matter in a product grouping system, no information on inputs need be collected.

(c) Accordingly, product groupings may cut across the producing relationships in establishments, or other producing units. Establishment outputs may be separated and assigned to different product groupings, as the principles of the product grouping system dictate.

Because it satisfies a different data use, a product grouping system is appropriately constructed on a different economic concept from the one that is used for an industry system. A product grouping system requires a market-oriented, or demand-based, economic concept. The market-oriented, or demand-based, concept for economic classifications is discussed at greater length in ECPC Issues Paper No. 1.

Moreover, there is no reason to integrate a product grouping system with an industry classification system, and there is every reason to avoid linking the two where they are in fact different. A product grouping system is intended to meet its own needs, and should meet those needs independently of the industry classification system, which is properly designed to serve a different purpose.[3]

The three North American countries have agreed that product grouping systems should be established on their own merits, as indicated in the following paragraph from their joint statement:

> "The statistical agencies of the three countries also agree that market-oriented, or demand-based, groupings of economic data are required for many purposes, including studies of market share, demands for goods and services, import competition in domestic markets, and similar studies. Each country will provide product data compiled within the framework of its respective statistical system, to meet the need for such information. Recognizing the increasing international trade in goods and services, each country will work cooperatively to help improve commodity classification systems, including the Harmonized System (HS) of the Customs Co-operation Council and the United Nations provisional Central Commodity Classification (CPC) system for services, by coordinating efforts and keeping each agency informed of proposals for changes" [5].

---

[3] Some controversy exists on this point, which I believe has arisen out of failure to distinguish between the purposes for which a product grouping system is needed and the quite different functions of a system that lists the index items in the industry classification system (see footnote 1).

V. Is a Conceptual Classification System Practical?

The approach followed in constructing NAICS involves: (a) taking the economic uses of industry data as a starting point, and (b) deriving an economic concept for industry classifications, making use of the economic theory that underlies the economic analyses that use industry data. Though it is clearly desirable that data be constructed in accordance with the implications of economic theory, and constructed so that the data meet the requirements for economic analysis, there has been some justifiable concern about the practicality of such an endeavor. Can we analyze pragmatically and empirically our present economic classifications with respect to the theoretical requirements for a conceptually based classification system? Can we design new and improved classification systems making use of the theory?

The ECPC and Statistics Canada have produced a number of studies that suggest that the task we have set ourselves is indeed practical.

A. The matrix papers

In two separate studies [6] [10], U.S. and Canadian 4-digit SIC industries were reviewed. Teams in each country asked whether individual industries embodied a production-oriented economic concept, or a market-oriented economic concept.

As explained in ECPC Report No. 1, "Economic Concepts Incorporated in the Standard Industrial Classification Industries of the United States" [6], these two reviews combined understanding of the economic concepts, as developed in ECPC Issues Paper No. 1, with informed judgments about the technologies and the markets that pertain to each detailed 4-digit SIC industry. The reviews were, first, tests to see whether the economic concepts could be implemented in a pragmatic way, using mainly the type of information about industries that has been used in the past to make decisions about the U.S. and Canadian SIC systems. These reviews use the available information to assess economic concepts.

Secondly, the two reviews provide a preliminary assessment of the concepts embodied in the U.S. and Canadian systems by past decisions. Their results are subject to revision on the basis of industry expertise.

Some present 4-digit SIC industries are already constructed along production-oriented lines, or could be, with relatively small adjustments to definitions. In the United States, the study suggests that a little under a fifth (19 percent) of manufacturing shipments come from 4-digit industries that are fully defined on the production-oriented concept, and another two-fifths (actually 45 percent) originate from industries that

could be made consistent with the concept by combining and/or subdividing existing industries (the details for these estimates are contained in [6]).

On the other hand, one could emphasize the other side of the picture. Fully two-fifths of manufacturing industries have no discernible production-oriented basis and nearly as large a proportion of manufacturing shipments arise in these industries; to this one could add the two-fifths of the manufacturing industries that require some adjustments to be fully consistent with the production-oriented concept (as noted in the previous paragraph). From those numbers, it is evident that the U.S. system as presently developed does not conform to the production-oriented concept.

The situation is about the same for the 150 services industries that were reviewed in the U.S. study. Actually, a slightly higher proportion of services industries has been defined to be consistent with the production-oriented concept, but much additional refinement of service industry definitions will be required to produce adequate industrial data.

A little under a quarter (23 percent) of U.S. manufacturing shipments come from SIC industries that have been defined on a market-oriented basis. Nearly half of those (10 percent of manufacturing shipments) are industries that meet the conditions for both production-oriented and market-oriented conceptual systems: These were designed "Ideal" industries in the review, because statistics for them are appropriate for both of the major classes of economic analysis.

Another 35 percent of shipments arise in industries that have some market-oriented basis in their definitions. Many of those are cases where production-oriented and market-oriented reasoning has been combined into a compromise industry definition that fully satisfies neither.

In the traditional view, the classification problem is to find ideal industries--those that are satisfactory for both production and market analysis--on the implicit assumption that deviations from ideal in practice can be handled as "special cases," for which case-by-case compromises can be effected. That ideal industries have been found in the United States in only 10 percent of the cases is a measure of how far the traditional view of the classification problem is from the empirical reality of actual industry structure.

B. Heterogeneity index

The ECPC has developed a new statistical approach that will assist in determining production-oriented economic groupings. This method is explained in ECPC Report No. 2, "The Heterogeneity

Index: A Quantitative Tool to Support Industrial Classification" [7], which applies the new method to 4-digit manufacturing industries in the United States.

The heterogeneity index is based on the following regularity: When producing units have the same production function and face the same input prices, each producing unit will exhibit the same proportionate expenditure on each productive input (shares of inputs in total cost) as will every other producing unit. When producing units have different production functions, their input expenditures will differ. The heterogeneity index measures the dispersion in relative expenditures on inputs among the establishments in an industry, or in a proposed industry. When the establishments have the same production functions, they will have the same input shares in total costs, and the heterogeneity index will be zero. The value of the index rises as establishment heterogeneity within the industry increases; that is, the index takes on a larger value as establishments with dissimilar production processes are combined into a single category.

The heterogeneity index can be used, in conjunction with other information, to judge how closely existing industries correspond to a production-oriented grouping. It can also be used to evaluate proposals to form new production-oriented industries, or to break apart or combine existing ones.

ECPC Report No. 2 also compares the results from the new heterogeneity index with the judgments that were incorporated into the matrix of ECPC Report No. 1. Note that the matrix judgments were formed before the heterogeneity index was computed, so that the matrix and the index could be used as independent evaluations. The degree of correspondence between these two completely independent evaluations, though not perfect, is both intriguing and promising (see ECPC Report No. 2).

The heterogeneity index is an important new tool that is available for implementing a production-oriented economic concept in a classification system.

C. Services classifications

The three North American countries have agreed to give special attention to classifications for services industries, as well as for high-tech and new and emerging industries. The classification of services poses special difficulties and because of this the ECPC has released a paper (ECPC Issues Paper No. 6, "Services Classifications" [4]) that discusses the application of a production-oriented economic concept to services industries.

The ECPC has been especially challenged by those who have said that our approach may be practical for goods but will not work

for services. We believe the application of production-oriented reasoning to services industries is practical, and ECPC Issues Paper No. 6 discusses practical interpretations of the economic concept. Moreover, the matrix exercise (ECPC Report No. 1) also applied the production-oriented concept to services industries in a pragmatic way, and we believe that this exercise showed that the production-oriented concept can be applied to services.

The task of classifying services industries will, however, be especially difficult. Additional special reports on the classification of services will be released as the work proceeds.

V. Applications of the ECPC's Research Approaches to the Classification Systems of Other Countries

We believe it would be especially rewarding to know the economic concepts that have been incorporated into industry definitions in classification systems outside the United States and Canada. It would also be valuable to test the heterogeneity index on the industry classifications of other countries. Exchanging the results of similar studies carried out on classification systems in use in different countries would provide a good way to determine where--that is, in which classification systems--the best ideas for industry groupings are to be found.

In the past, comparisons of different classification systems have more or less given the result: We do ours this way and we think ours is best, and you do yours that way and you think yours is best. However, we can now do better: Carrying out analysis of classification systems along the lines of ECPC Reports Nos. 1 and 2 and the Statistics Canada study [10] potentially provides a much more productive exchange of information than has been possible in the past. Rather than "splitting the difference" between mutually exclusive classification outcomes, performing some economic analysis on classification systems, of the type incorporated into the Statistics Canada and the two ECPC reports described in this paper, would produce new and valuable information for improving industry classifications.

Moreover, explicit conceptual analyses of classification systems would offer the potential for melding the international desire for comparability in industrial statistics with the goal of improving the available industrial statistics for the needs of users. Rather than setting the two goals against each other, or elevating the one over the other, as has sometimes inadvertently been true in the past, we need to gain wider understanding and support for a new approach: Constructing internationally comparable industrial statistics--where internationally comparable economies exist--that conform to a consistent economic concept provides the worldwide best course for the future of industrial statistics.

REFERENCES

[1] Bureau of the Census, *Proceedings*, International Conference on Classification of Economic Activities, Williamsburg, Virginia: U.S. Department of Commerce, November 6-8, 1991.

[2] Coats, R. H., "The Classification Problem in Statistics," *International Labour Review*, April 1925, pp. 511 and 520.

[3] Economic Classification Policy Committee, Issues Paper No. 1, "Conceptual Issues," *Federal Register*, March 31, 1993, pp. 16991-17000; Economic Classification Policy Committee, Bureau of Economic Analysis (BE-42), U.S. Department of Commerce, Washington, D.C. 20230.

[4] _____, Issues Paper No. 6, "Services Classifications," March 1994. Available from Economic Classification Policy Committee, Bureau of Economic Analysis (BE-42), U.S. Department of Commerce, Washington, D.C. 20230.

[5] _____, *Federal Register* (forthcoming). Available from Economic Classification Policy Committee, Bureau of Economic Analysis (BE-42), U.S. Department of Commerce, Washington, D.C. 20230.

[6] _____, Report No. 1, "Economic Concepts Incorporated in the Standard Industrial Classification Industries of the United States," May 1994. Available from Economic Classification Policy Committee, Bureau of Economic Analysis (BE-42), U.S. Department of Commerce, Washington, D.C. 20230.

[7] _____, Report No. 2, "The Heterogeneity Index: A Quantitative Tool to Support Industrial Classification," May 1994. Available from Economic Classification Policy Committee, Bureau of Economic Analysis (BE-42), U.S. Department of Commerce, Washington, D.C. 20230.

[8] U.S. Department of Commerce, Bureau of Economic Analysis, *Survey of Current Business* 70 (February 1990): 2.

[9] U.S. Executive Office of the President, Office of Management and Budget, *Standard Industrial Classification Manual, 1987*, 705 pages. (For sale by: National Technical Information Service, 5285 Port Royal Road, Springfield, Virginia 22161, Order no. PB 87-100012.)

[10] Young, Kenneth, "The Conceptual Basis of the Standard Industrial Classification," Standards Division, Statistics Canada, February 1994.

REVISING THE UNITED STATES STANDARD OCCUPATIONAL CLASSIFICATION (SOC) SYSTEM

Thomas J. Plewes
U. S. Bureau of Labor Statistics

I. Background

Historically, various United States Federal agencies, primarily the Department of Labor and the Bureau of the Census, have developed their own separate occupational classification systems, designed to meet their own specific statistical and programmatic needs. The lack of comparability between these various sources of occupational information and data led to multi-agency interest in and action to develop a Standard Occupational Classification (SOC) system, beginning in 1966. The SOC, first published in 1977 and revised once in 1980, was intended to provide a mechanism for cross-referencing occupation-related data collected by various economic and social statistics programs in order to maximize the analytical utility of these data.

The major underlying principle of classification in the SOC is work performed, not skills, training, education, licenses or other credentials. More specific occupations are grouped into the most detailed SOC categories based on their similarity in terms of work tasks and activities. Other classification principles include the following: SOC groupings are independent of the work setting, unless it alters the nature of the job; supervisors are identified separately from workers; large or small size is not a determinative factor for separate identification; and comparability to the international standard classification of occupations (ISCO).

The SOC was intended to be comprehensive in coverage, including all occupations for which work is performed for pay or profit, including unpaid farm work. The 1980 SOC was comprised of 664 distinct occupations at the most detailed level. It was not intended to meet all specialized analytical or organizational management purposes, but to serve as a general tool for reconciling various sources of occupational data.

In 1983, the major sources of U.S. occupational employment data -- the establishment-based Occupational Employment Statistics (OES) survey and the Current Population Survey (and Decennial Census) of households -- became more comparable when each adopted a new classification structure based on the SOC.

II. The Need for a New SOC

The SOC, unfortunately, never was implemented fully across all Federal occupation-related data collection efforts. Various

federal agencies continue to use their own distinct occupational classification structures. For example, the Department of Labor's Employment and Training Administration uses the Dictionary of Occupational Titles (DOT); the Department of Education uses its Classification of Instructional Programs (CIP); and the Office of Personnel Management has its own occupational classification structure. As a result, reconciling different occupational data sources continues to be difficult at best. In addition, the 1980 version of the SOC is outdated, as new occupations -- particularly in technical and health-related fields -- have emerged since that time (and are incorporated into some of the current occupational classification structures).

There are other reasons that attention recently has focused on occupational information. Concern with the quality of the U.S. workforce, skill formation issues, and changes in occupational structures due to new technology and shifts to "high-performance" work organizations, all highlight the importance of accurate, timely, and comparable occupational information to support program planning, career guidance, and training development. As such many users and producers of occupational data feel that it is time to re-examine the SOC and to develop a classification structure that meets the occupational information needs of the twenty-first century.

III. Actions to Inform the SOC Revision

In November 1991, the Office of Management and Budget (OMB) designated the Department of Labor as the lead agency to coordinate the development of a new U. S. Standard Occupational Classification (SOC) system by 1997, in time for implementation in the 2000 Census. Since that time, the Bureau of Labor Statistics' Office of Employment and Unemployment Statistics and the Dictionary of Occupational Titles (DOT) staff of the Employment and Training Administration (ETA) have been working together to organize activities aimed at developing information and alternative approaches related to classification principles for the new SOC. These activities have included commissioning contract papers on major occupational classification issues.

An International Occupational Classification Conference was held in June 1993, sponsored by the Bureau of Labor Statistics. The Conference provided a forum for the discussion of new ideas and alternative approaches to occupational classification issues. It included many individuals and agencies directly involved with the occupational classification user community, as well as international occupational experts from numerous countries. The papers, discussions, and ideas generated at the Conference are serving to inform revision activities for the SOC.

Some of the major issue areas addressed at the Conference are described below.

1. New challenges and alternative approaches to occupational classification: Currently, all federal occupational classification systems are based on work performed or job titles. As the pace of occupational change has increased, many people are becoming more concerned with issues of skills transferability between jobs or occupations in order to facilitate transitions in an increasingly volatile economic environment. An important issue raised during the conference is whether a new U.S. SOC should be based primarily on skill type and skill level, rather than work performed.

2. The feasibility and desirability of creating a unified occupational classification structure for government statistical and programmatic purposes: Although some Federal agencies may prefer to maintain their separate classification structures, others feel that net value could be provided to users of occupational data by developing a more unified Federal classification structure. At a minimum, there seems to be consensus that a more unified Department of Labor occupational classification structure is desirable, and movement in this direction has been occurring, even prior to the Conference. In its final report, the Secretary of Labor's Advisory Panel on the Dictionary of Occupational Titles recommended that a revised Dictionary conform to the classification structure of a revised SOC system and, in the interim, conform to the Bureau of Labor Statistics' Occupational Employment Statistics (OES) system.

3. How a revised SOC could meet the needs of users of occupational information who are dissatisfied with the current classification systems: Due to the current system of multiple occupational classification structures, users must obtain important related information -- such as demographic characteristics, industry and geographical distribution, worker attributes and skill requirements, and wages -- from different sources with different underlying classification structures. As a result, the information obtained from one source is not compatible with information derived from another source, leading to frustration on the part of many users. Another source of dissatisfaction lies with the perceived currency and accuracy of current occupational classification structures. Some structures, including the SOC, have not been updated for more than a decade, and therefore, many new occupations that have emerged as a result of new technology and changed forms of work organization are not included in current classification structures.

4. International perspectives on occupational classification and lessons for the U.S. SOC revision: A full day of the conference was devoted to international occupational classification issues. The international experience is important for two reasons: One relates to the international comparability of data, and the other relates to lessons that can be learned from the experience of other countries. A decision to move towards a common international classification system, such as the International Standard Classification of Occupations (ISCO-88), would

inevitably result in a loss of nation-specific occupational detail that many users of national data regard to be critical. In addition, there are questions about the degree to which ISCO-88 is structured on clear, consistent, and appropriate principles. The second reason to examine the international experience is to try to draw lessons from other countries, many of which have recently made substantial revisions to their national occupational classification systems. Issues explored included new approaches to principles of occupational classification (e.g., skill type and skill level); the level of effort and resources required and methodologies used to develop new systems; and the feasibility and desirability of developing a unified national classification system to replace existing disparate ones.

IV. SOC Revision Process

Following the Conference, the Office of Management and Budget established an SOC Revision Policy Committee, chaired by the Bureau of Labor Statistics, with representatives from the Bureau of the Census, the Employment and Training Administration, the Office of Personnel Management, the Defense Manpower Data Center, and, *ex officio*, the Office of Management and Budget (OMB). A Charter for the Committee recently has been approved.

The Policy Committee is charged with an examination of the Federal Government's various occupational classification systems for statistical and administrative uses, and with providing recommendations to OMB on the structure and implementation of a new SOC. The charge to the Committee includes: (1) identifying the major statistical uses of occupational classifications; (2) identifying and developing new concepts, structures, and methodologies to determine what constitutes an occupation; (3) developing a standard classification system based on these concepts; (4) planning the implementation of the new classification system; and (5) ensuring that there is ample opportunity for widespread public participation in the revision process.

The principal use of a revised SOC would be statistical, but it also would serve as a framework for administrative purposes and other occupational classifications. The Policy Committee will evaluate the utility of alternative classification structures in consideration of the following: (1) Ensuring compatibility between the descriptive material of the new Dictionary of Occupational Titles (DOT) and the revised SOC; (2) current public interest in a skills-based classification system; (3) users' needs for historical comparability of data; (4) the expertise of other countries in revising national classification systems; (5) desirability, but not necessity, of compatibility with international occupational classification systems; and (6) the need for all Federal Government occupational classification systems to be part of the SOC framework.

The Policy Committee will adopt processes that ensure ample opportunity for public participation. These processes will involve all stakeholders, including the range of occupational data users, both government and private, as well as data collectors and data providers. The Policy Committee will consider forming a Consultation Group, composed of Federal agencies not represented on the Policy Committee and interested public and private parties (e.g., States, associations, private individuals). Such a group would meet on a flow basis, as necessary, to provide input to the work of the Policy Committee. Notice of the Policy Committee's work will be widespread and will be published in the *Federal Register*, and all interested parties will be given the opportunity to be included on a mailing list.

The conceptual framework for the new SOC is to be completed prior to July 1995 to allow for testing related to the 2000 Census, as well as for the administration of the 1996 DOT National Content Test. The completed occupational classification structure should be available by July 1997 to coincide with development of the 2000 Census.

COMMENTS ON THE REVISIONS OF THE STANDARD INDUSTRIAL AND OCCUPATIONAL CLASSIFICATIONS

Joel Popkin
Joel Popkin and Company

The beginnings of the present effort to revise the Standard Industrial Classification (SIC) originated in the Census Advisory Committee structure in the mid 1980s. At that time, I represented the American Economic Association. At those committee meetings, I recall getting the assignment to comment on papers about how the SIC ought to be revised, and asking myself why I had not drawn a more interesting assignment. I clearly did not recognize that a revolution in economic classification was afoot.

Charles Waite, Associate Director of the Census Bureau, was handing out those assignments, and the papers I commented on were written by his staffers, Pamela Powell-Hill and James Monahan. That was 10 years ago, and I think marks the birth of this much needed and very important current effort to conduct a "clean-slate" revision of the SIC. Following those developments, Charles Waite planned and convened in Williamsburg what turned out to be a seminal international conference on economic classification. About the time the plans for the conference were being initiated, Jack Triplett, Chief Economist of the Bureau of Economic Analysis, presented at the 1990 Census research conference a very important paper illuminating the conceptual issues relevant to the classification of economic activity. Hermann Habermann, then Chief Statistician of the U.S. government, lent support to a continuation and formalization of those efforts.

The project to revise the SIC now has a full head of steam with a target for implementation in the 1997 economic censuses. OMB appointed the BEA lead agency, and Jack Triplett is chairman of the government-wide Economic Classification Policy Committee (ECPC). There are three elements of U.S. leadership in this significant undertaking. The first was the Williamsburg conference itself. The second was the successful negotiation among the countries of NAFTA of an agreement to develop a common, North American industrial classification system (NAICS). The third key element in the pervasiveness of this effort, and its enhanced chances of success, was that the North American plan and the Williamsburg conference were instrumental in prompting Eurostat, the statistical agency of the European Community, to reconsider some decisions it had made about industrial classification and to explore moving in the direction of fundamental rethinking of classification systems that the United States has promulgated.

There are three fundamental kinds of decisions that have to be made in designing an industrial classification system. The first is the selection of the unit of observation. The second is the concept by which individual observations should be grouped. And the third is the hierarchy along which groups should be aggregated.

Decisions about two of those three elements have already been made. The first is that the classification system will retain the establishment as the unit of observation except in cases where its use is not appropriate or feasible. The second is that the underlying concept for classification will embody a production-side approach. Establishments will be grouped together that share the same kind of production techniques and processes. The third issue, that of hierarchy, is still being studied.

Part of my role here today is that of discussant of the SIC revision process that is now underway. Some of you will be familiar with my views and recommendations if you have seen the paper I wrote for the Williamsburg conference.

With respect to the unit of observation, my recommendation was to change from an establishment based system to one which I characterized as focusing on divisions, departments or subsidiaries (DDS) within companies as units of observation. I made that recommendation for three reasons. The first is that the establishment is not as prevalent an economic unit of observation as it once was. That is at least partly due to the advances in telecommunications which permit output to be produced with more inputs obtained from different establishments within the company. That leads me to the second reason I recommended a larger unit of observation such as the DDS. It is that at a higher level of aggregation, the matching of inputs and outputs and the full accounting of all inputs may be more feasible and data collection simplified. The problem posed in using the establishment is that not all inputs can be accounted for, especially some purchased services and inputs of information, technology, and management skills from central offices and other establishments within the company. I felt that by moving the unit of observation to a higher level of aggregation within the company, those inputs could be captured and the activity of separate business units (SBO) or DDSs could be relatively well accounted for and measured. That approach is not new. It is used currently in the Census M3 report on "Inventories, Shipments, New and Unfilled Orders" in which data are collected directly from divisions of companies; and it is also being utilized in the Annual Capital Expenditures Survey (ACES) that the Census Bureau has developed. The third reason I made this recommendation was that it seemed as though it might ease reporter burden to the extent that establishment records are increasingly being consolidated at company or division levels. Nonetheless, the ECPC recommendation to use the establishment, but with a recognition that there may be exceptions, goes some distance to alleviating my concern about the use of the establishment.

With respect to classification concept there were two candidates. The production-oriented approach or the market-oriented approach. Each approach serves many legitimate uses, and both can be justified. I thought it would be inappropriate to recommend multiple classification systems simply because the resources are limited. So I thought it was necessary to recommend one approach. For me, it was the demand approach. The ECPC has adopted the production approach, but also has indicated that it is undertaking work to develop a structure in which outputs can be

classified by market grouping for both goods and services. Such market groupings (commodity product classes) already exist in the government. For example, in the price statistics program, the PPIs classified by stage-of-process indexes reflect a market-oriented measure, while the PPIRs represent the kind of price index that would be used to deflate shipments or outputs and measure productivity. I also thought that the market-oriented system would fit better into the harmonized system being used internationally to collect trade statistics.

As I mentioned, the hierarchy issue is still undecided. One recommendation I have made, described more fully in an article in the November 1993 issue of the Survey of Current Business, is to break the large service sector, which as currently defined accounts for two-thirds of the economy, into two sectors. One part would be called "distribution networks" covering retail, wholesale, transportation, communication and other network suppliers. The other grouping would consist of the traditional kinds of services which tend to be labor intensive--such as personal and business services.

As if undertaking the revision of one classification was not enough, the researchers of the federal statistical system have assumed yet another undertaking--a clean-slate look at the way we classify occupations. This effort was lead by the Bureau of Labor Statistics which in June 1993 also convened an international conference. I gave a paper at that one, too. It stressed the need to define the unit of observation--the job as I see it; then to develop an underlying concept for the grouping of jobs; and finally, a hierarchical framework within which those groupings can be aggregated. Among my conclusions in that paper was that occupational classification would serve more purposes if it could be thought of in a three-dimensional context. Jobs should first be aggregated by both type and skill level. The third dimension, though not as well defined, could be along the line of whether the job involves symbolic logical work, production process work, or in-place personal service, a classification scheme developed by Robert Reich in his book, The Work of Nations. Perhaps information, goods, and services would be another way to view such a classification at higher levels of aggregation. Perhaps, this third dimension would capture, a classification index, which in concert with the other two dimensions, would approximate how employers view or define the labor markets in which they buy factors of production. In any event, if we move in that direction, our occupational classification would resemble a three-dimensional matrix, a Rubik's cube. That would facilitate not only the analysis of markets for certain kinds of occupations, but also provide a reading on the skill level required for those occupations and the kinds of training that individuals might need to reach that skill level.

In closing, I think these two efforts to completely revise the SIC and SOC are major statistical developments with considerable impacts. I am most pleased to see U.S. government statisticians take the lead in achieving progress in these fundamental areas.

Comments on Economic Classification Revisions

Joe Mattey [1]

May 24, 1994

[1]The author is on the staff of the Board of Governors of the Federal Reserve System and is a research associate at the Center for Economic Studies (CES), U.S. Bureau of the Census. These remarks are for presentation at the seminar on classification sponsored by the Council of Professional Associations on Federal Statistics, May 25, 1994, in Bethesda, Maryland. The comments reflect the author's own views, not the official views of the Federal Reserve System or Census Bureau.

My remarks will be devoted to the industrial classification revision plans discussed by Carole Ambler in her presentation of Jack Triplett's paper. My background in this area stems both from my experience using the existing classification system–in, for example, analyzing productivity developments by industry–and from my work with plant-level data as a researcher at the Census Bureau's Center for Economic Studies.

Triplett's paper addresses six issues. First, he argues that the United States has mounted this effort to revise the classification system because users demand it; for example, users find the existing classification system inadequate for productivity studies. Second, Triplett explains that this revision of the SIC, unlike those in the past, is not going to be riddled with compromises among competing uses. The third section explains why a production-oriented concept has been chosen, and the fourth section discusses what could be done to appease those most interested in the classification concept that ran a closest second, the market-oriented concept that groups products according to their degree of substitutability. The fifth section argues that a conceptually-based classification system is practical, and the final section advocates that the research approaches of the Economic Classification Policy Committee be applied to the classification systems of other countries.

I would like to elaborate on several of these issues. First, from the perspective of productivity studies in manufacturing, I believe that the need for an improved classification system largely arises from the difficulties we have in implementing the existing system consistently over time and across surveys. For example, the four-digit SIC classification of an individual manufacturing establishment often differs depending on whether the code has been assigned on the basis of product detail collected by the Census Bureau or on the basis of information available to others who initially identify the birth of new establishments, such as the Social Security Administration, Bureau of Labor Statistics or the IRS. Moreover, even within the Census Bureau's SIC assignment system, the industry affiliation of multi-product plants can switch frequently over time. These classification difficulties (and deterioration in sampling frames for a broader range of reasons) cause published individual industry-level time-series to change too abruptly from year-to-year. Users often cope by modelling productivity at more aggregate levels, following the SIC hierarchy for the aggregation. But the current SIC hierarchy was not designed to preserve similarity of input structures upon aggregation, and the resulting aggregate analyses often do not make much sense. Thus,

for productivity analysis there is a demand for an improved SIC system in two respects; we will be much better off if the new system achieves greater continuity at detailed levels (over time and across data sources) and if the hierarchy for aggregation better preserves similarities of the production process upon aggregation.

Whether this consistency goal will be achieved ultimately is an empirical matter. The production-oriented unifying concepts of the revised classification system offer some promise of greater consistency, particularly if the process of how things are made in a given establishment tends to be more fixed than what an establishment makes. In other words, there is some hope that the new technologically based classification system will reduce the extent of SIC switching because it is easier for, say, a manufacturing plant to alter its product mix among goods that are not close substitutes than it is for that plant to change the basic manufacturing process. Ultimately, then, a fundamental task of classification is to find meaningful characteristics of establishments that are relatively fixed.

As an economist, any discussion of fixed factors of production automatically evokes images of the capital stock in place and also, to a certain extent, the human capital embodied in a firm's employees. The production orientation favored by the committee seems quite natural. My only advice is that when production processes are analyzed, particular attention should be paid to the fixity of the elements when deciding whether they are defining features of the industry.

Triplett discusses how the committee would be likely to proceed in determining the defining features of the industry in the section of his paper on whether a conceptually-based classification system is practical. He mentions three studies that demonstrate how classification decisions could be made. Two of these are "matrix papers" that offer subjective descriptions of the extent to which the existing classification systems in the United States and Canada fit the production orientation. A third paper presents a quantitative heterogeneity index for use as a diagnostic tool.

I have had the opportunity to read drafts of these papers. The overall impression that they leave is that alot of work remains to be done, particularly on achieving a consensus on the defining features of industries. The Canadian paper puts it well in saying: "In the United States, the E.C.P.C. has analyzed part of the SIC. There is an official concordance between the two classifications, so the results could be compared for similarly defined in-

dustries. The initial comparisons showed numerous differences in the way the two countries had applied the concepts. (p. 14)" Given this illustration that in a subjective process of assigning characteristics to industries, experts will differ in their application of concepts, they conclude that it would be desireable to have objective measures.

My final remarks concern the one objective measure of heterogeneity that has received the most attention, the heterogeneity index originally proposed by Frank Gollop in 1986. The second report of the Classification Committee presents this heterogeneity index for selected manufacturing industries. The basic process of using the index for classification starts with a tentative grouping of establishments into various industries. Then, for each industry, a weighted average of the differences in input cost shares among establishments in the tentative industry is computed. The relative sizes of the establishments in terms of, say, shipments, can be used as weights.

Thus far, this index has been calculated using only ten types of inputs, each of which is a very aggregate concept: production workers, other labor, fuel, electricity, purchased services, agricultural materials, mineral inputs, nondurable materials, durable materials and capital. I must confess that when inputs are defined at such an aggregative level, I find the heterogeneity index relatively useless for classification. To see this, one can contrast the results of the heterogeneity index for the fluid milk industry with the subjective process illustrated in the U.S. matrix paper. The latter paper states that "...the physical properties of fluid milk dictate many of the processing methods and the types of machinery and equipment that must be used to handle it (p. 11)." I interpret this as meaning that if a plant has the types of machinery and equipment specially designed for handling fluid milk, than it must be a milk processing plant. In contrast, the quantitative heterogeneity index just looks at the overall cost of capital among plants within the industry, without regard to the type of capital equipment. Similarly, the heterogeneity index as computed just looks at the overall cost of agricultural materials, whether or not these materials have anything to do with the defining features of milk production.

In the case of capital equipment and structures the use of the aggregative data can be defended on the grounds that detailed information is not available. However, detailed information on materials use is available from the Census of Manufactures. In some of my own work with the plant-level microdata, I have gone to the opposite extreme, singling out specific detailed

materials that comprise a large fraction of total materials costs in the industry in question. For example, I analyzed the degree of heterogeneity in the use by fluid milk processors of whole milk from dairy farms.

My own statistical work demonstrates some of the difficulties one encounters when attempting to develop a quantitative index of heterogeneity, and whether or not the Gollop index can be successfully applied depends on how these issues are resolved. For example, not all plants report data on specific materials use. Small plants, in particular, omit information on detailed materials use because the Census forms instruct them to do so if a minimum value threshold is not surpassed. Moreover, in any given industry, the inquiries on specific materials are restricted to only a few pre-selected materials. Which Census form a plant receives depends on the tentative classification of the plant. So, anyone trying to develop a quantitative index of heterogeneity for re-classifying plants faces the problem that the data needed to make such a reclassification might not be collected, exactly because the initial classification was inappropriate.

In summary, the revised classification system has the potential for helping users of the data quite a bit, particularly those interested in production function relationships. However, it seems like alot of work remains to be done to develop the consistency needed to achieve this benefit.

Session 2

DISCLOSURE LIMITATION METHODOLOGY

RESTRICTED DATA VERSUS RESTRICTED ACCESS: A PERSPECTIVE FROM PRIVATE LIVES AND PUBLIC POLICIES

George T. Duncan
H. John Heinz III School of Public Policy and Management
Carnegie Mellon University
Pittsburgh, PA 15213
Phone/FAX: (412) 268-2172/7036
email: George.Duncan@cmu.edu

1994 July 8

A paper presented to the Council of Professional Associations on Federal Statistics Seminar on New Directions in Statistical Methodology, Bethesda, MD, 1994 May 25-26. This paper draws directly on Duncan, G., Jabine, T., and de Wolf, V. (eds.) Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics, the final report of the Panel on Confidentiality and Data Access of the National Research Council and the Social Science Research Council. Thanks are extended to the panel members for their many contributions to the report. Special thanks go to Thomas Jabine and Virginia de Wolf for both their contributions to the report and for thoughts on this paper. It is dedicated to the memory of Roger Herriot, who in his work in the federal statistical system demonstrated so clearly the value of innovative thinking.

RESTRICTED DATA VERSUS RESTRICTED ACCESS: A PERSPECTIVE FROM "PRIVATE LIVES AND PUBLIC POLICIES"

George T. Duncan[1]
Carnegie Mellon University

1. Stewardship of Statistical Agencies.

A statistical agency is more an art museum than a confessional booth. Certainly the three institutions are similar in eliciting valuables under pledges of protective stewardship--indeed both the survey respondent and the penitent entrust their personal information. But more consequentially, the statistical agency shares only with the art museum a commitment to responsible dissemination to the legitimately curious. Alike, the statistical agency and the art museum must address the tension between protection and access.

Long before statistical agencies had ever sponsored a survey to obtain personal facts, the cloak of confidentiality had been extended in a religious setting. In 1215, the Latern IV Council decreed that "all the faithful, of both sexes, when they have reached the age of discretion, are to confess all their sins at least once a year to their own priest." (Bok 1983: 78) Traditionally, the received confession is treated as protected personal information, with the priest serving as an instrument of God. On the statistical front, it was not until 1890 that U.S. census legislation required census workers to swear under oath not to disclose census data except to their superiors. Likewise, art museums view protection of their treasured works as an essential function. Motivating the extension of protection by all three is a pragmatic footing: without assurances of security, each would be severely hampered in obtaining the largely voluntary contributions they require.

How does each institution protect its data? The priest silent to the curious is honorable. Contrarily, the art museum hidden to the inquisitive is ineffectual. Likewise, the statistical agency in secreting its data fails its mission.

---

[1]This paper draws directly on Duncan, G., Jabine, T., and de Wolf, V. (eds.) Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics, the final report of the Panel on Confidentiality and Data Access of the National Research Council and the Social Science Research Council. Thanks are extended to the panel members for their many contributions to the report. Special thanks go to Thomas Jabine and Virginia de Wolf for both their contributions to the report and for thoughts on this paper. It is dedicated to the memory of Roger Herriot, who in his work in the federal statistical system demonstrated so clearly the value of innovative thinking.

Restricted Data versus Restricted Access

Whether for museums or statistical agencies, the dual role of protection and dissemination is challenging, but these two pillars cannot be compromised without risking institutional collapse. Original microdata as collected from statistical surveys can no more be provided to all who might want it than the new Andy Warhol museum in Pittsburgh could freely hand over one of his renderings of Marilyn Monroe.

Generically, two dissemination strategies are possible: provide the good in restricted form, i.e., as a transformation, to a quite general audience without preconditions on use, or provide access to the good itself, but only to a restricted audience under restricted conditions. For art museums, the first strategy calls for providing reproductions, while the second strategy calls for guarded galleries. For a statistical agency, the first strategy results in dissemination of restricted data. The second strategy results in restricted access. Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics (1993), the report of the National Research Council/Social Science Research Council Panel on Confidentiality and Data Access, explores these two strategies in its Chapter 6. The purpose of this paper is to provide some perspective on the ideas and recommendations of the report on these topics of restricted data and restricted access.

2. Restricted data

Restricted data is a confidentiality-motivated transformation of the original data; it results from the application of a statistical disclosure limitation technique. Before releasing a microdata file, for example, a statistical agency might go beyond removing explicit identifiers like name, address, and Social Security number. To limit disclosure risk, the agency could, for example, give people's ages in five-year intervals rather than by the exact date of birth.

Private Lives and Public Policies gives an overview of some key concepts and techniques of disclosure limitation:

- Disclosure risk, including identity, attribute, and inferential disclosure
- Statistical procedures for disclosure limitation, both for microdata and for tabular data
- Impact of improved computer and communications technology
- Recent research on disclosure limitation

A review and evaluation of statistical disclosure limitation techniques and their

application is given in the Report on Statistical Disclosure Limitation Methodology (1994) and in Dalenius (1988) (also see, Fienberg 1993), so the treatment here will not be detailed.

Disclosure risk

As explored in Duncan and Lambert (1989), disclosure occurs when a data subject is identified from a released file (identity disclosure), sensitive information about a data subject is revealed through the released file (attribute disclosure), or released data makes it possible to infer the value of an attribute of a data subject more accurately than otherwise would have been possible (inferential disclosure).

Statistical procedures for disclosure limitation

Statistical disclosure limitation techniques involve transformations of data to limit the risk of disclosure. Use of such a technique is often called masking the data, because it is intended to hide characteristics of data subjects. Some statistical disclosure limitation techniques are designed for data accessed as tables (tabular data), some are designed for data accessed as records of individual data subjects (microdata), and some are designed for data accessed as computer databases. Common methods of masking tabular data are deleting table entries (cell suppression) and altering table entries (random error, or noise introduction). Common methods of masking microdata are deleting identifiers, dropping sensitive variables, releasing only a small fraction of the data records, and grouping data values into categories. Direct access of computer databases raises new statistical disclosure limitation issues which are only recently being addressed (see, e.g., Duncan and Mukherjee 1992; Keller-McNulty and Unger 1993).

In the case of a public-use microdata file, statistical disclosure limitation techniques can be classified into five broad categories (Duncan and Pearson, 1991):

1. Collecting or releasing only a sample of the data: For example, the Bureau of the Census first released a public-use microdata file with a 1-in-1000 sample from the 1960 Census of Population and Housing.

2. Including simulated data: This technique has not been implemented, but it is conceptually akin to including several identical limousines in a motorcade that is under threat of terrorist attack.

3. "Blurring" of the data by grouping or adding error to the individual values: Presenting subjects' ages in 10-year intervals is an example of grouping. A microdata file prepared by the Census Bureau for the National Opinion Research Center from the 1980 census masked census tract characteristics (e.g., percentage of blacks, unemployment rate) by adding random noise (Kim 1990).

4. Excluding certain attributes: Information on a doctoral graduate field of specialization might be omitted.

5. Swapping of data by exchanging the values of certain variables between data subjects: The value of some sensitive variable could be exchanged for that in, say, an adjacent record.

For data released as tables, the blurring and swapping techniques described above have been used. Three other statistical disclosure limitation techniques are unique to tables (Cox 1980):

1. Requiring each marginal total of the table to have a minimum count of data subjects.a

2. Using a "concentration" rule, also known as the (N, K)-rule, where N entities do not dominate K percent of a cell; for example, requiring that the reported aspects of two dominant businesses in a cell comprise no more than a certain percentage of a cell.

3. Using controlled rounding of table entries to perturb entries while maintaining various marginal totals.

Statistical disclosure limitation practices of federal statistical agencies
The practices of federal statistical agencies regarding statistical disclosure limitation is well-covered in Jabine (1993b), a paper commissioned by the Panel on Confidentiality and Data Access. Based on a detailed study of twelve statistical agencies, their basic finding is that, although most have standards, guidelines, or formal review mechanisms, there is great diversity in policies, procedures, and practices among them.

This finding provides the basis for the Panel's first recommendation in this area (all eight recommendations are given for convenience in the Appendix):

Recommendation 6.1. The Office of Management and Budget's Statistical Policy Office should continue to coordinate research work

> on statistical disclosure analysis and should disseminate the results of this work broadly among statistical agencies. Major statistical agencies should actively encourage and participate in scholarly statistical research in this area. Other agencies should keep abreast of current developments in the application of statistical disclosure limitation techniques.

Beginnings have been made in implementing this recommendation. In early 1992 the Statistical Policy Office convened an ad hoc interagency committee of ten persons to be chaired by Nancy Kirkendall of the U. S. Energy Information Administration. The mandate of the committee was to review and evaluate statistical disclosure limitation methods used by federal statistical agencies and to develop recommendations for their improvement. Subsequently, the ad hoc committee became the Subcommittee on Disclosure Limitation Methodology, operating under the auspices of the Federal Committee on Statistical Methodology. Its final product, the Report on Statistical Disclosure Limitation Methodology, notes, "the development and publication of this report is directly responsive to the CNSTAT Panel's Recommendation 6.1, which says, in part, that 'The Office of Management and Budget's Statistical Policy Office should continue to coordinate research work on statistical disclosure analysis and should disseminate the results of this work broadly among statistical agencies.'" In the report's Chapter VII, a research agenda is laid out for disclosure limitation methodology. A reasonable expectation is that further progress on dissemination will be made by the dissemination of the Subcommittee's report, the presentations at the Council of Professional Associations on Federal Statistics (COPAFS) Seminar on New Directions in Statistical Methodology, and publications in the OMB Statistical Policy Working Paper series.

The Panel was concerned with the impact of statistical disclosure limitation procedures on the quality of the data as it is disseminated to data users. Statistical disclosure methods can hide or distort relations among study variables and result in analyses that are incomplete or misleading. Because of this possibility, policy researchers have expressed serious reservations about the implementation of statistical disclosure limitation (e.g., Smith 1991). Further, data masked by some disclosure limitation methods can only be analyzed accurately by researchers who are highly sophisticated methodologically. Based on these findings, the panel made the following recommendation:

> Recommendation 6.2. Statistical agencies should determine the impact on statistical analyses of the techniques they use to mask data. They should be sure that the masked data can be accurately analyzed by a range of typical researchers. If the data cannot be

> accurately analyzed using standard statistical software, the agency should make appropriate consulting and software available.

Unfortunately, this recommendation has yet to be addressed, or to appear on the research agenda of statistical agencies. The Report on Statistical Disclosure Limitation Methodology is moot on this topic.

Given the potential difficulties that certain statistical disclosure limitation techniques can cause for analysts, it is important that federal statistical agencies involve data users in selecting such procedures. As Greenberg (1991:375) notes, "survey sponsors and data users must contribute to the decision making process in identifying areas in which some completeness and/or accuracy can be sacrificed while attempting to maintain as much data quality as possible." These thoughts led to the Panel's third recommendation:

> Recommendation 6.3. Each statistical agency should actively involve data users from outside the agency as statistical disclosure limitation techniques are developed and applied to data.

Steps toward implementation of this recommendation are being made through the inclusion of individuals outside the agency on microdata review panels. It remains to be seen whether the views of data users will be adequately represented.

Finally, over the past thirty years various agencies have released public-use microdata files successfully. Based on experience, such data dissemination has met a two-pronged test: (1) the microdata files have been useful to researchers and policy analysts and (2) confidentiality has been protected. Based on this finding, the panel made a final recommendation in this area:

> Recommendation 6.4. Statistical agencies should continue widespread release, with minimal restrictions on use, of microdata sets with no less detail than currently provided.
> Given an increased public concern over privacy and confidentiality issues,

Recommendation 6.4 presents a real challenge to statistical agencies. Far easier it would be to turn inward and protective. To do so, however, would be to abdicate the statistical agency's responsibility to provide the data a democratic society needs.

The panel noted that expansion of the number and richness of public-use microdata files to be disseminated would be better justified if all users were

subject to sanctions for disclosure of information about individually identifiable data subjects. Reference was made to a recommendation, in another chapter, as follows:

> Recommendation 5.3 There should be legal sanctions for all users, both external users and agency employees, who violate requirements to maintain the confidentiality of data.

3. Restricted access: Administrative procedures to protect confidentiality

Procedures for providing restricted access to data typically establish eligibility requirements for access and impose a variety of conditions governing the purposes for which the data can be used, which organizations and individuals can have access, the location of access, physical security measures, and the retention and disposition of initial and secondary data files.

Arrangements for providing restricted access to federal data for statistical purposes do exist. Jabine (1993a) provides 19 examples, including both interagency data sharing and arrangements with data users external to the federal government.

Interagency data sharing

There have been instances of agreements to permit interagency sharing of identifiable, or potentially identifiable, personal records for statistical purposes. Some of the instances involved transfers of administrative records; others involve transfers of data collected in statistical surveys. As identified in Private Lives and Public Policies, the mechanisms used to insure confidentiality in a selected set of instances included the following:

- Making data users in the receiving agency special sworn employees of the sharing agency
- Restricting further dissemination of data and follow up with respondents
- Periodic on-site inspections of the receiving agency's security measures by the sharing agency
- Regular review of the benefits of the sharing arrangement
- Written agreement that a specified data match would not be used for any other purpose and that the receiving agency would return the shared data

file when the match was completed

- Minimizing the possibility of using linked data to identify an individual in a public-use file and then using the survey information in the identified individual's record for administrative purposes by data masking

In general, an obvious requirement for interagency data sharing is that the statutory requirements for confidentiality of all of the agencies involved must be observed. A second requirement is that the transfer of data among agencies must be consistent with statements made to data providers when the data were obtained from them.

Developing arrangements for interagency data sharing can be a complex and time-consuming process, especially if more than two agencies are involved or if novel applications of the data are planned. New initiatives are likely to pose new legal, ethical, administrative, and policy questions. The expected benefits in cost savings or better quality data must be substantial to justify the level of effort and perseverance needed to find acceptable answers. It helps if the proposed data-sharing arrangements offer benefits to all of the parties concerned.

The success of the instances examined in efficiently using data resources while protecting confidentiality support the panel's first recommendation regarding restricted access.

> Recommendation 6.5. Federal statistical agencies should strive for a greater return on public investment in statistical programs through carefully controlled increases in interagency data sharing for statistical purposes and expanded availability of federal data sets to external users.

Full realization of this goal wil require legislative changes, as discussed in Chapter 5 of Private Lives and Public Policies, but much can be accomplished within the framework of existing legislation.

External data users

The availability of high-speed computers and sophisticated analytic techniques and software have generated vastly increased appetites for federal statistical data. In many cases if the data are restricted sufficiently to ensure confidentiality, the released data will not satisfy the needs of users. Appropriate to such cases, several modes of restricted access for external data users have

been developed by statistical agencies. Some of the important features of these access modes are eligibility criteria, location of access, cost and convenience for agencies and users, and methods of protecting confidentiality. Particular modes of restricted access include the following:

- Use of a fellows program with access at the agency's central facility, for a limited term, and only for projects that the host agency deems to be of interest

- Remote access to computer databases with automated screening of batch process programs

- Use of encrypted CD-ROM products which have statistical software that is restricted so as to prevent the user from obtaining unencrypted individual records or statistics that would tend to disclose individual information.

- Release of microdata under licensing agreements that provide for special sworn employee status, authorize unscheduled site visits to the data user, provide for prepublication review by the disseminating agency, and require return or destruction of the data when the research is completed.

- Ease on-site access of data users by providing access at agency regional centers.

Given this history and the value to society of broad dissemination of federal statistical data, the panel made the following two recommendations:

> Recommendation 6.6. Statistical agencies, in their efforts to expand access for external data users, should follow a policy of responsible innovation. Whenever feasible, they should experiment with some of the newer restricted access techniques, with appropriate confidentiality safeguards and periodic reviews of the costs and benefits of each procedure.

> Recommendation 6.7. In those instances in which controlled access at agency sites remains the only feasible alternative, statistical agencies should do all they can to make access conditions more affordable and acceptable to users, for example, by providing access at dispersed agency locations and providing adequate user support and access to computing facilities at reasonable cost.

Restricted Data versus Restricted Access

Finally the panel supported archiving of important statistical data:

> Recommendation 6.8. Significant statistical data files, in their unrestricted form, should be deposited at the National Archives and eventually made available for historical research uses.

This recommendation is intended to cover statistical databases from censuses and surveys and those, like the Statistics of Income and the Continuous Work History Sample databases, that are derived from administrative records. The panel was purposely not specific as to the content of such archived databases and the length of time for which confidentiality restrictions should continue to apply. Some databases, like the economic and population censuses, might include explicit identification of data providers. Others, especially those based on samples, might not include names and addresses, but would not be subject to statistical disclosure limitation procedures of the kind that are applied to public-use microdata sets for contemporary use.

4. Conclusions

There is an inverse relationship between restrictions on data and restrictions on access: as data restrictions increase, fewer restrictions on access are needed and vice versa. A given level of confidentiality can be achieved with various combinations of restricted data and restricted access. Just as an art museum may sell reproductions, provide carefully monitored access to galleries, and allow qualified art historians considerable latitude in examination of a work, a statistical agency must choose an appropriate mix of data products to disseminate that will serve the needs of their various data users. A strong beginning has been made by the federal statistical system in developing a research and implementation agenda for restricted data. This is evident from the important contribution of the Report on Statistical Disclosure Limitation. I ponder the contribution that might be made through a comparable effort in developing a research and implementation agenda for restricted access. No less, I ponder the restricted data and restricted access procedures that will be required to ensure data access with confidentiality in the computer databases of the Global Information Infrastructure.

REFERENCES

Bok, S. (1983) Secrets: On the Ethics of Concealment and Revelation New York: Random House.

Dalenius, T. (1988) Controlling Invasion of Privacy in Surveys Department of Development and Research, Statistics Sweden.

Duncan, G. T., Jabine, T., and de Wolf, V. (1993) Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics Washington, DC: National Academy Press.

Duncan, G. T. and Lambert, D. (1986) The risk of disclosure for microdata. Journal of Business and Economic Statistics 7(2):207-217.

Duncan, G. T. and Mukherjee, S. (1992) Confidentiality protection in statistical databases: a disclosure limitation approach. Proceedings of the International Seminar on Statistical Confidentiality, International Statistical Institute.

Fienberg, S. (1993) Conflicts between the needs for access to statistical information and demands for confidentiality. Technical Report #577, Department of Statistics, Carnegie Mellon University.

Jabine, T. (1993a) Procedures for restricted data access. Journal of Official Statistics 9(2):537-589.

Jabine, T. (1993b) Statistical disclosure limitation practices of United States statistical agencies. Journal of Official Statistics 9(2):427-454.

Keller-McNulty, S. and Unger, E. (1993) Database systems: inferential security. Journal of Official Statistics, 9(2)475-499.

Kim, J. (1990) Masking Microdata for National Opinion Research Center. Final Project Report. Bureau of the Census, Washington, D.C.

Report on Statistical Disclosure Limitation Methodology (1994) Statistical Policy Office, Office of Management and Budget, Washington, DC.

Smith, J. P. (1991) Data confidentiality: a researcher's perspective. American Statistical Association 1991 Proceedings of the Social Statistics Section. Alexandria, VA: American Statistical Association.

APPENDIX. Recommendations

Recommendation 6.1. The Office of Management and Budget's Statistical Policy Office should continue to coordinate research work on statistical disclosure analysis and should disseminate the results of this work broadly among statistical agencies. Major statistical agencies should actively encourage and participate in scholarly statistical research in this area. Other agencies should keep abreast of current developments in the application of statistical disclosure limitation techniques.

Recommendation 6.2. Statistical agencies should determine the impact on statistical analyses of the techniques they use to mask data. They should be sure that the masked data can be accurately analyzed by a range of typical researchers. If the data cannot be accurately analyzed using standard statistical software, the agency should make appropriate consulting and software available.

Recommendation 6.3. Each statistical agency should actively involve data users from outside the agency as statistical disclosure limitation techniques are developed and applied to data.

Recommendation 6.4. Statistical agencies should continue widespread release, with minimal restrictions on use, of microdata sets with no less detail than currently provided.

Recommendation 6.5. Federal statistical agencies should strive for a greater return on public investment in statistical programs through carefully controlled increases in interagency data sharing for statistical purposes and expanded availability of federal data sets to external users.

Recommendation 6.6. Statistical agencies, in their efforts to expand access for external data users, should follow a policy of responsible innovation. Whenever feasible, they should experiment with some of the newer restricted access techniques, with appropriate confidentiality safeguards and periodic reviews of the costs and benefits of each procedure.

Recommendation 6.7. In those instances in which controlled access at agency sites remains the only feasible alternative, statistical agencies should do all they can to make access conditions more affordable and acceptable to users, for example, by providing access at dispersed agency locations and providing adequate user support and access to computing facilities at reasonable cost.

Restricted Data versus Restricted Access

Recommendation 6.8. Significant statistical data files, in their unrestricted form, should be deposited at the National Archives and eventually made available for historical research uses.

Statistical Disclosure Limitation Methodology

by

Nancy J. Kirkendall, Energy Information Administration

Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology was released in May 1994. This working paper reflects the efforts of the Subcommittee on Disclosure Limitation Methodology of the Federal Committee on Statistical Methodology. I was the chair of the Subcommittee. The other members are William Arends, National Agricultural Statistics Service; Lawrence Cox, Environmental Protection Agency; Virginia de Wolf, Bureau of Labor Statistics; Arnold Gilbert, Bureau of Economic Analysis; Thomas Jabine, Committee on National Statistics; Mel Kollander, Environmental Protection Agency; Donald marks, Department of Defense; Barry Nussbaum, Environmental Protection Agency; and Laura Zayatz, Bureau of the Census.

Working Paper 22 presents a basic introduction to statistical disclosure limitation, describes the methods used by 12 Federal Statistical Agencies, provides more detail on techniques used to protect tables and microdata, and discusses needed research. It also presents the Subcommittee's recommendations. The previous Statistical Policy Working Paper on the subject of disclosure limitation was *Statistical Policy Working Paper 2*, which was published in 1978. While *Working Paper 22* is an update of *Working Paper 2* in some sense, one of our primary purposes was to summarize and describe the current techniques which are used to protect data, and to make recommendations concerning what the subcommittee felt should be done. It is primarily intended to serve as a practitioner's handbook.

The purpose of this paper is to summarize the information and the recommendations made in *Working Paper 22*.

Disclosure Limitation

"Federal agencies and their contractors who release statistical tables or microdata files are often required by law or established policies to protect the confidentiality of individual information. This confidentiality requirement applies to releases of data to the general public; it can also apply to releases to other agencies or even to other units within the same agency. The required protection is achieved by the application of statistical disclosure limitation procedures whose purpose is to ensure that the risk of disclosing confidential information about identifiable persons, businesses, or other units will be very small."[1]

The historical method of providing data to the public is via tables. Beginning in 1962 with the advent of the computer age, agencies also started releasing microdata files. In a microdata file, each record contains a set of variables that pertain to a single respondent. The variables relate to that respondent's reported values. However, there are no identifiers on the file, and the data may be disguised in some way to make sure they do not reveal the respondent's identity.

---

[1]*Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology*, p. 1.

For our purposes there are two types of disclosure. **Identity disclosure**, occurs when a specific respondent can be identified from the data. Identity disclosure is particularly important to microdata files, and the solution is to limit or modify the identifying information on the file. **Attribute disclosure** occurs when confidential information about the respondent is revealed. This type of disclosure is particularly important to tables (where it is assumed that one might know if a person is represented in the table), and the solution is to make sure a sufficient number of respondents contribute to each cell in the table.

A distinction is also made between tables of **frequency** data and tables of **magnitude** data. A simple example illustrates the difference. Assume that a survey provides data on a person's profession, his salary, and the county in which he lives. Let us assume that in Franklin county, we had the following three respondents who reported that they were doctors.

Example Cell in Profession x County table
{Doctors, Franklin county}.

Number	Count	Salary
1	1	$600,000
2	1	$ 75,000
3	1	$ 75,000
Total	3	$750,000

With this example, if we publish the total for counts (3), we say we have count data. If we publish the percent of people surveyed who were doctors, we say we have frequency data. With frequency or count data every respondent contributes exactly the same amount to the cell, and methods of identifying sensitive cells depend only on the number of respondents contributing to a cell.

On the other hand, the salaries are called magnitude data. Here the respondent's contribution to the cell total depends on his reported value. Let us assume that the two doctors who are less well paid are local general practitioners, and the third is a heart surgeon who works in the city, but lives in Franklin County. Publishing the total salary would allow each of the local doctors to make a very good estimate for the salary of the heart surgeon. If they can estimate his salary "too closely", we would say that we have attribute disclosure. Thus, for tables of magnitude data, the method of determining sensitive cells depends on the values reported by each respondent.

In the next few sections of this paper, we will illustrate the methods used to protect data and present the Subcommittee's recommendations. Section 1 concerns tables of frequency or count data; Section 2 tables of magnitude data; and Section 3 microdata. Section 4 is a summary.

1.0 Tables of Frequency (Count) Data

A cell in a table of frequencies or counts is sensitive if there are too few respondents. The methods used to protect such cells include:

1. Collapse categories (combine rows or columns).

2. Suppression.

3. Controlled (random) rounding.

4. Confidentiality edit.

Both collapsing categories and suppression are widely used by Federal agencies, and have been for years. Random rounding and controlled rounding have not actually been used by Federal agencies. The confidentiality edit is a new method which was used to protect tables from the 1990 decennial Census.

Assume that cells are defined to be sensitive if they have three or fewer respondents. The following table is an example we will use to illustrate different ways of protecting the sensitive cells. The cells which are sensitive are printed in bold with an asterisk.

Table 1 -- Example -- with Disclosure

Household Head Education Level

County	Low	Med	High	Very High	Total
A	15	**1***	**3***	**1***	20
B	20	10	10	15	55
C	**3***	10	10	**2***	25
D	12	14	7	**2***	35
Total	50	35	30	20	135

Source: Numbers taken from Cox, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column heading are fictitious.

1.1 Combine categories

As noted above, one way of protecting the sensitive cells is to combine rows and/or columns. In the following table, the education levels are combined into two categories. Clearly, the result is that there are no sensitive cells. However, a lot of information is lost.

Table 2 -- Example Without Disclosure Protection Provided by Combining Rows or Columns

Household Head Education Level

County	Low/Medium	High/Very High	Total
A	16	4	20
B	30	25	55
C	13	12	25
D	26	9	35
Total	85	50	135

1.2 Suppression

The second method of providing protection is to simply withhold from publication the sensitive cells and a combination of other cells in each row and column so that it is not possible to derive the value of the sensitive cells by subtraction using the published marginal totals. Clearly, we need at least two suppressed cells in every row and column, but is that enough? The answer is no, and here is the counter example.

Table 3 -- Example With Disclosure Protection Not Provided By Suppression

Household Head Education Level

County	Low	Med	High	Very High	Total
A	15	S_1	S_2	S_3	20
B	20	S_4	S_5	15	55
C	S_6	10	10	S_7	25
S	S_8	14	7	S_9	35
Total	50	35	30	20	135

Source: Numbers taken from Cox, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column heading are fictitious.

To show that this table still contains disclosures, consider the sum of row 1 and row 2 minus the sum of column 2 and column 3. This reduces to the following equation:

$$(15 + S_1 + S_2 + S_3) + (20 + S_4 + S_5 + 15) - (S_1 + S_4 + 10 + 7)$$

$$- (S_2 + S_5 + 10 + 7) = 20 + 55 - 35 - 30$$

or

$$S_3 = 1$$

This illustrates that selection of cells for complementary suppression is not a trivial matter. Methods of linear programming are used to select the set of cells which are "optimal" in some sense and which protect the sensitive cells. The following table with suppressions does protect the sensitive cells.

Table 4 -- Example Without Disclosure Protection Provided by Suppression

Household Head Education Level

County	Low	Med	High	Very High	Total
A	15	S	S	S	20
B	20	10	10	15	55
C	S	S	10	S	25
D	S	14	S	S	35
Total	50	35	30	20	135

Source: Numbers taken from Cox, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column heading are fictitious.

This example leads to the first of our recommendations. When suppression is used to protect tabular data, whether frequency or magnitude data, the table with suppressions should be *audited*. Auditing involves applying a linear programming algorithm to calculate the largest value a suppressed cell can take and the smallest value it can take. If the largest value and the smallest value are equal, the cell total is revealed exactly. If they are "too close" then the cell value can be estimated "too closely".

1.3 Random Rounding or Controlled Rounding

With random or controlled rounding, each cell count is rounded using some base value. In the following example, the base value is 5. In this case each cell count can be written as $X = 5q + r$. For random rounding each cell is rounded at random. This cell would be rounded up with probability r/5, and down with probability 1-r/5. The problem with this procedure is that tables do not add, as illustrated in the Table 5.

Table 5 -- Example Without Disclosure Protection Provided by Random Rounding

Household Head Education Level

County	Low	Med	High	Very High	Total
A	15	0	0	0	20
B	20	10	10	15	55
C	5	10	10	0	25
D	15	15	10	0	35
Total	50	35	30	20	135

Source: Numbers taken from Cox, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column heading are fictitious.

Random rounding has been used by Statistics Canada and was used by the New Zealand Department of Statistics before they moved to controlled rounding. The New Zealand Department of Statistics moved to controlled rounding primarily because users complained that the randomly rounded tables did not add (George and Penny, 1987.)

Controlled rounding is like random rounding except that a linear programming method is used to impose the constraint that the table must add. Controlled rounding was a topic of research during the 1980's, and for two dimensional tables and most three dimensional tables current methods work very well. It was proposed for use with the 1990 decennial census (Greenberg, 1986), but has not yet been used by any Federal statistical agency. An example of our tab' protected with controlled rounding is presented below.

Table 6 -- Example Without Disclosure Protection Provided by Random Rounding

Household Head Education Level

County	Low	Med	High	Very High	Total
A	15	0	5	0	20
B	20	10	10	15	55
C	5	10	10	0	25
D	10	15	5	5	35
Total	50	35	30	20	135

Source: Numbers taken from Cox, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column heading are fictitious.

1.4 Confidentiality Edit

All of the above methods are applied to a specific table. If the table is changed in some way, or another table containing data from the same data file is constructed, another detailed analysis must follow to assure that consistent protection is applied.

The confidentiality edit is a new method which was developed at the U. S. Census Bureau and used to protect tables from the 1990 Census (Griffin, Navarro, and Flores-Baez, 1989). With this method the original microdata file is manipulated (much as it would be if it were going to be released for public use). After manipulation the microdata file can be used directly to make tables. Other tables made from the same manipulated microdata file will also be protected, and the protection will be consistent. The approach described below was used for the regular decennial Census data file (the 100 percent data file), it uses a microdata protection technique called "data swapping" or "switching" (Dalenius and Reiss, 1982).

To apply the confidentiality edit the following steps are applied.

1. Take a random sample of records from the microdata file;
2. Find a match with them in some other county, based on a set of key variables;
3. Swap all other variables on the matched records;
4. Make tables

After the confidentiality edit, our table might appear as below.

Table 7 -- Example Without Disclosure Protection Provided by Confidentiality Edit

Household Head Education Level

County	Low	Med	High	Very High	Total
A	13	2	4	2	20
B	18	12	8	17	55
C	5	9	11	0	25
D	14	12	8	1	35
Total	50	35	30	20	135

The only disadvantage I have seen quoted is that the table does not look as if disclosure limitation has been applied.

1.5 Recommendation

While each of these methods has advantages and disadvantages, the Subcommittee was unable to determine which of these methods were preferable in terms of the level of protection applied, and the usefulness of the result. Our recommendation is that further research should be done to address this question, and the result widely disseminated.

2.0 Tables of Magnitude Data

For tables of magnitude data only two methods can be used to protect sensitive cells. They are combining categories, and suppression. Each has the same strengths and weaknesses as discussed above, and if suppression is used the table should be audited. For tables of magnitude data, the new question is how to identify sensitive cells?

We indicated above that the respondents' reported values are used. In fact, cells are identified as sensitive if a simple linear combination of respondent level data is positive. The linear equation is called a linear sensitivity rule and the coefficients depend on the specific rule used and the parameters chosen. There are three rules which are commonly used:

(n,k) rule -- a cell is sensitive if n respondents contribute k% or more to the cell total;

p-percent -- a cell is sensitive if the published total can be used to estimate any respondent's data more accurately than p-percent;

pq -- like the p-percent rule, but acknowledges that before data are published, common knowledge allows estimation of any respondents' data to within q percent ($q > p$).

Recommendations

The Subcommittee's recommendations for tables of magnitude data are:

1. Only subadditive linear sensitivity measures should be used to identify sensitive cells. Subadditivity is a mathematical property that assures that if two or more cells are not sensitive, then their sum (union) is not sensitive either. Fortunately, all three commonly used linear sensitivity rules are subadditive.

2. The committee prefers the p-percent or pq rules as providing more consistent protection.

3. Suppression or collapsing categories are the only accepted methods of protecting sensitive cells.

4. The parameter values used in practice should not be revealed.

5. Tables containing suppressions should be audited.

For tables of magnitude data research is needed into identifying summary statistics to publish as a replacement for a sensitive cell total. If it could be shown that the summary statistics do not reveal individual data, they could be used instead of suppression and provide users with more information.

3.0 Microdata

For tables, we have associated "disclosure" with the publication of "sensitive cells", and have justified a simple way to identify which cells are sensitive. Once that is done, several approaches have been used to protect the sensitive cells. Unfortunately, for microdata files there is no standard agreed to definition of what constitutes "disclosure", other than uniquely identifying an individual in a data file.

The following four common ways to protect microdata files are used by virtually every agency which releases microdata files.

1. Use only a sample of the population. (A sample protects an individual's data, because it is not generally known whether or not a particular individual is included in the file.)

2. Remove obvious identifiers (eg. name, address, social security number).

3. Limit geographic detail (detailed data about an individual from too small a geographic region increases the risk of identification.)

4. Top code, bottom code and\or recode continuous high visibility variables. (Recoding continuous variables essentially makes them discrete. The larger values are shown only as greater than some number, the smaller values are shown as less than some number, and the intermediate values are assigned to a range.)

Salary is an example of a high visibility continuous variable. It may take many different values, and either very large ones, very small ones, or very precise recording of the value may reveal a respondent's identity. (Like our highly paid heart surgeon.) Other ways of protecting microdata are also applied to high visibility continuous variables. They include:

5. Masking (add or multiply by random numbers);

6. Swapping or rank swapping (find two records which match on a selected set of variables and exchange (swap) the remaining variables);

7. Blank and impute for randomly selected records. (randomly select a set of records, eliminate specific reported variables and replace them by imputed values);

8. Blurring -- aggregate values across small groups of respondents. (find a group of respondents, average some of their variables, and replace the reported values by the average.)

Recommendations

The subcommittee could only make one fairly obvious recommendation for protecting microdata files.

Remove direct identifiers and limit other identifying information.

Research is needed into defining disclosure or an unacceptable likelihood of disclosure for microdata files. Another area of needed research is into the impact of disclosure limitation techniques on the usefulness of the resultant data file. The subcommittee believes that research into these topics was of the highest priority.

4.0 General Recommendations and Summary

In addition to the specific recommendations above, the subcommittee had the following general recommendations. Agencies should

1. Seek advice from respondents and data users. Respondents should be asked about variables they consider sensitive and those they do not consider sensitive. It would be better if agencies applied disclosure limitation methods only to variables considered sensitive by respondents. Data users should be offered the opportunity to comment on disclosure limitation methods. Agencies should use this information in selecting the disclosure limitation methods to use.

2. Centralize review of disclosure limited products within an agency. A centralized review of disclosure limited products assures consistency in the application of disclosure limitation within an agency. In addition, a centralized review provides greater assurance that the data are adequately protected.

3. Share software and methodology. Agencies need to help each other to assure consistency in practice, and to make more advanced methodology and software widely available.

4. Agencies which release the same or similar data sets should cooperate in the application of disclosure limitation to those data sets. If there is no coordination, it is more likely, for example, that cells selected for complementary suppression by one agency, might not be suppressed by the other agency. This would lead to disclosure.

This paper has provided an elementary description of statistical disclosure limitation methodology and the principle recommendations of the Subcommittee on Statistical Disclosure Limitation Methodology. *Working Paper 22* provides considerably more detail on statistical disclosure limitation methodology, agency practices and needed research. It also provides an extensive annotated bibliography. The Subcommittee hopes that you find the information useful.

References

Cox, L. H., Johnson, B. McDonald, S. Nelson, D. and Vazquez, V. (1985 Confidentiality Issues at the Census Bureau," Proceedings of the Bureau of the Census First . ual Research Conference, Bureau of the Census, Washington D.C. pp 199-218.

Dalenius, T. and Reiss, S. P. (1982), "Data Swapping: A Technique for Disclo Control", Journal of Statistical Planning and Inference, Vol 6, pp. 73-85.

George, J. A. and Penny, R. N. (1987), "Initial Experience in Implementing Controlled Rounding for Confidentiality Control, Proceeding of the Bureau of Census Third Annual Research Conference, Bureau of the Census, Washington DC., pp 253-262.

Greenberg, B. (1986), "Designing a Disclosure Avoidance Methodology for the 1990 Decennial Census," presented at the 1990 Census Data products Fall Conference, Arlington, VA.

Griffin, R. A., Navarro, A. and Flores-Baez, L. (1989), "Disclosure Avoidance for the 1990 Census, Proceedings of the Section on Survey Research Methods,, American Statistical Association, Alexandria, VA, pp. 516-521.

Discussion of Presentations on Statistical Disclosure Methodology[1]

Stephen E. Fienberg[2]

1 Prologue

This past weekend my wife and I were attending a Bat Mitzvah and the daughter of our friends read her portion of the Torah from the *Book of Numbers* dealing with the census of the Israelites in the desert. As I listened to her, I read this passage from the bible again with special care with the hope of some divine inspiration for my discussion of the two papers presented today. Let me share with you what I learned about disclosure limitation.

First, the census seemed to be much easier to take than we have found to be the case in modern times in the United States. There is no mention of an undercount, differential or otherwise, although women and children were intentionally omitted from the count. It turns out that there were 603,550 Israelites aged 20 and above, and the bible gives various breakdowns of these totals, without any reference to or apparent concern for confidentiality.

[1]Presented at "Seminar on New Directions in Statistical Methodology," sponsored by the Council of Professional Associations on Federal Statistics, Bethesda, MD, May 25-26, 1994

[2]Stephen E. Fienberg is Maurice Falk Professor of Statistics and Social Science at Carnegie Mellon University, Pittsburgh, PA, 15213. The preparation of this discussion was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada to York University.

Second, the 12 tribes were organized around the tabernacle and in that sense we could think of the tribes as corresponding to geographic areas. Part of the reported data goes down to subgroups whose order of magnitude is a few thousand. Clearly, this would not meet the Census Bureau requirement for the release of identification of geographic codes for microdata sets where the threshold is now 100,000.

Third, while the bible contains no cross-tabulations as we know them today, it does include considerable information that could be displayed in cross-classified form. But even the smallest numbers reported, e.g., the 273 for the number of first born of the Levites, would not seem to provide an example requiring cell suppression.

Fourth, the bible actually releases the names of several individuals who participated in the census, especially the names of a number of the tribal leaders and their sons. This suggests that the Israelites didn't have any hang-up about the issue of uniques in the population for the release of census data. The idea seems to be that there is the need to distinguish whether or not the release is in fact harmful. After all, everyone knew that Moses, Aaron, and a number of others were included in the census and what their demographic classifications were. Therefore, identifying them by name did not compromise them in any way.

Finally, there were few or no subsequent releases from the biblical census

so we don't have much evidence about how the Israelites would have treated concerns about confidentiality. We do know, however, that there was nothing corresponding to Title 13 in the Torah or in the commentaries such as the Talmud.

Having noted all of this in the way of prefatory remarks, let me now turn to the two papers in this session.

2 Duncan on *Private Lives and Public Policies*

George Duncan has summarized the major recommendations from Chapter 6 of *Private Lives and Public Policies*, a report issued by the NRC-SSRC panel he chaired as they relate to statistical procedures to protect confidentiality. His paper begins with a discussion of a 1215 Latern IV Council decree on confidentiality and quickly shifts to statistical agencies' dual role of protector and disseminator of data. He then takes up the panel's themes of restricted data (via some transformation) versus restricted access. To do this, he needs to define disclosure and, in keeping with the literature, discusses this at three levels: individual disclosure, attribute disclosure, and inferential disclosure, and he lists some standard techniques for providing restricted data to achieve disclosure avoidance. This material is a brief introduction to that which is covered in much greater detail in chapter II of the draft Federal Committee on Statistical Methodology Working Paper 22, *Statistical Disclosure Limitation Methodology*, described by Nancy Kirkendall. In my remarks I will focus

on the panel's recommendations regarding restricted data and those aspects of the topic dealt with in Working Paper 22.

Because of the great diversity in policies and practices of the statistical agencies (documented in the panel report and in chapter III of Working Paper 22) the panel recommended that OMB should continue to coordinate research work on disclosure limitation and disseminate the results widely. The existence of the Subcommittee on Disclosure Limitation Methodology of the Federal Committee on Statistical Methodology and its recently released working paper represent OMB's and the agencies' positive response. The panel's second recommendation relates to agency assessments of the impact of their own data disclosure limitation techniques and Working Paper 22 remains silent on the matter, a point to which I will return in a few moments.

A few years ago I argued that the statistical agencies in the U. S. clearly were using techniques that were too conservative, i.e., that they erred too much on the side of restricting data in order to ensure that guarantees of confidentiality are not compromised as opposed to increasing the extent and utility of released data. I was immediately challenged and I offered as evidence to support my proposition the total absence of anecdotes where, despite agency actions, confidentiality had been breached. Agencies must remember that they are only public protectors and not owners of the data and they need to involve users in the choice of disclosure avoidance procedures. This is the third of the panel's recommendations and this, according to Duncan, is in the process of implementation by a number of agencies. The panel's final

recommendation encouraged the continued widespread release of microdata sets.

I have watched the NRC/SSRC panel from conception through the completion of its report. While the four recommendations I have singled out here from Chapter 6 of the report sound much like apple pie and motherhood, they and the other recommendations of the panel are clearly designed to move the practice of statistical data disclosure forward and encourage the development of a statistical basis for confidentiality practices. I heartily recommend the report and its companion volume of technical commissioned papers which appeared as a special issue of the *Journal of Official Statistics* in the fall of 1993.

3 Kirkendall on *Statistical Disclosure Limitation Methodology*

Nancy Kirkendall has described some of the ideas and materials from Working Paper 22 of the Federal Committee on Statistical Methodology Subcommittee on Disclosure Limitation Methodology, an activity which she chaired. This working paper needs to be considered against the backdrop of an earlier Federal Committee on Statistical Methodology working paper on the topic issued in 1978. What we have here is a major update with considerable detail and an extensive annotated bibliography. Depending on how we approach the topic, we find both good news and bad.

First, the good news. Much has happened in the intervening 16 years. The earlier working paper was technically innovative and it served as a catalyst to the development of new disclosure limitation methodology, especially in such agencies as the Bureau of the Census, but also by those in universities such as George Duncan and my former colleague Diane Lambert, and by Tore Dalenius, my fellow discussant today. The new working paper documents many of these advances and the extent of the research developed is impressive. So too are the advances in the uses of disclosure limitation methodology by federal statistical agencies. The current agency practices, as described in chapter III of Working Paper 22, are far more advanced thanks both to the methodological developments and to attendant advances in computation. In these senses, the new working paper represents a major progress report on the health of the federal statistical system.

Next, the bad news. I found the new working paper disappointing, largely because it represents an intellectual backsliding from the innovative stance staked out by its predecessor and because of its failure to adopt what I would argue is a badly needed statistical foundation for the very methods whose cause it advances. Let me explain.

Chapter II of the report captures some the current discussions in the literature about the the definitions of disclosure, but it fails to build on Dalenius' statistical definition of disclosure that formed the foundation for the structure of the 1978 paper. As a consequence, we have descriptions of methodology

such as cell suppression which, while seemingly advanced, represent mathematics but not statistics. The techniques have been honed so that they can be implemented for large collections of cross-classifications utilizing linear programming and other techniques but we are never told, either by those who developed the approach or by the Subcommittee preparing this working paper, what statistical criteria the methods attempt to optimize and the extent to which they succeed. Thus we are told, for example, about the need to keep the values of n and p in the cell suppression rules confidential, but there is no recognition that statistical learning by those outside the agency might easily make such a statement essentially moot. Similarly, in the discussion of three-way and multiway cross-classifications, there is no recognition of relevant statistical methodology that might inform the very methods under discussion such as the probabilistic theory for Fréchet bounds on cell values (e.g., see Kwerel, 1983). When we get to the discussion of research issues relating to cell suppression, we find more of the same: advances in optimization of network flow methods, more elaborate computer programs, faster software. Where is the statistics in statistical disclosure limitation methodology? Where is the recognition that the data collected by statistical agencies is not error free? I contend that this very measurement error ultimately drives the statistical properties of attempts to compromise otherwise confidential data and disclosure limitation methodology to counter such attempts.

Nancy Kirkendall presented an example of an application of cell suppression which produces through complementary suppressions the following table (in which S stands for a suppressed cell and x a released cell):

$$\begin{array}{ccccc} S_1 & x & x & S_2 & S_3 \\ S_4 & x & x & S_5 & x \\ x & S_6 & x & x & S_7 \\ x & S_8 & x & x & S_9 \end{array}$$

She uses this to illustrate the need for auditing tables prior to release since the cell with entry S_3 can be determined via the other cells. It is interesting to note that all of this is related to the theory of existence of maximum likelihood estimates under quasi-independence for two-way tables. (e.g. see Chapter 5 of Bishop, Fienberg, and Holland, 1975). That those developing methods in this area seem unaware of such links to the statistical literature serves to reinforce my point on the need to make statistical disclosure methods more statistical.

I have a similar reaction to the briefer materials described in the Working Paper on data swapping, especially as it was implemented in the 1990 decennial census. This method grew out of a novel notion suggested by Tore Dalenius, but there appears to be little recognition by those who implemented the approach regarding the effect that the method has had on the utility of the resulting data, for example, as it is to be used for enforcement of the

Voting Rights Act. I understand that considerable effort went into some of these considerations in advance, but we have little documentation and no post-censal evaluations.

The Working Paper also places what I believe to be a misguided emphasis on "population uniques." As various authors have noted, uniqueness in the population is a necessary but not sufficient condition for identity disclosure, and there is no reason to believe that identity disclosure necessarily compromises confidentiality guarantees. My example of the identity release in the biblical census I believe makes this point well. The Working Paper relegates the more interesting and more important statistical problems of inferential disclosure and measuring disclosure risk to the research agenda.

Finally, the report tries to make a clear demarcation between methods for microdata and methods for tabulations. What it fails to recognize is that many examples of tabulations are in fact restricted microdata. For example, tables of counts are microdata in which either the original variables are categorical or are continuous but have been disguised through the use of conversion through categories, and where the data have been truncated by the dropping of variables. Surely there should be some linkage between the methods for microdata and for tabulations. This is less a criticism of Working Paper 22 than it is of the state of the art of research on disclosure limitation. (See the related remarks in Fienberg, 1994).

There are interesting statistical ideas and proposals for a unified theory

of disclosure control in the research literature, such as those captured by the papers by Fuller, Lambert, Little, and Rubin in the recent special issue of JOS on privacy and confidentiality, but these are not given appropriate coverage in the Working Paper nor are they reflected in agency thinking. Perhaps this simply reflects the lag between research and practical implementation. Despite such shortcomings, Working Paper 22 is an excellent summary both of current methods and practices in the agencies. The Subcommittee should be applauded for its efforts.

4 Restricted Access or Expanded Access?

George Duncan's second major topic was the NRC/SSRC panel's recommendations on administrative procedures to protect confidentiality. The panel has emphasized the role of interagency data sharing as well as technological aids that facilitate such access. While the need for such restricted access clearly will continue, I believe that the future will be one of expanded rather than restricted access. Working Paper 22 is especially helpful in this regard. Chapter V on "Methods for Public-use Microdata Files" provides a concise primer on the developments in this area.

I've mentioned the role of technology in restricted access, but technology is even more important when we come to expand access. A number of federal statistical agencies are playing leadership roles in this regard. Nancy Kirkendall referred to the innovative approach being explored by the National

Center for Educational Statistics, but there are many other examples. For example, micro-data from the 1990 decennial census are currently available over the Internet via the Consortium for International Earth Science Information Network (CIESIN) in Michigan. Further, the Bureau of the Census has created SIPP-On-Call, a new interactive approach to allow access to files from the Survey of Income and Program Participation over the Internet. Special user-friendly access is available via Gopher or NSF's Mosaic. Even *Wired* magazine, in its June 1994 issue, describes such access to its readers and points out that one also has on-line access to the Privacy Act and Title 13 as hypertext documents!

The new world of immediate user and intruder access over the "information highway" will place greater demands on released microdata and it will test, in new ways, the appropriateness of disclosure limitation methods both for the preservation of confidentiality and for the increased utility of the released data. This, I predict, will be a major topic for the next Federal Committee Subcommittee effort in this area and I expect that new statistical approaches to disclosure limitation will accompany these emerging changes.

5 Summary

There is much meat for statistical thought in *Private Lives and Public Policies*, the report of the NRC/SSRC panel, and in both the original Working Paper No. 2 and the recently released Working Paper No. 22, *Statistical Disclosure Limitation Methodology*, produced under the sponsorship of the

Federal Committee on Statistical Methodology. I fully expect that the next COPAFS-sponsored seminar on new statistical methodology, will highlight new advances in this area that build on the substantial contributions to date, that will also better link to statistical ideas, and that will report on the enhanced utility of released data resulting from these new developments.

6 References

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. N. (1975). *Discrete Multivariate Analysis: Theory and Practice* MIT Press: Cambridge, MA.

Fienberg, S. E. (1994). Conflicts between the needs for access to statistical information and demands for confidentiality. *Journal of Official Statistics*, **10**, in press.

Kwerel, S. M. (1983). Fréchet bounds. In *Encyclopedia of Statistical Sciences*. Vol. 3, (S. Kotz and N. L. Johnson, eds.), Wiley: New York, 202-209.

Tore Dalenius
DISCUSSION

Introduction

Before around 1970, the main direction of the methodological development was on the development of survey designs enhancing the efficiency, i.e. increasing the amount of information provided by a survey by other means than increasing the size of the survey.

Around 1970, a decisive change may be observed. The attention of the survey statisticians was now gradually directed towards how to recognize and hopefully address the problem of invasion of privacy. To address that problem, it proved necessary to apply methods which in fact served to reduce the amount of information made available. The subject of this meeting – Disclosure Limitation Methodology – reflects the above-mentioned change of methodological direction.

Dr. Duncan's presentation is based on ch. 6 of the book "Private Lives and Public Policies". This chapter considers two main options for protection of the confidentiality of released data: providing 'restricted data' and providing 'restricted access'. Dr. Kirkendall's presentation is based on the report Statistical Disclosure Limitation Methodology, prepared by a subcommittee of the Federal Committee on Statistical Methodology. This report is limited to disclosure limitation by means of 'restricted data'. Obviously, both documents are *final* products, a fact of relevance for the shaping of my discussion.

In what follows, I will first discuss selected aspects of restricted data and restricted access, respectively, to be followed by brief accounts of some additional aspects.

SELECTED ASPECTS OF RESTRICTED DATA

1. Two Classes of Data

Dr. Duncan and Dr. Kirkendall discuss in some detail two classes of data, viz. tabular data (frequency data and magnitude data), and microdata.

2. Frequency Data

The data to be restricted are represented by a table $T(N)$ with $R \times C$ cells. The restriction is achieved by a two-step procedure:

i. the sensitive cells of $T(N)$, if any, are identified by subjecting the table to a threshold rule: cells with a small number of data subjects (such as $n = 3$) are considered sensitive;

ii. next, some cells are combined, suppressed or rounded.

3. Magnitude Data

Typically these data are non-demographic, such as income or sales, accounted for by a table $T(X)$ with $R \times C$ cells. The variable X has in most cases a skew distribution: a small number of data subjects may account for a large proportion of the cell values. These cells may accordingly be sensitive, i.e. make it possible to link the cells with the data subjects accounted for, that is, to identify the data subjects. Hence, some kind of a restriction has to be applied to these cells.

The restriction of the data is achieved by:

i. first identifying cells to which a small number of data subjects contribute a large percentage of the cell value - this may be done by using the p percent rule, the pq rule, or the (n, k) rule, also called the 'dominance rule';

ii. next, these cells are subjected to restrictions, such as top-coding.

4. Microdata

Most releases of microdata are made up by a set of records with data about *individuals*. Only in exceptional cases do the data refer to business establishments.

Before the records can be released, formal identifiers must be removed ('deidentified'). But it may still be possible to link a record with a data subject: unique combinations of data concerning some attributes may serve as 'quasi-identifiers'. Hence additional restrictions are necessary, such as:

i. sampling;

ii. excluding data for one or more variables;

iii. representing the data by broad classes; age may for example be represented by an interval (age class);

iv. releasing data only for large populations; and

v. confidentiality edit of the data,

to give but five examples.

SELECTED ASPECTS OF RESTRICTED ACCESS

5. A Wide Class of Procedures

Dr. Duncan includes in this class several disclosure limitation approaches. Common to them is that the statistical agency establishes eligibility requirements for the data users who are to be included in the group of users given access. I will briefly consider four procedures.

6. Interagency Data Sharing

This term is used to denote two cases:

i. transfer of *administrative* data from a government agency to the statistical agency; and

ii. transfer of *statistical* data from a government agency to the statistical agency.

7. Swearing In of Users

Formally, this kind of restricted access means that potential users are given status of employees of the statistical agency concerned, either at the main office, or at some local office near the place where the potential users live.

Clearly, the statistical office will have an opportunity of critically assessing the users' research projects and also the merits of the users.

8. Site Inspection

Assume that there is a government agency with authority to inspect how a statistical agency performs with respect to protection of the confidentiality of the data to be released. Then this "control agency" may implement a scheme for inspection of the performance of the statistical agency.

The scheme may call for inspection every kth month. A better scheme would, however, call for inspection at dates chosen at random. This would make it impossible for the statistical agency to perform well during an inspection but not between inspections.

SELECTED MISSING TOPICS

9. The Coverage of the Two Presentations

It goes without saying that it is possible to identify topics which have not been presented, or possibly only touched upon. I will provide three such examples.

10. Example No. 1 – Schemes for Rounding

Rounding the counts in a table may be carried out in several ways. The main ways are related to:

i. the choice of a base different from the standard $b = 5$;

ii. the simultaneous use of more than one base, especially if the table is large (many rows and columns);

iii. rounding all cells in the table rather than a subset of cells; this type of scheme has in fact been proposed for use in the British population census; and

iv. the use of deterministic rather than random rounding.

11. Example No. 2 – The Multi-Table Problem

Let T_1 be a table with no disclosure. And let T_2 be another similar table. Release of both T_1 and T_2 is not necessarily safe. Access to both tables may make it possible to derive a combined table T_3 which is disclosing.

12. Example No. 3 – Release by a Database

The statistical agencies should develop schemes for releasing statistics by means of a database. There is no reason to 'wait and see' what comes out with respect to a data superhighway.

TOPICS FOR RESEARCH AND DEVELOPMENT

13. Terminology

There is as yet no generally agreed upon terminology in the area under consideration here. It suffices to mention the following facts:

i. privacy is defined in a great many different ways;

ii. confidentiality is sometimes viewed as 'anonymity'; and

iii. what in the two presentations is called 'disclosure limitation' was called 'disclosure avoidance' in the 1978 report; an alternative term is 'disclosure control', which I prefer.

It is indeed high time to develop a standard terminology.

14. A Catalogue of Potential Research Topics

In the report from the subcommittee there are some suggestions about research topics. But additional topics are needed. I will suggest one topic, viz. design of microdata about business establishments.

15. Inventory and Analysis of Sensitive Topics

In the last two decades, the non-response rate in surveys has shown a tendency to grow, possibly reflecting an increasing unwillingness to answer questions about sensitive topics.

In my view, the survey statisticians should process surveys already carried out and generate an *inventory* of sensitive topics which may explain the development. Such an inventory would be useful in the design of future surveys, by drawing the statisticians' attention to the need for special measures (such as special measurement methods) to improve the rate of cooperation.

The inventory should be *analyzed* to identify groups of data subjects with very large non-response rates. Such groups may then be singled out for special action.

CONCLUDING REMARKS

By way of presenting a summary of my views about the two presentations, I want to say that I have found them very informative and helpful. Dr. Duncan and Dr. Kirkendall are to be congratulated to the contributions they and their cooperators have made.

Session 3

CUSTOMER SURVEYS

Quality Management for Customer Satisfaction Surveys

Richard M. Devens, Jr.
Executive Editor
Monthly Labor Review
U.S. Bureau of Labor Statistics

Introduction.

This is the story of a customer satisfaction survey done for the Employment and Unemployment Statistics Quality Council at the Bureau of Labor Statistics. This is was the program's first customer satisfaction survey, and we are still learning from it. What I hope to pass on in this paper are the lessons learned about serving the "other customers," the executives that sponsored the survey and the front-line staff at the survey's focus. In other words, how "fit for use" was the National Survey of Users of Employment and Unemployment Statistics? Before plunging into that, I will take a few minutes to explore the importance of customer surveys and to outline the technical process of designing and conducting this one.

The reason for conducting any customer satisfaction survey is the position customers hold in the guiding principles of total quality management (TQM):

- customer focus,
- employee involvement,
- continuous improvement.

Customer focus, in my mind, is both the most important of these principles and the most difficult to persuade many public-sector managers to accept.

Continuous improvement is normally accepted straight off, usually with the assertion that the organization is already practicing it. The TQM purist might quibble that managers often mean their organization is always on the lookout for the big breakthrough, rather than practicing Deming's Fifth Point. [Improve constantly and forever every process for planning, production, and service.] My own observation has been that managers really do want to improve their operations, one way or another.

Employee involvement is a bit harder to sell. Many executives are used to and, quite frankly, happy with a command-and-control structure. In the case of the statistical agencies, such organizations were tremendously successful at organizing the armies of data collectors, mail room clerks, document controllers, coders, key entry workers, data reviewers, statistical assistants, statistical typists, junior economists, computer operators, computer programmers, research assistants, supervisory statisticians, senior economists, printers, and Assistant Commissioners that it took, and still takes to a fair extent, to produce a few tables of accurate, timely, relevant numbers.

Whether these hierarchies will work as well when data collection becomes automated, databases are connected through electronic data interchange (EDI), and performance becomes more dependent on the commitment of highly-skilled, self-confident, and very independent professionals is the issue. I believe such developments will lead organizations to embrace employee involvement models sooner rather than later--and most executives realize it, however grudgingly.

Customer focus, in contrast, is a very difficult concept for public-sector managers to accept at all, let alone embrace. The first reaction is, "We don't sell anything, so we don't have customers." Even after getting over this "filthy lucre" barrier, there is, especially in "craft" or "engineering" cultures such as those of the statistical bureaus, a deep skepticism about the fitness of the customer to make rational decisions or even to know what they want. These reactions are evident deep down into the structure of such agencies. Where the first-line will quickly accept the notion of getting involved in and taking greater responsibility for technical improvement, there is little enthusiasm for treating their work as a customer-satisfying process, not an estimates- or analysis-producing process.

The upshot of all this for the manager of a customer satisfaction survey is that there are two other--and perhaps more difficult --customers that must be considered in parallel with the external customer: the executive-level sponsors and the front-line staff. The rest of this paper overviews the National Survey of Users of Employment and Unemployment Statistics and its findings, the interaction of the project with its sponsors, the interaction of the project with the front-line staff, and the reactions of these "other" customers to the survey.

Outline of the National Survey of Users of Employment and Unemployment Statistics

The National Survey of Users of Employment and Unemployment Statistics is based on the premise that customer satisfaction is measured by the discrepancy between the client's needs and expectations and the client's perception of our performance. In the marketing literature, this is known as "disconfirmation" theory. The survey measures expectations and performance in five broad factors:

- Data quality: The accuracy, relevance, and timeliness of our statistics.
- Tangibles: The appearance and understandability of our materials.
- Dependability: Our demonstrated ability to perform promised services reliably, correctly, and promptly.
- Assurance: The knowledge of our employees and their ability to convey trust and confidence.
- Empathy: The caring, courteous, individualized attention we provide.

Each factor is represented by specific statements in the questionnaire. (See box.) The questionnaire also provides for an independent ranking of the importance of the factors and for general evaluations of satisfaction with our statistics and associated services.

Quality Factors and Their Proxies
(Question number in parentheses)

Data Quality

(2) The demographic, geographic, and industrial coverage of the statistics is sufficient for my needs.

(7) The data provided meet my standards of accuracy and reliability.

(8) The data provided meet my standards of timeliness and currency.

Dependability

(1) Staff are always available during their normal working hours.

(4) My questions are answered promptly and dependably.

(5) It is easy to get in touch with someone who can answer my questions.

(14) The information I ask for is sent in the medium and format requested.

Tangibles

(6) Materials provided make sense and can be understood without additional information.

(14) The information I ask for is sent in the medium and format requested.

Assurance

(9) Staff are knowledgeable and competent.

(11) Staff can clearly explain conceptual and analytical issues without using overly technical language.

(12) Staff can clearly explain the technical limitations of the data.

Empathy

(3) Staff make me feel that I can call back for additional clarification or data.

(4) My questions are answered promptly and dependably.

(10) Staff are courteous.

(13) Staff go out of the way to understand and fulfill my requests.

(15) Apologies are rendered for inconveniences such as delay or misunderstanding of my needs.

Clients rated the statements on the quality of our statistics and services on 5-point scales for their expectations of quality and their perception of our performance. The expectation score is subtracted from the performance score to yield the "performance gap" for any specific statement. The performance gap for a factor is the mean gap for the set of statements that represent it.

In addition to the customer satisfaction scales, the survey asks how clients use our data, which programs they have utilized, and what channels of distribution were used to access data. We also provided space for comments.

Designing the Survey

We developed this user-friendly questionnaire using cognitive research methods including focus groups, think-aloud interviews, and a pilot test. Each of these methods identified errors and we were able to take corrective action before taking the final survey into the field.

In the field, Dillman's Total Design Method was followed closely, with the exception of experimental variations in the third and final follow-ups. Clients selected for the survey received several mailings:

- A notice arrived at the customer's address a few days before the primary questionnaire package.
- A thank-you/reminder letter followed the questionnaire by about a week.
- A second package went out two weeks after the "tickler."
- Final prompting, experimentally split between certified mail and telephone prompts, began 2 weeks after that.

This intensive data collection methodology yielded a usable response rate of 87.8 percent.

Two minor modifications to the Dillman method were necessary. First, the front cover of the questionnaire was not illustrated with graphics because of the limited space, and the stationery size was the ordinary 8 1/2 by 11. Second, the reminder/thank-you postcard was replaced by a reminder/thank-you letter because in-house constraints allowed letter production only.

The experimental exercise conducted in the third follow-up tested certain refinements to the Total Design Method for use in the establishment setting. Two weeks after the second follow-up, each of the remaining nonrespondents was randomly assigned to either certified mail follow-up or telephone prompting. In the control group, "holdouts" received the third follow-up packet by certified mail containing a replacement questionnaire, a business reply envelope, and a cover letter. The wording of the cover letter was different from the cover letters used in the preceding follow-ups; we softened and relaxed the wording but emphasized explanations of why this additional follow-up is important and is sent by certified-mail.

In the treatment group, nonrespondents were contacted by trained, experienced telephone prompters. We prepared a survey-specific training agenda, drawing on insights from nonresponse

conversion efforts in telephone follow-up surveys. The training included practice of scripted telephone procedures including appropriate reactions to specific reasons for refusal, discussion of persuasive techniques, and use of call record sheets. Approaches to locate the sample subject and find the best time to call back were also included in the training.

The survey's sample frame was constructed from two sources. First, client contact staff in the National program offices, the Regional Offices, and the Inquiries and Correspondence section logged contacts during September-November 1992. Program managers and senior executives provided separate lists of "regular" clients--persons maintaining on-going professional contact with our programs.

The lists were merged and duplicate entries removed. The resulting sample frame contained 3553 names which were stratified based on the program office that was the point of contact. Two additional strata were formed: one for all of the regular clients and another for all of the customers who were logged in by more than one program. The total sample of 999 clients was obtained by selecting samples of approximately equal size from all strata except one. Members of the stratum of regular users were included in the sample with certainty.

The response rate figures from the national survey of users of Employment and Unemployment Statistics are shown in Table 1. After the second replacement questionnaire mailout, the overall response rate had already reached 75%, which is the average overall response rate for TDM-based surveys. The third and final follow-up boosted the response rate by 13 percentage points to approximately 88%.

Table 1. Response and Conversion Rates

Mailing	Conversion Rate (%)	Overall Rate (%)	N*
Prenotice (Day 1)			999
1st mailout (Day 8)	28.65	28.65	998
Reminder (Day 15)	50.87	65.24	978
2nd mailout (Day 29)	28.53	75.77	970
3rd follow-up (Day 43)			
Certified mail	48.21		
Phone Prompting	44.44		
Overall	46.45	87.68	950
Close-out (Day 71)			

* The sample size declined as ineligibles were uncovered through the data collection process.

Summary of Findings. Despite averaging 4.08 out of a possible 5 points on the performance scale of our survey, we did not fully meet our customers' expectations. (Expectations averaged 4.46 out of 5.)

Considering the major factors displayed on chart 1, our "performance gaps"--the average difference between our performance and our customers' expectations across the statements that represented the factors--were:

1.	Data quality	(-0.66)
2.	Dependability	(-0.52)
3.	Tangibles	(-0.34)
4.	Assurance	(-0.28)
5.	Empathy	(-0.17)

Using an expectations/performance grid --a "customer window" in the most recent jargon--to analyze individual statements shows specific areas to concentrate our efforts on. (See chart 2.) In this graphic display, the intersection of the axes represents the grand means for customers' expectations (Y-axis) and their perception of our performance (X-axis). The points plotted for each statement are the ordered pair of Z-scores. According to this analytical tool, the important places to "Concentrate" corrective strategies are:

1. More timely (#8) and detailed (#2) data
2. Making it easier to find someone to answer your questions (#5)
3. Providing clearer materials (#6)

The survey report expressed these in terms of three strategic themes for improvement:

- Get Faster: Make statistical products available to the public more quickly.
- Basic Service First Time/Full Service Every Time: Have analysts able to answer broader ranges of inquiries, rather than transfer customers across program lines.
- Clarity, Clarity, Clarity. Make products and services easier to understand from the customer's point of view.

Chart 1

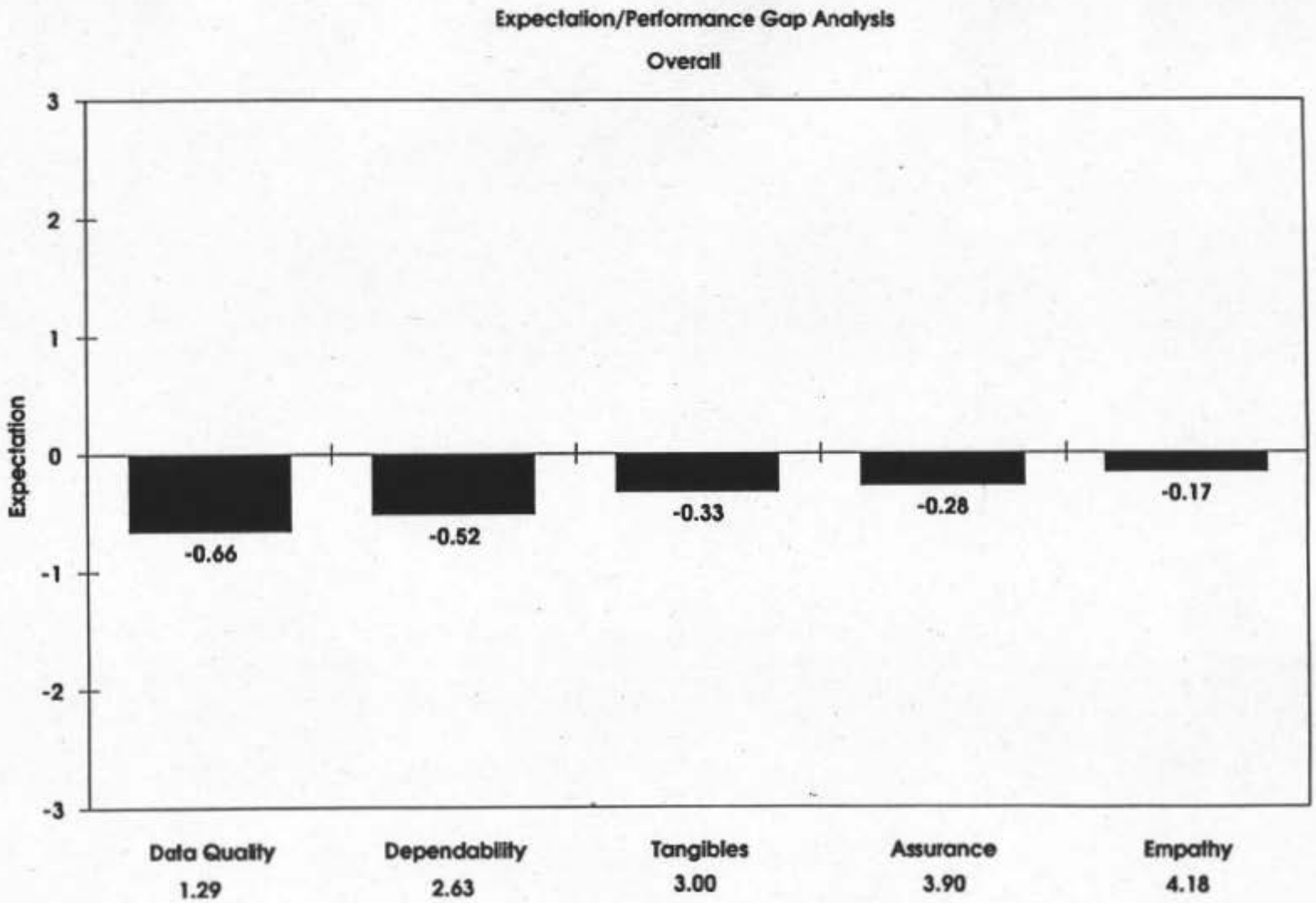
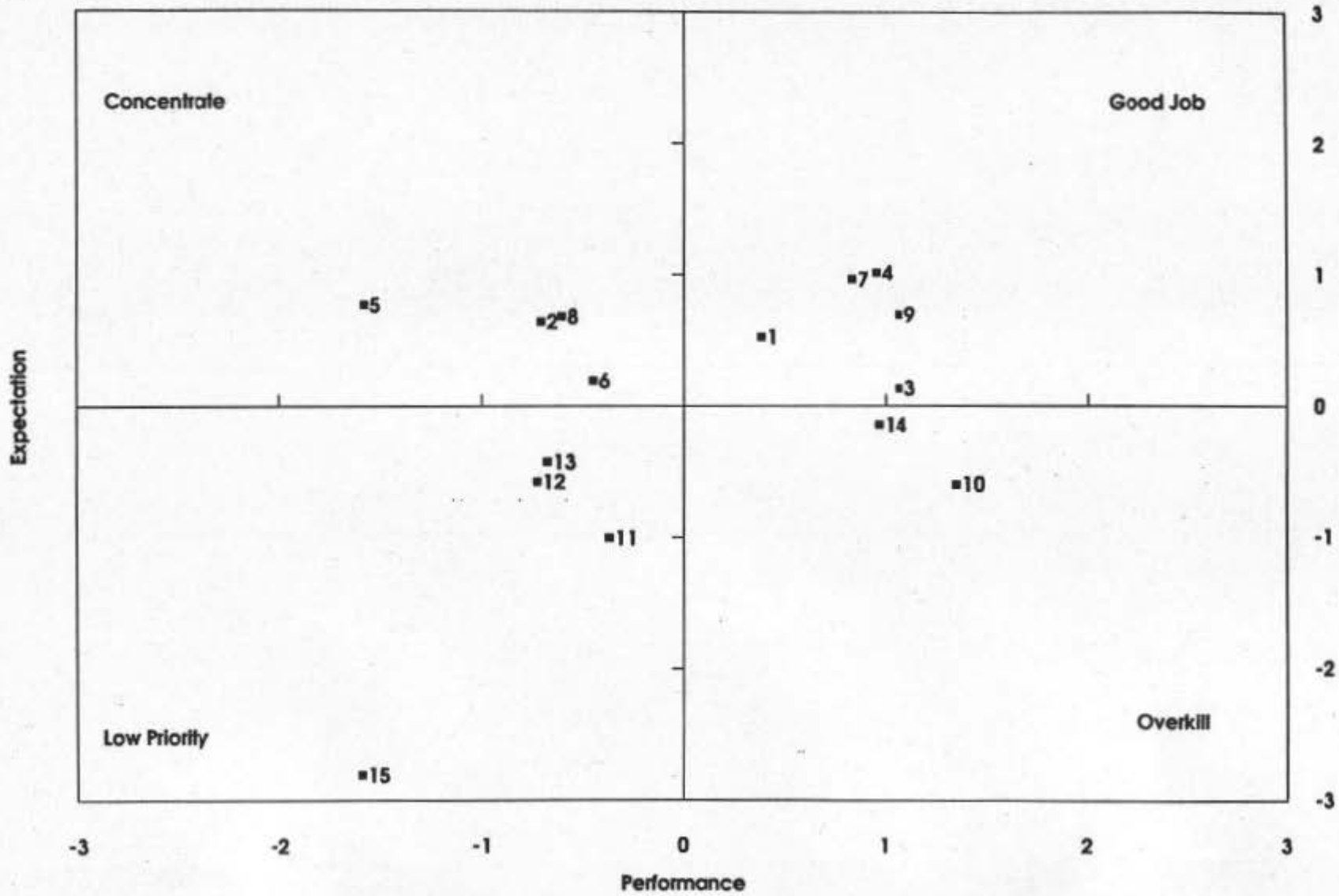

Expectation/Performance Gap Analysis
Overall
Expectation
3
2
1
0
-1
-2
-3
-0.66
-0.52
-0.33
-0.28
-0.17
Data Quality
1.29
Dependability
2.63
Tangibles
3.00
Assurance
3.90
Empathy
4.18

Chart 2

Expectation/Performance Comparison

Overall



Interaction with executive-level sponsors. The National Survey of Users of Employment and Unemployment Statistics is an "infrastructure" project on the part of the Employment and Unemployment Statistics (EUS) Quality Council. The Council itself is the executive-level coordinating body for the EUS Quality Improvement Program. The Council took its first look at conducting a customer survey at its August 1991 meeting. I was assigned the task of pulling together a few ideas on how such surveys were conducted and how they might apply to the EUS quality program. At their November 1991 meeting, the Council approved moving forward.

For the next several months, the minutes don't do the meetings justice. "..council members expressed concern," "..requested that a full proposal be prepared before proceeding," "...discussed the difficulties...." "After some discussion, it was agreed...." I am sure all of you are aware of what lurks behind these bland formulations.

After laboriously negotiating final approval, the survey team administered the instrument to the Quality Council itself, with the instruction, "Complete the questionnaire as if you were the 'average' customer." This exercise had the twin goals of further educating the Council on the survey and developing a baseline measure of the Council's awareness of customer needs and knowledge of the customers' perceptions of our performance.

As a baseline exercise, the Quality Council identified the same order of priority among the major quality factors that customers did. Interestingly enough, however, the absolute sizes of the perceived performance gaps were actually larger among Quality Council members than among customers. The primary source of the larger gaps, as shown below, was lower performance ratings by the Quality Council.

	Customers	**Council**
Average performance	4.08	3.80
Average expectation	4.46	4.36

How this exercise worked as an educational tool is a good question. My subjective evaluation is that the Council members themselves perceived it fairly narrowly in its baseline setting role, and would be surprised to find out about its covert objective of preparing them to more fully understand the survey's results.

When the final report of the project was drafted, the Quality Council was briefed on its contents and provided with copies for comment and approval. Once approval was obtained, final reports with more extensive technical documentation were published and circulated to the Council and the staff of the employment and unemployment statistics activity. The Business Research Advisory Council to the Bureau also expressed an interest in the survey and its results. Members of the Quality Council attended that briefing as well.

Interaction with customer-contact staff. Another set of customers for the survey and its results is the front line customer-contact staff. This is the group of our colleagues that provided roughly 3,500 customer names and addresses representing over 5,000 direct contacts over a 3-month period. These are also the people whose work product was put under scrutiny by the survey and the upon whom much of the burden of improvement would be likely to fall.

From the outset of the survey, five senior professionals from customer-contact units were assigned to the project. Their substantive contributions were critical and they also served as a "backchannel" of informal feedback between the staff and the survey team. That channel was, during the universe-building phase, our best means of helping the staff focus on keeping a complete log of contacts. (As a result of our debriefing of the representatives, one improvement we are likely to make in future surveys is a shorter log-in.) While the survey was in the field, the backchannel kept the staff informed about our progress.

Other interactions with the customer-contact staff included formal training sessions on the objectives, concepts, and methods of the survey and the procedures they would follow maintaining the universe log. In addition, the log procedures were documented on the forms themselves along with explanations of the purposes of the survey itself and of some of the most critical pieces of universe information--e.g., telephone numbers.

In general, interaction with this group of customers is something we should improve on. Some regional office information staffs had virtually no training or documentation of the survey or their role in it until what might be generously called the last minute. Interaction with the Inquiries and Correspondence Branch of the Office of Publications, while more timely, never reached the extent or intensity needed.

Reaction of sponsors. The most important measure of the success of a project such as the National Survey of Users of Employment and Unemployment Statistics is the action it prompts the organization to take. On this score the results are promising, but not overwhelming. Actions "moved onto higher priority time paths" to improve on the critical data quality factor include:

- Advancing the review and release of State-wide Local Area Unemployment Statistics (LAUS) data by 2-3 weeks (from a baseline of roughly 9 weeks after the reference period).

- Converting 20,000 late respondents to the Current Employment Statistics (CES) survey to automated self-reporting using an advanced touch-tone telephone data collection technology to improve the timeliness of these reports.

- Hosting the International Occupational Classification Conference to provide a forum for discussing new ideas and alternative approaches to the details of occupational categorization.

- Expanding service-sector detail for Current Employment Statistics by adding 108 new series to our most detailed publication and 20 seasonally-adjusted series to the employment news release.

In general, these have been projects that were on various burners to start with--the most the survey can claim is that some were completed more quickly and with more fanfare. There have been a few initiatives to start addressing some of the issues of service quality:

- Increasing the clarity of hard copy information sent to customers
- Resource book for information calls
- Developing new-employee training module for customer service
- Cross-program briefing on data availability.

My personal evaluation of the impact of the survey is that it was useful, but not nearly in proportion to the skills exercised or to the resources expended. That sense of disproportionality of effort leads me to the point of the paper--how well served are the "other" customers?

User-survey-users' surveys. To find out, I conducted a pair of informal surveys of the two groups of "other" customers. The survey of the Quality Council asked for their evaluation of the importance of the strategic directions the results pointed to, an evaluation of the communication processes between the Council and the project team, and an evaluation of the team's effectiveness at communicating the results. A similar survey was conducted among the front-line staff.

The results of the executive survey indicated that the group found that they rated the importance of the 3 strategic themes quite closely together between 5.3 and 5.8 on a scale of 1-to-7 (Not important at all to Extremely important). The highest score went to the Clarity, Clarity, Clarity theme.

The executives' evaluations of the effectiveness of our communication of the concepts and methods of the survey and of the results of the survey were devastatingly frank. On 5-point scales, the scores were 3.43 on effective communication of concepts (between "fairly" and "very" effective) and 3.57 on clear communication of results (again between "fairly" and "very"). These low scores, and remember they came from colleagues, probably reflect the reason the response to the survey was not overwhelming--the credibility of the product was not established and the results were not clearly communicated to the sponsors.

Reaction of front-line staff. The front-line's reaction to the survey is neatly summarized by the response rate to the survey included in the individual copies of the final report--almost 8 percent. Obviously, our efforts to engage this group fell short. The open question is why did it happen? I fear that the real reason was a fundamental failure to convince the front line that the customer satisfaction survey was serious. This may be the most significant quality issue for the National Survey of Users of Employment and Unemployment Statistics.

For what it is worth, the front-line survey found that among six respondents the importance ratings of the themes ranged from 4.7 to 5.7, with the clarity issue highest once again. The scores for effective, clear presentation of the concepts, methods and results of the customer survey--3.2 for effective presentation of objectives, concepts, and methods and 3.3 on clear communication of the results-- were even lower than those given by the sponsors.

Conclusions. Customer satisfaction is the "outcome" of any statistical or information service. This must often be measured quite separately from the "output" of programs. Output measures too often tell more about what is important to us than what is important to the customer. We in the Federal statistical community have always been concerned about hard measures of the output, "accurate data". We have only just now become aware that the soft outcome, "satisfaction with promptness and dependability of service," is perhaps even more important. That is why customer satisfaction surveys are useful--they are tools to measure and manage that outcome.

To be taken seriously as management tools, however, customer satisfaction surveys must be credible to the "other customers"--the people who should respond to the results. My point is that to obtain that outcome, the customer survey manager must establish credibility in advance and not think that good output--a clever report based on sound data--will suffice.

Note: All material in this article is solely the reponsibility of the author. The views expressed here do not necessarily reflect the policy of the Bureau of Labor Statistics or the views of other BLS staff members.

COMPARABILITY IN CUSTOMER SATISFACTION SURVEYS:

PRODUCTS, SERVICES, AND GOVERNMENT AGENCIES

by

Michael D. Johnson

National Quality Research Center
The University of Michigan
School of Business Administration
Ann Arbor, MI 48109-1234

Phone: 313-764-1259
Fax: 313-763-9768

April, 1994

To appear in *New Directions in Statistical Methodology*, Washington, DC: Office of Management and Budget, forthcoming. Do not quote or reproduce without permission. The author thanks the Swedish Post Office, the National Quality Research Center at the University of Michigan's School of Business Administration, and the International Center for Studies of Quality and Productivity at the Stockholm School of Economics for providing the data described in the paper.

COMPARABILITY IN CUSTOMER SATISFACTION SURVEYS: PRODUCTS, SERVICES, AND GOVERNMENT AGENCIES

ABSTRACT

This paper describes recent advances in customer satisfaction surveys and their implications for government agencies. Many agencies are in the process of implementing customer satisfaction monitoring systems and identifying appropriate private sector benchmarks. Satisfaction models and survey methods currently being used to produce national customer satisfaction indices are described. These efforts illustrate a number of important steps that should help government agencies produce meaningful measures of satisfaction and identify private sector industries that provide realistic agency benchmarks.

INTRODUCTION

Customer satisfaction has emerged as an important benchmark for gauging the performance of various economic agents over the past decade. Manufacturers of durable and nondurable products, retailers, service providers, utilities, and government agencies alike have implemented, or are in the process of implementing, customer satisfaction measurement systems. At a micro-level, these systems monitor a firm's or agency's primary asset - their customers - and provide important diagnostic information needed to improve or maintain satisfaction. At a macro-level, Sweden and Germany have implemented national customer satisfaction indices to monitor the major sectors of their economies while the United States, Taiwan, and New Zealand are in the process of doing the same.

There is disagreement, however, among psychologists, economists, consumer researchers, public policy makers, and others regarding the merits of comparing satisfaction across individual and industries. The ever broadening arena of customer satisfaction, in conjunction with recent advances in how satisfaction is surveyed and operationalized, shed light on this long-standing debate. The goal of this paper is to describe these developments and discuss their implications for government agencies. Many agencies now find themselves, for the first time, asking such questions as, "who are the customers served by our agency, who should be our customers, what standards should we use, and what comparable businesses should we benchmark on?"[1] Recent advances in satisfaction survey methods provide important insights into how government agencies should survey their customers. Sweden's experience with a national Customer

---

[1] Section 1 of President William J. Clinton's executive order, dated September 11, 1993, begins as follows:

> In order to carry out the principles of the National Performance Review, the Federal Government must be customer-driven. The standard of quality for services provided to the public shall be: Customer service equal to the best in business. For the purposes of this order, "customer" shall mean an individual or entity who is directly served by a department or agency. "Best in business" shall mean the highest quality of service delivered to customers by private organizations providing a comparable or analogous service.

Satisfaction Barometer (the SCSB) also illustrates which private sectors businesses provide benchmarks or standards of comparison for these agencies. First, however, the nature of the debate over the comparability of satisfaction and the issues involved are described.

SATISFACTION AND THE HAPPY SLAVE PROBLEM

Customer satisfaction is a customer's evaluation of their overall experience with a product or service to date (Johnson and Fornell 1991; Johnson et al. 1994). This definition of satisfaction is consistent with existing views in economic psychology, where satisfaction is often equated with notions of subjective well-being (Van Raaij 1981), and economics, where satisfaction is equated with post-purchase consumption utility (Meeks 1984). Because it describes the customer's total consumption experience, satisfaction predicts customer loyalty and a firm's subsequent "profitability." In the private sector this "profit" is bottom line return on assets (Anderson et al. 1994). For government agencies, the benefits of increased customer satisfaction range from budget considerations, to more efficient use of taxpayer dollars, to the creation of a more positive image, to compliance (e.g., for the Internal Revenue Service).

There is a long standing debate in economics over the comparability of satisfaction across individuals and industries (see Hammond 1991 for a review and extensive bibliography). Bentham (1802) defended the comparability of satisfaction as both possible and necessary from a public policy standpoint, though not without error. Subsequent economic theorists sought to eradicate satisfaction measurement and comparisons as value laden and unnecessary (Hicks 1939; Robbins 1938). Recently, satisfaction has again emerged as a basis for making meaningful comparisons across people and products. Virtually all policy recommendations require comparisons of welfare which is proof enough that they are possible (Scitovsky 1951). The important question has become how comparisons of satisfaction or well-being are and should be made (Hammond 1991; Jorgenson 1990; Sen 1979; Simon 1974; Tinbergen 1991).

Our interest is specifically with customer satisfaction. Economic theorists are more often concerned with comparisons of more global economic well-being, which includes not only customer satisfaction but job satisfaction and income evaluation (Poiesz and Grumbkow 1988). Broad based comparisons of customer satisfaction are not exactly new. Andreasen and Best (1977) report meaningful comparisons of customer satisfaction and complaint behavior across a variety of product and service categories, while Pfaff (Lingoes and Pfaff 1972; Pfaff 1977) has used subjective measures of satisfaction to construct an index for the purpose of comparing various food product categories. Wikström (1983) has even compared subjective levels of customer satisfaction across countries (Sweden and the U.S.) and argued that the observed differences can be traced to underlying differences in market performance between the two countries.

Yet some policy researchers have concluded that subjective measures of customer satisfaction are incapable of revealing any meaningful differences (Hunt 1988; Ölander 1988). Most notably, Ölander (1977a, 1977b, 1988) argues persuasively that subjective measures of customer satisfaction are fraught with problems. Foremost among these is the so-called "happy slave" problem. Because customers adapt to the levels of product and service performance available to them, no meaningful differences in satisfaction should emerge. Individual differences in the degree of adaptation within and across industries further compounds the problem. Other problems include the notion that customers may have different yardsticks by which they judge satisfaction. Even if they used the same standards, consumers may have very different product or service alternatives available to them, and/or differ in their knowledge of these alternatives. Finally, customers may fail to express true dissatisfaction or strategically express false dissatisfaction in hopes of receiving some retribution.

At some level Ölander's concerns are very real, as when one might compare satisfaction between customers in relatively wealthy and impoverished countries. At the same time, several considerations suggest that the concerns expressed over comparing

subjective measures of customer satisfaction are overstated. As Pfaff (1977) argues, who is in the best position to evaluate customer satisfaction but customers themselves? People are more similar than different, and these similarities are growing in an increasingly "global" economy. We live in an information age in which consumers from different countries and socio-economic strata of our society are increasingly aware of alternative products and services available in the marketplace.

Equally if not more important are recent advances in satisfaction survey methods and modeling which facilitate our ability to compare subjective measures of satisfaction. Sweden's Customer Satisfaction Barometer (SCSB) embodies these advances and is serving as the prototype for the American Customer Satisfaction Index (ACSI).

THE SWEDISH CUSTOMER SATISFACTION BAROMETER

Established in 1989, the SCSB was the first national customer satisfaction index for domestically purchased and consumed products and services (Fornell 1992). The index is constructed using survey measures obtained from representative customers in each of 32 major Swedish industries which themselves represent approximately 70% of Sweden's gross domestic product (GDP). Those companies that account for approximately 70% of combined industry sales are selected to represent each industry. In cases where a company sells multiple products or services, the "flagship" brand (the product or service with the highest sales in kronor) is chosen to represent the company. For example, Saab Scania is represented by the sales of its 9000 series automobiles while banks are represented by their money lending activities. Each year approximately 100,000 customers are contacted by telephone and screened to obtain a sample that has experience with the products and services in the index. The number of customers who pass the experience screen and agree to participate is approximately 25,000 each year. Industry level sample sizes range from about 250 to over 4000 depending on the number of competitors.

Comparability in the Model

The first step in assuring comparability in the SCSB involves the choice of satisfaction related constructs and how they are modeled. Survey respondents are asked a variety of questions to operationalize five key constructs: (1) customer perceptions of product or service performance, (2) their expectations regarding performance, (3) customer satisfaction, (4) whether they have complained ("voice"), and (5) customer loyalty. The SCSB model, which is presented in Figure 1, posits six relationships among these variables. These relationships are relatively universal in that they cut across all of the products and services in the barometer and are described briefly here (for more extensive descriptions and discussion see Fornell 1992).

- insert Figure 1 here -

Satisfaction is posited to be a function of two antecedent variables, perceived performance or quality and customer's expectations regarding performance (Fornell and Johnson 1993; Johnson and Fornell 1991). Customer satisfaction should increase with the degree to which a product or service provides net benefits that customers value (i.e., perceived performance). Because expectations embody past quality or performance information, they too should positively affect satisfaction. Expectations serve to anchor overall evaluations of satisfaction in the vicinity of the expectations (Oliver 1977, 1980). The size of this anchoring effect depends on the relative strength of the expectations versus performance information (Johnson, Nader, and Fornell 1994). As experienced customers can predict, to some degree, what levels of performance they will receive, expectations should also show a positive relationship to perceived performance.

There are two primary behavioral consequences of satisfaction. Increased customer satisfaction should reduce the incidence of customer voice or complaining behavior. Satisfied customers are also loyal customers, which is the key to the satisfaction-profitability linkage (Anderson, Fornell and Lehmann 1994). Finally, voice may increase loyalty. The size of this relationship reflects the degree to which customers

are allowed to voice their complaints and a firm's ability to address these complaints. That is, the relationship is positive when a firm can turn a complaining customer into a loyal customer. Overall the relationships in Figure 1 are well supported and appear to generalize across Swedish industries (Fornell 1992).

Comparability in Satisfaction Survey Items

The next step in assuring the comparability of satisfaction in the SCSB is to use a survey instrument whose questions are themselves universally applicable and help control for industry differences. This is quite different from what typically occurs in the context of a particular product category or industry where perceived performance is operationalized using customer ratings of a product or service on quality dimensions or attributes that are idiosyncratic to the industry (e.g., attributes of an automobile). In the SCSB, performance is operationalized using two measures of perceived value, the customer's perception of quality received relative to the price or prices paid (benefits relative to costs) and their perception of the price or prices paid relative to quality received (costs relative to benefits). Research demonstrates that this "value" is a common denominator that consumers use to compare even very dissimilar or "noncomparable" products and services (Johnson 1984, 1989). Using value perceptions to measure performance also controls for differences in income and budget constraints across respondents (Hauser and Shugan 1983; Lancaster 1971) which allows us to compare very high and very low priced products and services.

Satisfaction is also surveyed using comparable items. These include the customer's rating of overall satisfaction, how well the product performs relative to an ideal product or service in the industry, and whether performance fall short of or exceeds customer expectations. Theoretically, all three of these rating should reflect the underlying level of satisfaction independent of the particular product, firm, or industry involved (Johnson 1994; Johnson, Anderson and Fornell 1994). Customer voice is, meanwhile, measured in two ways: the incidence of formal complaints to company or

agency managers, and the incidence of informal complaints to personnel or service providers. As for performance and satisfaction, both measures are flexible and apply to a variety of organizations.

Finally, customer loyalty is measured using questions regarding repurchase likelihood and sensitivity to price increases. While very applicable to competitive product and service industries, these loyalty measures are more problematic in the case of government agencies and monopolies. The solution used in the SCSB is to make the questions hypothetical. That is, assuming some other organization could provide an agency's services, how likely would you be to use the agency again and how much more would the agency have to "charge you" before you would switch to the hypothetical competitor? Over time, these questions are becoming less hypothetical and more realistic as government agencies are being reinvented and subjected to increased competition and market pressures.

Comparability in Satisfaction Model Estimation

The third step in assuring the comparability of subjective satisfaction centers on just how the survey items described above are used to operationalize the constructs and estimate the relationships in Figure 1. An important aspect of the Swedish index is that satisfaction (as well as performance, voice, and loyalty) is operationalized as a latent variable within a system of equations. Johnson and Fornell (1991) argue that satisfaction, as a theoretical concept, is a common denominator on which very different people and products may be compared. As a latent theoretical construct, satisfaction is empirically measurable as a weighted average of multiple satisfaction indicators. As Ölander and others have argued, any individual rating or measure which uses a particular yardstick is at best an indirect proxy for satisfaction. Operationalizing satisfaction as the shared variance among a set of multiple satisfaction survey measures provides a more direct measure of latent satisfaction.

This latent variable is estimated with a system of equations, or causal model framework, using the SCSB model in Figure 1. The particular estimation method used to operationalize latent satisfaction and estimate the model is partial least squares or PLS (Fornell 1989; Lohmöller 1989; Wold 1982). PLS is an iterative estimation procedure that corrects for routine least-squares measurement problems and does not impose distributional assumptions on the data. This is particularly attractive in a satisfaction context where distributions are often highly skewed. As a result, PLS is better suited to causal model estimations involving small samples than is, for example, covariance structure analysis using LISREL. It also allows the researcher to operationalize latent variable scores and hence calculate an index value.

Another important feature of PLS is that it aims to explain variances at the observed (measurement) level while LISREL aims to account for observed covariances. In Figure 1, latent satisfaction should ultimately explain variance in loyalty across customers. PLS weighs the individual satisfaction survey items in the satisfaction index so as to maximize the index's ability to explain loyalty. This, in turn, provides a satisfaction index that is comparable in the following sense. In each industry, the satisfaction index explains an endogenous, dependent variable that is universally applicable across industries. The satisfaction index is itself explained by two antecedents that should affect satisfaction in a similar fashion across industries. If the satisfaction index behaves as it should behave according to the model in Figure 1, then its validity and value as a benchmarking source is supported.

Empirical Evidence of Comparability

The ultimate test of the "happy slave" problem and other questions raised regarding the comparability of satisfaction is an empirical one. A recent study by the author and Claes Fornell (Fornell and Johnson 1993) using the SCSB data explicitly examines this issue. In the study we argue that if one can explain differences in satisfaction across industries using some underlying difference in the industries, then the

observed differences are meaningful. Specifically, we argue that product or service differentiation in an industry is one logical basis for explaining differences in expectations, performance, and satisfaction across industries.

Differentiation, in this context, refers to the availability of predictably different options to customer. Differentiated industries offer predictably different options that more directly meet the needs of a heterogeneous population of customers. In contrast, undifferentiated industries offer "no choice." The automobile industry in Sweden is, for example, highly differentiated. Customers can choose among a wide variety of options and are confident in their ability to evaluate differences among them. Police, telecommunications, and public postal services are, in contrast, relatively undifferentiated due to the lack of variety from which to choose. At a more intermediate level of differentiation are banks and insurance companies, where alternatives exist yet customers have difficulty judging their differences. The study found that the level of differentiation across the industries explained fifty-percent of the variance in aggregate perceived industry performance. This performance, in turn, explained over half of the variance in aggregate industry customer satisfaction.

This study has important implications for government agencies who must now benchmark their customers' satisfaction to that observed in private sector industries. In the past, public utilities, monopolies, and government agencies had no competitors on which to benchmark satisfaction levels. Because the industry level differences in satisfaction are meaningful, national indices such as the SCSB provide these agencies and firms with useful benchmarks. The Satisfaction Index scores for the Swedish industries are presented in Table 1. Government owned industries include the pharmacies, local police services, business post, public post, railroads, business telecommunications, public telecommunications, and state sponsored TV broadcasting. To illustrate the differences among industry types, the industries in Table 1 were grouped into three classes: (1) products and product retailers, (2) services, and (3) government owned agencies and

businesses. The average satisfaction indices for each of these three groups from 1989 to 1993 are plotted in Figure 2.

- insert Table 1 and Figure 2 here -

The figure illustrates several interesting points. First, following Fornell (1992; Fornell and Johnson 1993), products and product retailers show systematically higher levels of satisfaction than do competitive services and government owned agencies and businesses. Both of the latter groups are service-oriented, which makes it inherently more difficult to meet specific customer needs. While products meet customer needs largely through their physical means of production, the production of services involves more of the human resources of the firm and customers themselves. This creates greater heterogeneity, on average, in the production of services versus products and lower average performance (Fornell and Johnson 1993; Zeithaml et al. 1988). In Figure 2, products and product retailers show the highest satisfaction and it stays relatively stable over time. Competitive services are below the products and retailers, which is consistent with the nature of service production. The drops in service satisfaction in 1992 and 1993 are due primarily to the recent poor performance in the banking sectors. Finally, the government owned agencies and businesses are generally lowest in satisfaction. This is due both to their service orientation, which makes it difficult to provide consistent quality, and monopoly positions, which limits customer choice.

More important from a benchmarking standpoint is the steady increase in satisfaction for the government sector over the five years in which the index has been in operation. Average satisfaction has increased from 54 to 61 (on a 0 to 100 scale) in this five years. Some of this increase is due to the addition of the high performing state pharmacies in 1990. Even without the pharmacies, however, there is a steady increase in this sector (from 54 to 59). Importantly, the differences between competitive and government owned services is decreasing over time. Following Fornell and Johnson (1993), this suggests that competitive services provide government agencies with a useful

benchmark for industry satisfaction that were not previously available. It appears that government agencies in Sweden are using these attainable benchmarks to improve performance. It would be more unrealistic to expect agencies, on the whole, to achieve the satisfaction levels that we observe for competitive products where the means of production is quite different. A second implication is that individual government owned or regulated businesses, such as the pharmacies, are capable of achieving even higher satisfaction levels. Overall, the SCSB results thus provide government agencies in Sweden with both attainable benchmarks and role models for setting satisfaction goals.

THE AMERICAN CUSTOMER SATISFACTION INDEX

The SCSB serves as the prototype for the American Customer Satisfaction Index (ACSI) which will be released for the first time in October of 1994. The ACSI is a quarterly, national index of customer satisfaction. Sponsored by the University of Michigan, the National Quality Research Center at the Michigan Business School, and the American Society for Quality Control, the index will, in its first year, survey approximately 50,000 customers of approximately 200 companies and government agencies which comprise about 49 percent of U.S. Gross Domestic Product.

There are important differences between the ACSI and the existing SCSB. First, the ACSI is larger in scope given the greater size of the U.S. economy. American firms are also more diverse in that a single firm is more likely to compete in multiple industry sectors. Sampling is, therefore, being done at the "firm level" rather than the "product or service" level. Finally, the ACSI (and future versions of the SCSB) include an expanded set of survey items. In addition to the original SCSB questions, customers will be asked both their expectations and perceptions of performance regarding two key quality components: (1) "fitness for use," or the degree to which a product or service provides those things that the customer personally requires from the product or service, and (2) "things gone wrong," or the degree to which a product or service is free from defects.

Both factors are germane to quality across all U.S. industries and will provide interesting bases for comparison.

Conclusions: Implications for Government Agencies

Our recent experience in the development of national customer satisfaction indices illustrates a number of important principals and concepts that should help government agencies as they actively implement satisfaction measurement systems. The first is that there is a relatively "universal" model of the antecedents and consequences of customer satisfaction. The challenge that agencies face is in translating the constructs in Figure 1 to the particular agency context. Customer loyalty, for example, may be "repurchase" in some agencies (e.g., the buying of Census Bureau data) and "compliance" in others (e.g., with an IRS regulation or rule). A second implication is that there are universal ways of asking the survey questions needed to operationalize such things as performance, satisfaction, and loyalty. This involves a focus on common denominators, such as "value" when operationalizing perceived performance, and using multiple standards of comparison, as when measuring satisfaction.

Once a flexible model and a set of survey measures are in place, the measures should be used to develop indices of the key constructs. This is especially true for satisfaction where any single survey item is at best a proxy for a customer's overall evaluation of their experience with a firm or agency. Ideally, the satisfaction index should be estimated within the context of a model (e.g., Figure 1) where, for example, performance and expectations explain satisfaction and satisfaction, in turn, explains customer voice and loyalty. These steps address many of the criticisms raised by consumer and policy researchers such as Ölander and Hunt toward the use of subjective measures of satisfaction. They help assure comparability in satisfaction measures across people and industries.

However, the ultimate test of this comparability is an empirical one. As the Swedish experience shows, customer satisfaction is empirically comparable. When

customer satisfaction is properly surveyed, measured and modeled, it allows one to compare "apples and oranges." The resulting comparisons provide useful benchmarks for government agencies as they improve quality. Competitive service industries provide a very straightforward benchmark that, based on the Swedish experience, appears attainable for government agencies as a whole. A national index also allows one to identify a particular agency or agencies to serve as role models and provide even higher satisfaction goals.

At another level, having established the comparability of satisfaction surveys, government agencies can use satisfaction index results to make better decisions and resource allocations. Existing productivity measures and price indices are limited in the way they account for quality changes (National Economic Research Associates 1991). Resources could be allocated more effectively by targeting industries or agencies that rate particularly low on satisfaction to help improve overall consumer welfare. For example, if the IRS rates particularly low on satisfaction, allocating resources toward improving customer satisfaction should more than pay for itself in terms of increased efficiency, compliance and resulting revenue generation. Finally, agencies will benefit by having a more complete picture of their organizations. Any comprehensive strategic plan for a public or private organization must integrate the organization's goals for achieving customer, employee, and owner satisfaction. In government agencies, taxpayers are the ultimate "owners." As agencies strive to meet customer needs, build customer loyalty, and save taxpayer dollars, these owners are the ultimate winners.

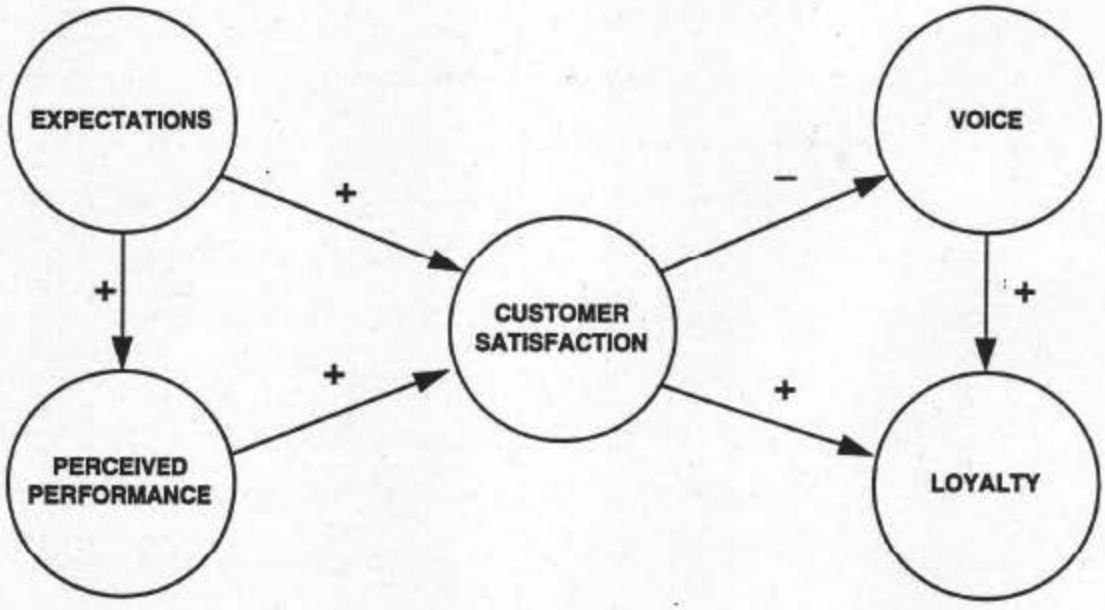


Figure 1. The SCSB Model

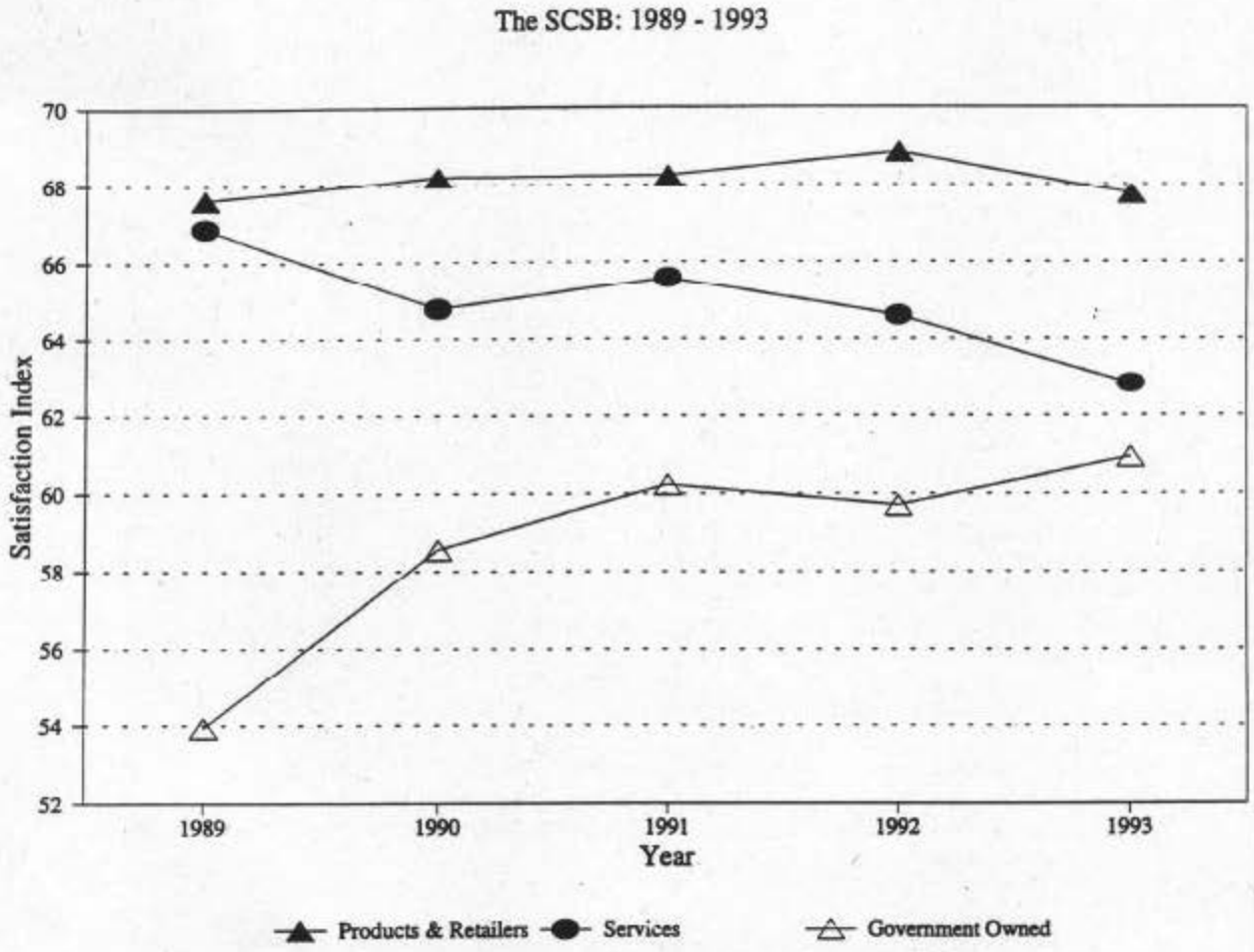


Figure 2. Year-to-Year Changes in the SCSB

Industry	Satisfaction by Year*				
	1989	1990	1991	1992	1993
Airlines	67	67	68	63	65
Automobiles	77	78	78	76	77
Banks (Business)	70	67	64	65	59
Banks (Public)	69	69	67	67	63
Clothing Retailers	63	63	62	63	63
Computers (Main Frames)	69	63	63	64	62
Computers (Business PCs)	70	66	66	67	64
Department Stores	62	63	61	61	60
Food Processors**	68	71	71	72	70
Furniture Retailers	64	63	65	65	64
Gas Stations	67	68	70	70	70
Grocery Stores	66	68	65	67	66
Insurance (Business)	64	62	64	62	61
Insurance (Automobile)	66	63	66	64	62
Insurance (Life)	65	65	63	61	54
Mail Order	na	64	63	64	64
Newspapers	na	60	64	63	62
Pharmacies	na	76	73	72	74
Police	56	55	58	59	58
Postal Service (Business)	59	62	65	61	66
Postal Service (Public)	65	61	67	63	65
Railroad	44	55	54	54	54
Shipping	na	65	69	67	69
Travel (Charter)	68	67	68	68	68
Telecom. (Business)	53	57	57	61	61
Telecom. (Public)	55	59	61	59	61
TV Broadcasters	44	43	47	48	49

* Satisfaction index is on a 0 to 100 scale.
** The averages for Food Processors include six separate food industries (basic foods, candy and coffee, baked goods and dairy products, beer, meat products, and canned and frozen foods).

Table 1. Swedish Customer Satisfaction Barometer Results

REFERENCES

Anderson, Eugene W., Claes Fornell, and Donald R. Lehmann (1994), "Customer Satisfaction, Market Share, and Profitability: Findings from Sweden," *Journal of Marketing*, forthcoming.

Andreasen, Alan R. and Arthur Best (1977), "Consumers Complain - Does Business Respond?," *Harvard Business Review*, July-August 1977, 93-101.

Bentham, J. (1802), in E. Dumont, *Theory and Legislation*, London: Trubner, 1871.

Fornell, Claes (1989), "The Blending of Theoretical and Empirical Knowledge in Structural Equations with Unobservables," in H. Wold (ed.), *Theoretical Empiricism*, New York: Paragon House, 153-173.

_____ (1992), "A National Customer Satisfaction Barometer: The Swedish Experience," *Journal of Marketing*, 56 (January), 6-21.

_____ and Michael D. Johnson (1993), "Differentiation as a Basis for Explaining Customer Satisfaction Across Industries," *Journal of Economic Psychology*, 14 (4), 681-696.

Hammond, Peter J. (1991), "Interpersonal Comparisons of Utility: Why and How they are and Should Be Made," in J. Elster and J. E. Roemer (eds.), *Interpersonal Comparisons of Well-Being*, Cambridge: Cambridge University Press, 200-254.

Hauser, John R. and Steven M. Shugan (1983), "Defensive Marketing Strategies," Marketing Science, 2 (4), 319-360.

Hicks, J. R. (1939), *Value and Capital*, London: Oxford University Press.

Hunt, H. Keith (1988), "Consumer Satisfaction/Dissatisfaction and the Consumer Interest," in E. Scott Maynes (ed.), *The Frontier of Research in the Consumer Interest*, Columbia, MO: American Council on Consumer Interests, 731-747.

Johnson, Michael D. (1984), "Consumer Choice Strategies for Comparing Noncomparable Alternatives," *Journal of Consumer Research*, 11 (December), 741-753.

_____ (1989), "The Differential Processing of Product Category and Noncomparable Choice Alternatives," *Journal of Consumer Research*, 16 (December), 300-309.

_____ (1994), "The Four Faces of Aggregation in Customer Satisfaction Research," Working Paper, Ann Arbor, MI: National Quality Research Center.

_____, Eugene W. Anderson and Claes Fornell (1994), "Rational and Adaptive Performance Expectations in a Customer Satisfaction Framework," Working Paper, Ann Arbor, MI: National Quality Research Center.

_____ and Claes Fornell (1991), "A Framework for Comparing Customer Satisfaction Across Individuals and Product Categories, *Journal of Economic Psychology*, 12 (2), 267-286.

_____, Georg Nader and Claes Fornell (1994), "Expectation, Perceived Performance, and Customer Satisfaction for a Complex Service: The Case of Bank Loans," Working Paper, Ann Arbor, MI: National Quality Research Center.

Jorgenson, D. W. (1990), "Aggregate Consumer Behavior and the Measurement of Social Welfare," *Econometrica*, 58, 1007-1040.

Lancaster, Kelvin (1971), *Consumer Demand: A New Approach*, New York: Columbia University Press.

Lingoes, James C. and Martin Pfaff (1972), "The Index of Consumer Satisfaction: Methodology," in M. Venkatesan (ed.), *Association for Consumer Research: 3rd Annual Conference Proceedings*, 689-712.

Lohmöller, J.-B. (1989), *Latent Variable Path Modeling with Partial Least Squares*, Heidelberg: Physica-Verlag.

Meeks, J. G. Tulip (1984), "Utility in Economics: A Survey of the Literature," in Charles F. Turner and Elizabeth Martin (eds.), *Surveying Subjective Phenomena, Volume 2*, New York: Russell Sage Foundation, 41-91.

National Economic Research Associates (1991), *Developing a National Quality Index: A Preliminary Study of Feasibility*, Report prepared for The American Quality Foundation, White Plains, NY: NERA, Inc.

Ölander, Folke (1977a), "Can Consumer Dissatisfaction and Complaints Guide Public Consumer Policy?" *Journal of Consumer Policy*, 1 (Spring), 124-137.

_____ (1977b), "Consumer Satisfaction - A Skeptic's View," in H. Keith Hunt (ed.), *Conceptualization and Measurement of Consumer Satisfaction and Dissatisfaction*, Cambridge, MA: Marketing Science Institute, 409-452.

_____ (1988), "Consumer Satisfaction/Dissatisfaction and the Consumer Interest," in E. Scott Maynes (ed.), *The Frontier of Research in the Consumer Interest*, Columbia, MO: American Council on Consumer Interests, 753-759.

Oliver, Richard L. (1977), "Effect of Expectation and Disconfirmation on Post-exposure Product Evaluations: An Alternative Interpretation," *Journal of Applied Psychology*, 62, 480-486.

_____ (1980), "A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions," *Journal of Marketing Research*, 17 (November), 460-469.

Pfaff, Martin (1977), "The Index of Consumer Satisfaction: Measurement Problems and Opportunities," in H. Keith Hunt (ed.), *Conceptualization and Measurement of Consumer Satisfaction and Dissatisfaction*, Cambridge, MA: Marketing Science Institute, 36-71.

Poiesz, Theo B. C. and Jasper von Grumbkow (1988), "Economic Well-Being, Job Satisfaction, Income Evaluation and Consumer Satisfaction: An Integrative Attempt," in W. F. van Raaij, G. M. van Veldhoven and K. E. Wärneryd (eds.), *Handbook of Economic Psychology*, Dordrecht, the Netherlands: Kluwer Academic Publishers, 570-593.

Robbins, L. (1938), "Interpersonal Comparisons of Utility: A Comment," *Economic Journal*, 48, 635-641.

Scitovsky, Tibor (1951), "The State of Welfare Economics," *American Economic Review*, 41, 303-315.

Sen, A. K. (1979), "Interpersonal Comparisons of Welfare," in M. J. Boskin (ed.), *Economics and Human Welfare: Essays in Honor of Tibor Scitovsky*, New York: Academic Press, 183-201.

Simon, John L. (1974), "Interpersonal Welfare Comparisons can be Made - and Used for Redistribution Decisions," *Kyklos*, 27, 63-98.

Tinbergen, Jan (1991), "On the Measurement of Welfare," *Journal of Econometrics*, 50 (1), 7-13.

Van Raaij, W. Fred (1981), "Economic Psychology," *Journal of Economic Psychology*, 1 (1), 1-24.

Wikström, Solveig (1983), "Another Look at Consumer Dissatisfaction as a Measure of Market Performance," *Journal of Consumer Policy*, 6 (1), 19-35.

Wold, Herman (1982), "Systems Under Indirect Observation Using PLS," in Claes Fornell (ed.), *A Second Generation of Multivariate Analysis: Methods*, New York: Praeger, 325-347.

Zeithaml, Valerie A., Leonard L. Berry, and A. Parasuraman (1988), "Communication and Control Processes in the Delivery of Service Quality," *Journal of Marketing*, 52 (April), 35-48.

Session 3 Customer Surveys

Discussion

Robert M. Groves
University of Michigan and Joint Program in Survey Methodology

In addressing the notion of customer service standards and customer measurement programs, the U.S. government is attempting to import a set of ideas tried in the commercial sector. It is useful to note that the ideas, once tried, do not always prove themselves to be cures of the ailments of modern commercial organizations. Why they work sometimes and why they don't other times is the topic of much current debate. We are now living through the period of time when most of you in the audience are determining whether this is the management philosophy of the week or the beginning of a new perspective on agency functioning.

First, let's take a minute to review the recent history of the commercial sector. the common lessons of the "customer satisfaction movement" are:

- external threats help shock organizations into paying attention to their customers
- customer orientation succeeds only when top management forces it, repeatedly, in every forum, relentlessly
- measurement of satisfaction only once is nearly useless
- measurement of satisfaction without simultaneous measurement of production/service activities related to satisfaction is nearly useless

Now let's see whether these lessons are relevant to the papers we have heard presented today.

The Devens Paper

There are really three parts to the Devens paper -- a commentary on the customer-orientation movement, the description of a survey, and a review of the feedback loop to managers.

Commentary on customer-orientation movement

Devens notes that managers do "want" to improve their operations, one way or another. Clearly, the question is whether the operations *are* being improved. It seems clear that all change is difficult to induce in government agencies, but if anything, continuous improvement changes may be more difficult in ongoing statistical operations than in other areas. The problem stems from the need to maintain comparable measurement

systems over time in ongoing series. The concern is change that affects the bias properties of estimates, not just the variance properties (yet even changes that theoretically affect only variance properties (eg. a new sample) can affect bias).

Why is that a concern? Rarely do we have information that bias change goes in the right direction. Exceptions are the higher victimization rates of the NCS and higher unemployment rates of CPS, where there is a strong model of tendencies to underreport those phenomena, but even there the model can be easily challenged.

So incremental change in statistical operations may be harder than incremental change in other fields because the product of today has more value if it is comparable to the product of yesterday.

Devens notes that many managers are skeptical about customer focus because customers aren't qualified to judge the quality aspects of statistical series. This comment on the surface sounds familiar to those in charge of the design of the 1994 Chevrolet Caprice, but statistical agencies may have greater challenges than manufacturers of other products. The General Motors managers did have access to many marketing research studies about the concerns and interests of their customers, but apparently discounted them. Most statistical agencies have no equivalent of the market research function, and thus customer desires are only indirectly and erratically communicated. As with automobiles it is easy to confuse the fact that only the customers know what information they need, but only the statisticians may know how best to produce it.

Description of the Survey

In this section there is very little concern about issues concerning the sampling frame and inference. This unfortunately is a serious lacuna in most of the literature about the Total Design Method. Despite its name it does not address issues of coverage error in surveys. These issues are complex and largely uncharted in surveys of customers of statistical agencies because the target population of customers has usually not be fully enumerated at the time of a survey. Even the definition of "customer" becomes a complex one, when considering information as a product.

The most important point of this section is that careful planning of a mailed questionnaire can yield high response rates. For this reason alone, this is an important paper. When government agencies are telling others and are being told that response rates in the 20% range are the highest to be expected, this work has shown that careful planning and execution can obtain high participation rates. High response rates are doubly important in this area, because of the finding that nonrespondents to satisfaction surveys tend to be dissatisfied with the service or product.

Review of Feedback Loop to Managers

Despite some commentary that the survey did not yield clear findings, conclusions were indeed drawn from the data collection -- the need for more timely faster products, one stop service for questions, and clearer presentations of information. Of these, it appears that the organization has addressed timeliness of products most directly. It is noteworthy

that this area was not the one with the largest gap between expectations and perceived performance. This might be an example of management either ignoring the empirical findings or management supplementing the empirical work with other external information about performance. The paper does not reveal which is the appropriate interpretation.

The Johnson Paper

This paper describes a large effort to construct a useful measure of consumer satisfaction across all sectors of the economy. It is conceptualized as another macroeconomic indicator, measuring an outcome of production-- in one term "post-consumptive utility" The paper is divided into three sections: a) Can satisfaction be compared across sectors/industries?; b) a description of the Swedish Customer Satisfaction Barometer; and c) an announcement of the American Customer Satisfaction Index

Can satisfaction be compared across sectors/industries?

How would you know whether you'd have the answer to this question? Would it depend on the ability to predict behavior? What behavior? The evidence of comparability presented includes the finding that 50% of the variance in perceptions of performance is explained by the amount of differentiation in the industry, and performance explains 50% of variance in satisfaction. Clearly, one would like to assemble more evidence: behavioral outcomes measured on same persons over time, stable relationships between satisfaction levels and growth rates, complaint rates, etc.;

Swedish Customer Satisfaction Barometer

This is a large data collection and estimation series, running since 1989, covering 70% of sales in each industry, and measuring one product or service per company. Clearly the process of sampling firms and products/services is a nontrivial problem, as those in the consumer and producer price index measurement process know. The sampling problem facing this index is multi-level (sectors, firms, products/services, customers) there are important sources of variance at each level and important sources of information about customer satisfaction.

The concepts measured include: perception of performance, expectations regarding performance, satisfaction, reports of whether the customer has complained about the product or service, and customer loyalty. These are difficult measurement issues. For example, the approach is forced to use hypothetical questions on loyalty, using words like "if another agency could provide the same service."

The paper presents findings from the Swedish effort that are stimulating, given the current effort at measuring customer satisfaction in U.S. government agencies. For example, there is the finding that Swedish government agencies (police, pharmacies, post office, railroads, telecommunications, tv broadcasting) started with lower satisfaction and rose in satisfaction faster than other sectors. One wonders whether that finding will be duplicated in the US. The finding itself illustrates one of the challenges to the measurement process. To what extent is measurement of satisfaction with services of a government agency affected by general feelings of civic pride, trust in government,

political efficacy?

American Customer Satisfaction Index

The last section of the paper sketches out the plans for a U.S. customer satisfaction index. This effort will be different from the Swedish experience in that the U.S. population of firms offer more diverse products and services.

In both this and the Swedish index description there seems to be most emphasis placed on the psychometric properties of the measurement and little concern with traditional survey issues of coverage of the target population (telephone surveys are planned), nonresponse error, and measurement errors associated with social desirability, mode of data collection, etc.

Summary

These two papers, although seemingly disparate in the focus, can serve to remind us of two important debates in customer satisfaction:

- Is satisfaction merely a function of the difference between reported expectations and performance ratings?

- Do expectations cause perceived performance?

Let me summarize my reactions to the papers:

1. In government agencies, we are at the beginning of the customer measurement process. Its value rests on repeated measurement, empirical assessment of relationship between actions of employees and satisfaction, and change in satisfaction over time. We are a long way from this status of measurement and innovation.

2. Both of these papers appear to miss the connection to actual activities of the units to increase satisfaction. They are more heavily focused on measurement than how measurement can lead to change and then later to improved satisfaction.

3. The papers flow from different conceptual bases; a debate that is not joined by the two. One stems from the notion that satisfaction is in some sense the gap between expectations and performance. The other attempts to add another concept, perceptions of an ideal service or product, in order to calibrate the gap between expectations and performance. These conceptual differences are part of the debate now ongoing in the satisfaction measurement field. These are important issues for the practical import of satisfaction measurement. If, for example, performance at time 1 sets expectations at time 2, then poor performance lowers expectations, and in one perspective, would yield higher satisfaction, as expectations and performance were in sync. From the other perspective, departures between performance and the concept of the ideal, would be larger at time 2 and lead to large "performance gaps."

If government agencies take seriously the measurement of customer satisfaction, they will

inevitably be forced to attend to such issues. They are key to the meaningful tracking of how satisfaction can change with improved performance of organizations. We are at the beginning of this process for government agencies, and we are in the debt of these two papers for alerting us to such issues.

DISCUSSION

Elizabeth Martin
Bureau of the Census

Both papers presented in this session raise issues which are relevant to the current efforts by federal agencies to respond to a presidential order to survey their customers, measure satisfaction, and use the results to set service standards and provide customers with greater choice in services. Significantly, the executive order further states that "as information about customer satisfaction becomes available, each agency shall use that information in judging the performance of agency management and in making resource allocations."

If this aspect of the executive order comes to pass, then the issue of the comparability of customer satisfaction measurements among agencies and across diverse products and services is of more than academic interest. I'd like to start by focussing on the issue of comparability of measurement as addressed in the Johnson paper, which describes a customer survey conducted across 32 industries in Sweden, including government-run industries such as the postal service and railroads. I'll be drawing on my recent involvement in an effort to design a generic customer satisfaction questionnaire for use by all of the agencies of the Department of Commerce. Next, I'll discuss the Devens paper, which discusses a customer satisfaction survey targetted much more narrowly, to users of employment and unemployment statistics produced by the Bureau of

Labor Statistics. It raises some interesting issues about the utility and consequences of customer satisfaction surveys.

Johnson and his colleagues are concerned with a very ambitious effort to develop a customer satisfaction index (the Swedish Customer Satisfaction Barometer) which can be applied across industries, products, and, ultimately, in different countries. Customers were sampled from 32 Swedish industries, including several which were government-run, which represented 70 percent of Sweden's gross domestic product. The companies that accounted for approximately 70 percent of combined industry sales were selected to represent each industry. Each company was represented by its product or service with the highest sales. Each year, 100,000 customers were contacted by telephone and screening questions were asked to determine if they had experience with the products and services chosen to represent the sample companies. About one quarter were eligible and were asked questions to measure their perceptions of performance, their expectations about performance, their satisfaction, whether they ever complained, and their loyalty.

Johnson argues that the survey measurements may be used to compare satisfaction and performance across industries on the following basis: that the model of customer satisfaction which informs their measurements is universally applicable across industry; the measurements used in their survey are universally applicable across all industries; and that there exist meaningful differences in satisfaction between industries which can be explained by industry-level differences in degree of product or

service differentiation--the a. ıt of choice offered customers. He also suggests that that the index provides meaningful and comparable information about customer satisfaction for government agencies and industries as well as private sector ones.

Although it may be possible to design a survey which yields comparable measurements of satisfaction for customers of diverse products and services produced by different industries, Johnson and his colleagues have not satisfactorily made the case for the comparability of their measurements. They need to address the following questions: First, is the definition of a <u>customer</u> comparable across different industries? Second, are the sampling frame and response rates comparable across industries and over time? Finally, are their satisfaction measurements comparable and meaningful across different industries and over time?

The first issue, of what is a customer, usually is not terribly ambiguous in the private sector, but it bedevils attempts to measure customer satisfaction in a government setting. To most of us, a customer is someone who purchases a commodity or service, usually by choice or voluntarily. In a government setting, many products and services are not purchased directly by their users, but subsidized in whole or in part by taxes. Many government products and services are not received voluntarily on the part of the "user" or recipient. Many "customers" of police services or tax collection services no doubt would, if they had the choice, choose not to obtain the service at all. The Johnson paper does not address this issue. In their survey, they defined as customers persons who <u>had experience</u> with the products and services surveyed.

This is a reasonable strategy, but it is important to include measures in the survey to permit the analyst to separately identify voluntary customers of government services, who obtained a product or service by choice, and involuntary customers, who did not choose to obtain the product or service. Comparisons of the former group with customers of private industry may be meaningful, but comparisons involving the latter group probably would not be.

Another definitional problem arises because many services offered by the government are not intended to benefit those who experience them directly, but to protect or benefit others, such as the public, who may not even be aware of their existence. For example, one service provided by a Department of Commerce agency is the inspection of fisheries. Presumably this service is ultimately intended to benefit fish-eaters by ensuring the quality of fish, and only indirectly benefits the fisheries themselves. In this example, it would be difficult to measure the satisfaction of customers who may be unaware that a service exists, much less that they are recipients of it. This sort of issue can make it difficult to identify who should be regarded as the customers of government services, and this category of "customer" would be ruled out by Johnson's screening criterion of "having experience" with a product or service.

A second set of very difficult issues affecting comparability of data has to do with the identification and sampling of customers. In order to make comparisons across industries, one must be certain that the samples are comparable. Johnson reports that each year customers were identified in surveys of the public in

which 100,000 individuals were asked about their experiences with the target products. It is not clear how or whether the sample represents organizational customers. For many of the industries being evaluated--such as banks, railroads, main frame computers--much if not most of their business would be with other organizations or businesses, not with individual consumers. Organizations, and their experiences as customers, do not appear to be represented in the index. Their absence reduces the meaningfulness of customer satisfaction measures for industries in which organizational customers represent a large share of all customers, and reduces comparability of measures across industries which differ in their customer base.

Even if one accepts the limitation that only individual customers are represented in the sample, it is still unclear what universe the results represent. Johnson surveyed customers of the leading products of companies representing 70 percent shares of each of a set of industries which together accounted for 70 percent of the Swedish GDP. One would expect the companies, products, and industries included in the index to change over time with changes in the economy. This implies that there are two potential sources of change in the value of the index: changes in customer satisfaction for a given set of products, and changes in the composition of products, companies, and industries which make up the index. Given the uncertain interpretation which could be put on any given change in its level, it is not clear how a customer satisfaction index defined this way can provide useful information about trends. Moreover, Johnson's definition appears to leave out most customers,

since 75 percent of his sample was ineligible for the survey. The limited and rather peculiar constraints on the set of customers included in the survey seem to reduce its usefulness as a general index of customer satisfaction with wide applicability across different countries.

The quality and comparability of the results of customer satisfaction surveys depends not only on the quality of the sampling frame, but also on response rates. Johnson presents no information about response rates in the customer survey he reports. In order to make comparisons among industries, one would want reasonably high response rates for all the industries being compared. If response rates varied among industries, then artifactual differences in satisfaction may result from greater nonresponse bias for some than others.

In general, the construction of sampling frames for customer surveys is problematic. If there exist records of purchases, orders, or logs of telephone or other contacts, then these may be used as a sampling frame. However, for many services and products, there are no records which identify customers, especially if no formal or recorded transaction takes place. Customers who pay cash for a product or service, or who listen to the weather station or look up information in a census publication in the library cannot be readily identified. Thus, sampling from records or logs of customer transactions is likely to provide uneven coverage of customers depending on the nature of the industry, how it conducts business with its customers, and the quality and completeness of the records it keeps about customers or transactions. In some

customer surveys, samples are drawn from lists of customers provided by an agency or firm specifically for the survey. Such lists can be very vulnerable to selection bias, since organizational representatives who know that customer satisfaction is to be evaluated are likely to overrepresent satisfied customers in their lists. This selection bias may vary among agencies or organizations and could have a very adverse effect on the comparability of satisfaction measures across agencies or organizations.

The third issue which needs more attention in the Johnson paper is the comparability and meaningfulness of the measurements, especially when applied in a government context. Performance was measured as value, or benefits relative to costs, which seems not to apply very well to products or services which have no specific or direct cost attached to them, as is the case for many government products and services. The key construct of customer loyalty was operationally measured by intention to repurchase and insensitivity to price. Because these measures do not fit government's transactions with its customers, Johnson and his colleagues changed the measures to hypothetical ones for government agencies. It is questionable whether this modification yields results comparable to the original measure. Finally, customer expectations were measured retrospectively, that is, customers were asked to report what their expectations had been at the time of purchase. Retrospective reports of past attitudes are notoriously biased toward present attitudes, and it is highly likely that this measure of "expectations" is contaminated by respondents' subsequent experiences with

a product. This flaw would make it impossible to test the effect of prior expectations on satisfaction or perceived performance.

In summary, customer surveys which aim to compare across diverse industries and products (such as the SCSB discussed by Johnson) potentially are affected by very serious problems of data comparability, including lack of comparability arising from sample design, differential nonresponse, and the measurements themselves. There appears to be a considerable amount of careful methodological and statistical work that still needs to be done to ensure that customer surveys are designed to yield meaningful comparisons of customer satisfaction across industries and over time. Until that groundwork is done, such comparisons should be made cautiously.

The Devens paper raises a different set of issues. The Customer Satisfaction Survey he reports on was much narrower in scope and purpose than the satisfaction index discussed by Johnson, and the issues of data comparability are not nearly as serious, if they exist at all. The survey of customers of employment and unemployment statistics was well done, and obtained an admirably high response rate (88 percent) using reminders and mailings of follow up questionnaires. The survey assessed several aspects of the statistical product, including data quality, ability of staff to answer technical questions, etc. Devens reports that the survey results moved the agency to take several actions to improve the timeliness of the release of its statistics, which the survey showed customers thought was very important.

The paper includes a couple of telling comments by the author, who personally evaluates the survey as "useful but not nearly in

proportion to the skills exercised or the resources expended." He also notes the very low response rate--8 percent-- obtained from a survey of front-line employees upon release of the customer survey, and voices his suspicion that this low rate is due to "failure to convince the front line that the customer satisfaction survey was serious."

Devens' remarks remind us of a couple of key points about customer surveys of this type. First of all, there needs to be clear specification of the goals of the survey, and an understanding of how the information from the survey will be used, in order for the survey to be useful. (This point applies to any survey, not just customer surveys.) To be taken seriously by employees, a customer survey should be designed to address questions to which managers and employees need or want the answers. In the case of the survey Devens reports, it appears that the survey was not credible to managers, and was used in a very limited way by them, reducing the meaningfulness of the survey.

The second point is that customer surveys can themselves affect the expectations of customers and employees. Carrying out a customer survey may raise the expectations of customers (and employees) that a company or agency is going to do something to improve service. If that doesn't happen, and if the survey turns out to be an empty exercise, then the indirect effect of a customer survey may be to reinforce the cynicism of customers or employees or both.

Taking these two points together, and returning to the earlier discussion, we can draw several general conclusions about customer

surveys: Unless an agency or company plans to actually use the results of a customer survey, it shouldn't conduct the survey. The survey should be planned with clear goals and uses in mind. It should be designed to provide fairly specific information that represents useful feedback to managers and employees, and that has implications for action. If the intent is to compare customer satisfaction over time, among products, or among industries and agencies, then the survey should be designed and data evaluated to ensure that results are comparable and can support the comparisons to be made.

Session 4

ADVANCES IN DATA EDITING

IMPROVING OUTLIER DETECTION IN TWO ESTABLISHMENT SURVEYS

Julia L. Bienias,[1] David M. Lassman, Scott A. Scheleur, and Howard Hogan
U. S. Bureau of the Census

I. Introduction

One step in producing estimates from survey data is editing. In many settings, trained analysts examine the data to find unusual or unexpected values, which may be the result of errors made by the respondent or in the data-capture processes. Having found a questionable case, the analyst then tries to verify its accuracy by checking the original form, obtaining related data from other sources, and/or contacting the respondent. One would like to correct as many errors as possible within the time limitations for a given survey. Thus, accurately identifying the cases whose values are most likely to be the result of errors is an essential part of efficient editing.

Previous researchers have successfully used various graphical methods to improve both the efficiency and accuracy of the editing process (e.g., Esposito, Fox, Lin, & Tidemann, in press; Granquist, 1990; Houston & Bruce, 1992; Hughes, McDermid, & Linacre, 1990). We describe the application of graphical methods from exploratory data analysis to the task of identifying potentially incorrect data points. Our report is the result of a working group of analysts, research statisticians, and programmers devoted to this effort.[2] We illustrate the methods with data primarily from the Annual Survey of Communication Services and the Monthly Wholesale Trade Survey. We first describe the two surveys and the current methods used for editing.

2. Descriptions of the Two Surveys

2.1 The Annual Survey of Communication Services

The Annual Survey of Communication Services (ASCS) is a mail survey covering all employer firms that are primarily engaged in providing point-to-point

---

[1]This paper reports general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau. Address correspondence to: Julia L. Bienias, Economic Statistical Methods and Programming Division, Bureau of the Census, FOB 3015-4, Washington, DC 20233.

[2]We thank the other members of our working group, without whom the work described here would not have been possible: Thomas Bell, Willard Caldwell, Vicki Garrett, Imelda Hall, Donald Hundertmark, Jennifer Juzwiak, William Knowlton, David Stachurski, and Georgeann Wright. We also acknowledge the many other members of Business Division and of Economic Statistical Methods and Programming Division who have supported this effort and who continue to support it.

communication services (e.g., telephone, television, radio), as defined in Major Group 48 of the 1987 edition of the *Standard Industrial Classification Manual*. The ASCS provides detailed revenue and expense statistics from a sample of approximately 2,000. The Census Bureau introduced the survey in 1991 to track the explosive growth and change in the industry. The Bureau of Economic Analysis is the primary federal user of the data collected; other users are the Bureau of Labor Statistics and private industry (U.S. Bureau of the Census, 1992.)

2.2 The Monthly Wholesale Trade Survey

The scope of the Monthly Wholesale Trade Survey (MWTS) is all employer firms engaged in wholesale trade, as defined by Major Groups 50 and 51 of the 1987 edition of the *Standard Industrial Classification Manual*. Particularly, the survey covers merchant wholesalers who take title to the goods they buy and sell, collecting sales and inventory information. The MWTS, conducted since the 1940's, is a mail survey of approximately 7,000 firms, of which 3,500 receive forms in a given month. As with the ASCS, the Bureau of Economic Analysis is the primary federal user of the data. (See U.S. Bureau of the Census, 1994.)

3. Issues Involved in the Current Editing Procedures

After the data from the questionnaires are keyed, a computer program flags cases failing completeness, internal consistency, and/or tolerance edits. Editing review is divided among several analysts for a given survey. Each analyst finds which edits have failed for a case through an interactive correction system or a paper listing, on a case-by-case basis. They can also use a database query system to try to find problem cases that have not already been identified.

There are several disadvantages to this approach. Examining one case at a time does not permit the analyst to obtain a broad view of the behavior of the industry as a whole, and such a view can be of great benefit in determining the impact of an individual unit on the aggregate estimate. In addition, it undoubtedly leads analysts to examine more cases than necessary. Finally, for a few of the ASCS tolerance edits, constant parameter levels derived from previous surveys have been hard-coded into the programs. This implicitly assumes the relationships among the variables are static over time, which may not be the case.

4. Application of Exploratory Data Analysis Methods

4.1 Background

Exploratory data analysis (EDA) can be described as "a set of tools for finding what we might have otherwise missed" in a set of data (see Tukey, 1977). These tools, combined with the analysts' subject-matter expertise, are particularly well-

suited to the task of data editing. In this setting, we are not interested in ascertaining the truth of a postulated economic model or a similar estimation or hypothesis testing problem. Rather, our goal is to determine which cases are unusual with respect to the bulk of the cases and to follow up those cases. In addition to providing methods for displaying data in a variety of ways, EDA emphasizes fitting data using methods that are relatively insensitive to the presence of outliers in the data ("resistant" methods). Such fitting is a way to define and then account for (remove) certain aspects of the data so the analyst can concentrate on other aspects. (See Hoaglin, Mosteller, & Tukey, 1983; Velleman & Hoaglin, 1981.)

EDA fits well with the survey processing environment. Because in the editing stage we expect to find wild observations that might be off by orders of magnitude from the bulk of the data, transformations and resistant techniques are particularly useful in helping us find order amid the chaos. In addition, these techniques allow for efficient examination of large amounts of information at once, an aspect that is particularly valuable in the time- and resource-constrained survey production environment.

From the arsenal of tools collectively called "exploratory data analysis," we considered both univariate boxplots and the more general bivariate fitting. We describe boxplots first, followed by scatter plots and some methods for fitting. In addition, although transformations are applicable to all tools, we describe them in the context of scatter plots, because that is where we used them most.

4.2 Boxplots

Boxplots allow quick visual analysis of the location, spread, and shape of a distribution. Our boxplot has its box spanning the lower and upper quartiles, with whiskers extending from the box to the furthest data point within a distance of one-and-one-half times the interquartile range from the box. We considered data values beyond the whiskers as potential outliers. If the data are reasonably symmetric, then these cutoffs provide a good working definition of cases which may need review. See Tukey (1977) for a discussion of boxplots in general, and Hoaglin, Mosteller, and Tukey (1983) for a discussion of the expected number of outliers for different sample sizes. Note that the whisker definition could be modified to suit the needs of a particular survey operation (e.g., one could use 2 times the interquartile range instead of 1.5).

Figure 1 demonstrates the use of the boxplot for operating ratio (expenses/revenue) data from the ASCS.[3] The boxplot shows that the median operating ratio is .7978 and fifty percent of the points lie between .7269 and .9811. The left and right whisker values are .3760 and 1.3401. The cases flagged by the

[3]To protect the confidentiality of our data, we have not provided details about the particular subset of data analyzed in each plot, nor have we labeled axes when such information could be revealing.

use of the boxplot are different (and fewer in number) than the cases that would have been flagged by the current hard-coded edit parameters, .9 and 1.1. Those parameters fail to help us isolate the "true" outlier cases, as they result in too many cases being flagged. Alternatively, we could flag cases that would appear beyond the whiskers as in our boxplot, an approach that is "dynamic" in that it relies on incoming data to set parameters. At minimum, we could use values from Figure 1 as new hard-coded edit bounds, noting that these revised bounds would no longer be symmetric around one (consistent with the findings of Granquist, 1990).

4.3 Scatter Plots

A scatter plot of two variables is a simple and particularly useful technique. When the data are appropriately transformed, one can use a variety of methods to remove linearity in the scatter and then examine the residuals from the linear fit. This allows us to see patterns that we might otherwise miss when looking at the original data; looking at the residuals from a fit allows us to examine the data on a finer scale (see Section 4.5).

As a vivid illustration of the kinds of problems encountered in editing data, we used another survey for which we had raw responses to a particularly problematic question. One item in the Motor Freight Transportation and Warehousing Survey is the percent of revenue derived from local trucking, a question believed to be confusing to respondents may define "local" in different ways. Figure 2, a scatter plot of these unedited data for the current versus prior period, shows a weak linear relationship. Cases along the 45 degree line are companies whose year-to-year reports are consistent. The reports become more inconsistent the further they are from the 45 degree line. Some of the cases along the vertical axis are "births" to the survey (cases selected during the current period to reflect new firms). Births should be analyzed separately, because they have only current-year data.

4.4 Transformations

Transforming the data so patterns can be more easily discerned is a technique that is important to all graphical and data-fitting methods. It is used to obtain symmetry in the data, to promote linearity, and to equalize spreads between data sets. These properties are assumed, implicitly or explicitly, by many of the techniques we use to analyze data. For example, when we look for outliers by examining a boxplot, we are implicitly assuming the data are supposed to be symmetric. If the data are naturally skewed, many of the points in the tail that appear to be outliers are actually values that are consistent with the underlying distribution. Thus, "discovering" such outliers in the long tail would not be very meaningful. With skewed data, we want to spend our time investigating those data points that are particularly unusual, given that we expect many points far from the bulk of the data. If we transform skewed data to be generally symmetric, we can then find those points.

Because economic data are typically positively-skewed, transformations that lead to the expansion of lower data values and to shrinking the spread of larger data values are particularly useful. (See Hoaglin, Mosteller, & Tukey, 1983, for more details on types of transformations.)

Figure 3 is a an example of the use of transformations for the ASCS. The scatter plot of untransformed revenue data (Figure 3a) reveals little, as one case is many times larger than the other cases. Hiding the large case was unsuccessful, as the next largest case was still many times larger than the remaining cases. Instead, taking logs of the data showed a useful scatter plot (Figure 3b). We see a strong linear relationship, which is what we expect for a plot of current and prior data. Cases that do not appear to be following this linear relationship would thus be considered unusual. We also see that the case that appeared to be an outlier in Figure 3a is, in fact, very much in line with the rest of the data.

For the MWTS, a scatter plot of the current inventory data against the current sales data shows that most of the data are bunched in the lower left corner (see Figure 4a). Because both variables are skewed, we first tried a natural log transformation ($\log(x+1)$). (We added one because a value of 0 for inventory data does not indicate the case is a birth, and thus it may be of value to include such cases.) This overtransformed the data, skewing them in the opposite direction (Figure 4b).[4] This is because there was a big gap in values between 0 and the next largest value. Such an effect would also occur if there were many establishments with very small reported data and a few with very large values. We then tried taking the square root (Figure 4c) and fourth root (Figure 4d). The latter resulted in the most useful transformation, as most of the data can be seen clearly.

4.5 Fitting

In this section we describe two approaches to fitting, ordinary linear regression and resistant regression. Both were useful, in different ways.

In analyzing ASCS data, we considered the relationship between revenue and payroll for current year data. Figure 5a shows the ordinary least squares regression of revenue on payroll; there are many points clustered near the origin and two cases in the upper right corner. First, we tried removing the two large cases. Again the distribution showed points clustered in the left corner. Such an approach, of iteratively hiding points and refitting, has the disadvantage of being subjective and of essentially requiring analysts to identify outliers first.

One alternative is to use ordinary least squares on transformed data. In this example, logs were useful. Figure 5b shows the fit to the logged data, depicting a

[4]If the cases with 0 reported inventory are ignored, as they might be for other variables, then the logarithm transformation provided a useable picture of the data.

strong linear relationship. The point labeled A is an obvious outlier. Examination of the residuals revealed a pattern, which allowed us to discover that tax-exempt cases were inadvertently being included in the analysis. Tax-exempt cases should be examined separately from taxable cases, because our revenue item only includes taxable receipts. Removing both those cases and point A and refitting the data (Figure 5c) led to the distribution of absolute residuals shown in Figure 5d. This plot can be used to detect outliers, as with a cutoff level C:

$$C = K * \text{(median absolute residual)}.$$

We found $K=4$ (corresponding to $C = .7868$) to be the best. All cases above .7868 were examined and most were "true" outliers. For our example, this method was judged by the survey analysts to be excellent for finding outliers.

Unfortunately, ordinary least squares (OLS) can give great weight to fitting a few wild values. It may work well, as in our example, when there are only a few wild cases and the demarcation between usual and unusual is clear. As an alternative, we investigated resistant fitting using the biweight function developed by Tukey (Mosteller & Tukey 1977; McNeil, 1977). This widely-tested iterative weighted-least-squares fitting procedure uses a weighting function defined as:

$$W_i = \begin{cases} (1-u_i^2)^2, & u_i<1 \\ 0 \text{ otherwise}, \end{cases}$$

where $u_i = (r_i / (c*s))$

$r_i \equiv$ Residual from previous fit for point i

$s \equiv$ mean absolute residual from previous fit

$c \equiv$ scale factor.

Setting $c = 4$ is quite resistant, $c = 8$ is moderately resistant. We stopped iterating when the proportionate change in s was less than 0.01. This required few iterations; resistant regression is a very efficient and fast procedure.

We applied resistant regression to the MWTS, predicting logged current inventory data by logged inventory data from the prior year. We expect a linear relationship. Figure 6a shows the data and the line from the OLS fit, and Figure 6b shows the residuals from that fit. It is easy to see the OLS fit misses the central tendency of the point cloud. Figure 7a shows the fit resulting from resistant regression ($c=4$). This fit more effectively removed the linearity from the data. The residuals now cluster around 0, as we would want (Figure 7b).

5. A Note on Using Ratios

In many instances, data review has relied on calculating ratios (e.g., sales/payroll) and looking for unusually large or small ratios. There is nothing wrong with this approach per se, but it would be wrong to rely too strongly on it.

The use of ratios assumes a rather simple model of the true relation between the two variables, specifically a straight line through the origin. The true relation may

differ markedly, there may be data clouds following different straight lines. For example, the relationship might be different for a small company than for a large company. It is essential that the data reviewer plot the data and look at the shape. Further, the "acceptable ratios" are often set from historic data, last year's or last census'. The relationships can change systematically throughout the business cycle. One could iterate, calculate the average ratio from the current survey, calculate its standard deviation, identify and remove outliers, and start again. However, given the existence of rather fine iterative resistant fitting tools, it is hard to see the advantage of this approach.

6. Summary and Extensions

We have described how principles and methods from EDA can be used to improve the efficiency and accuracy of editing, by helping analysts see patterns in the data and use that information to prioritize cases for follow-up. Building a successful editing system using this approach is more than just selecting the correct statistical tools. The system must be acceptable to the people who will use it. Creating such acceptance requires training the analysts in the methods described here, as well as incorporating the tools into the current production environment and existing computer systems. To date, we have been successful in getting many people to try the methods on several surveys. In addition to the surveys described previously, these methods are currently being applied to the Motor Freight Transportation and Warehousing Survey, the Service Annual Survey, and the Commodity Flow Survey.

Analysts for these surveys reported that being able to ascertain the effect of a given case on the estimate was quite useful. Other specialized programs written for data editing provide this feature (e.g., Esposito, Fox, Lin, & Tidemann, in press; Houston & Bruce, 1992). Incorporating sampling weights in the procedures described here provides a similar utility.

The EDA approach can be combined with batch-type edits (e.g., SPEER, Draper, Greenberg, & Petkunas, 1990; Lee, in press). One could examine the data flagged from a batch program along with the unflagged data using the tools described here. Or, the graphical-based methods could be the basis for batch-type dynamic edits. For example, a program could transform the data to be more symmetric and then flag all cases that would be beyond the whiskers of a boxplot. Finally, in settings in which hard-coded edit parameters must be used, these methods can be used on a subset of data to help find or evaluate such cutoffs.

7. References

Draper, L ., Greenberg, B., & Petkunas, T. (1990). On-line capabilities in SPEER (Structured Programs for Economic Editing and Referrals). *Proceedings of Statistics Canada Symposium 90: Measurement and Improvement of Data Quality*, pp. 235-44. Ottawa: Statistics Canada.

Esposito, R., Fox, J. K., Lin, D., & Tidemann, K. (in press). ARIES: A visual path in the investigation of statistical data. *Computational and Graphical Statistics*.

Granquist, L. (1990). A review of some macro-editing methods for rationalizing the editing process. *Proceedings of Statistics Canada Symposium 90, Measurement and Improvement of Data Quality*, pp. 225-34. Ottawa: Statistics Canada.

Hoaglin, D.C., Mosteller, F., & Tukey, J. W. (Eds.) (1983). *Understanding Robust and Exploratory Data Analysis*. NY: Wiley.

Houston, G., & Bruce, A. G. (1992, February). Graphical editing for business and economic surveys. Technical report, New Zealand Department of Statistics, Mathematical Statistical Branch.

Hughes, P.J., McDermid, I., & Linacre, S. J. (1990). The use of graphical methods in editing (with discussion). *Proceedings of the 1990 Bureau of the Census Annual Research Conference*, pp. 538-54. Washington, DC: U.S. Department of Commerce.

Lee, H. (in press). Outliers in survey sampling. In B. Cox et al. (Eds.), *Survey Methods for Business, Farms, and Institutions*. NY: Wiley.

Mosteller, F., & Tukey, J. (1977). *Data Analysis and Regression*. Reading, MA: Addison Wesley.

McNeil, D. R. (1977). *Interactive Data Analysis*. NY: Wiley.

Office of Management and Budget. (1987). *Standard Industrial Classification Manual*. Available from National Technical Information Service, Springfield, VA (Order no. PB 87-100012).

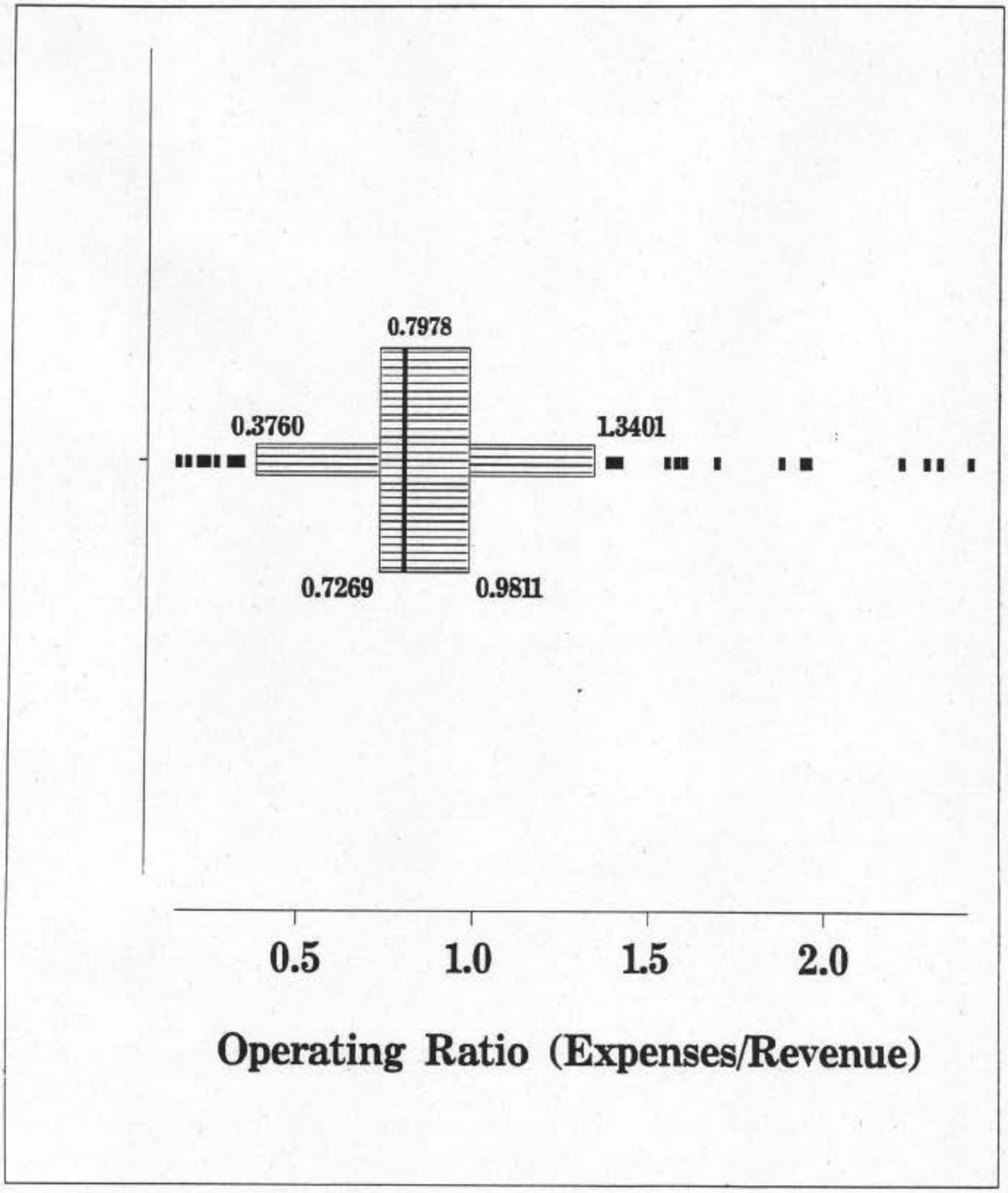
Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

U.S. Bureau of the Census. (1992). *Annual Survey of Communication Services: 1992*. Washington, DC: U.S. Government Printing Office (Current Business Reports, Item BC/92).

U.S. Bureau of the Census. (1994, April). *Combined Annual and Revised Monthly Wholesale Trade, January 1987-December 1993*. Washington, DC: U.S. Government Printing Office (Current Business Reports, Item BW/93-RV).

Velleman, P.F., & Hoaglin, D. (1981). *Applications, Basics, and Computing of Exploratory Data Analysis*. Boston: Duxbury Press.

Figure 1


0.7978
0.3760
1.3401
0.7269
0.9811
0.5
1.0
1.5
2.0
Operating Ratio (Expenses/Revenue)

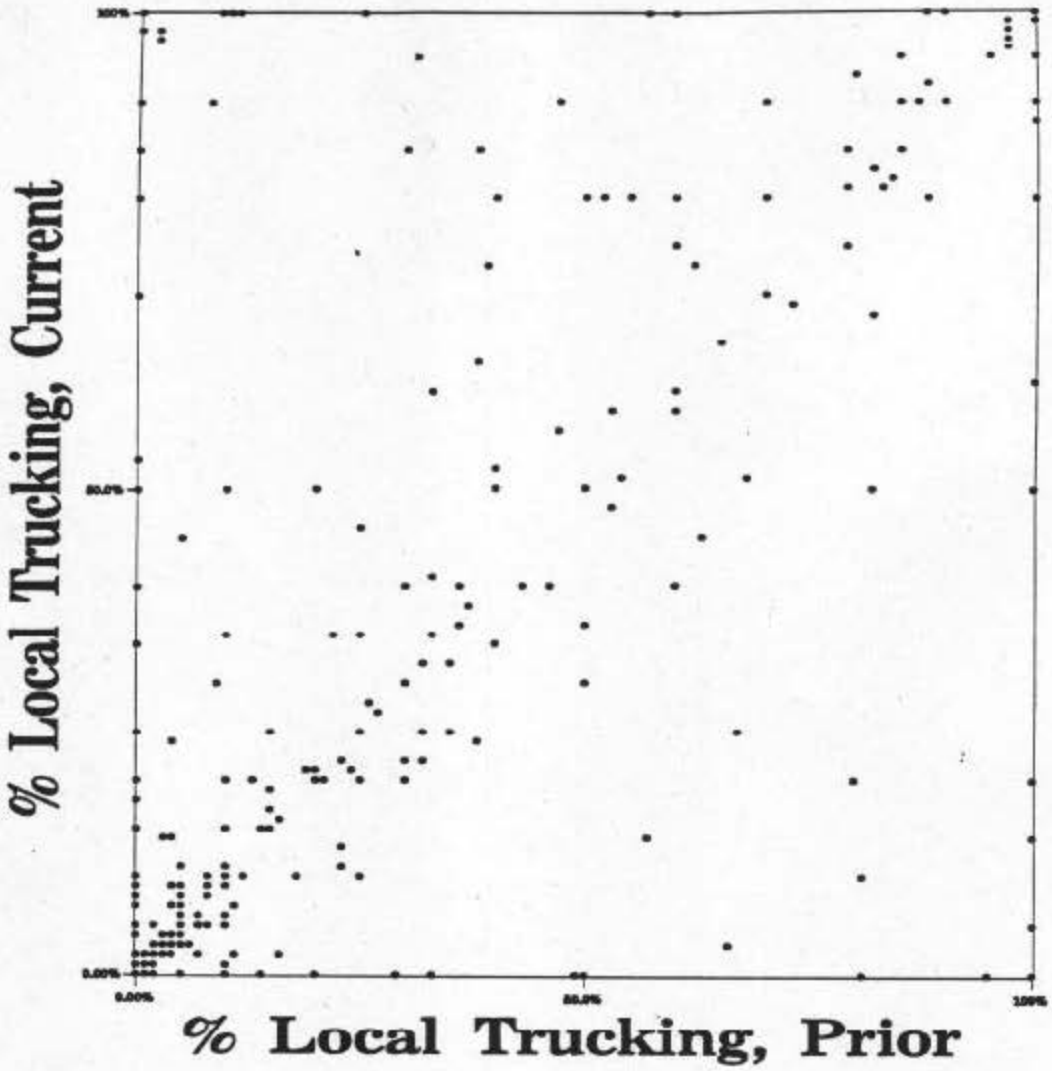
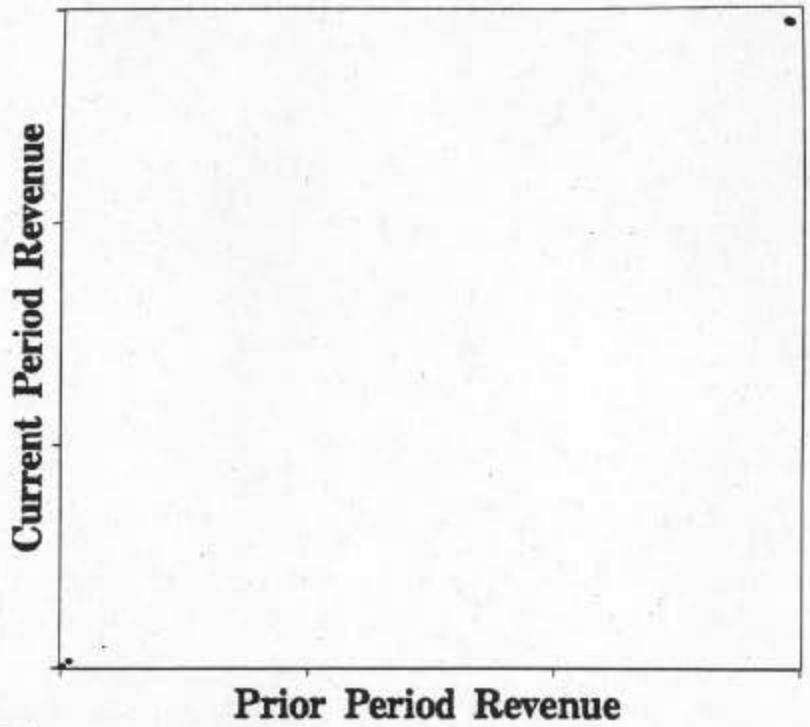
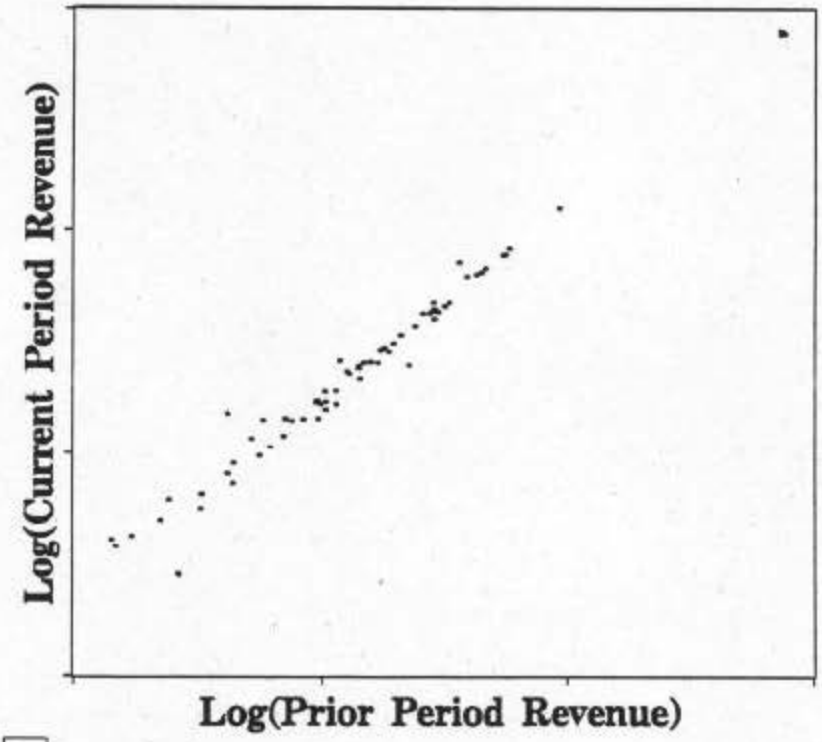


Figure 2

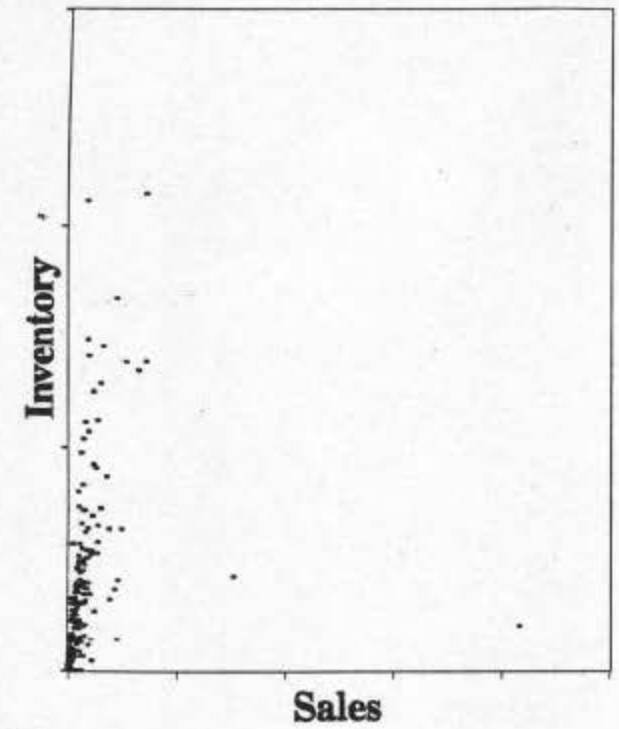

Current Period Revenue
Prior Period Revenue

a

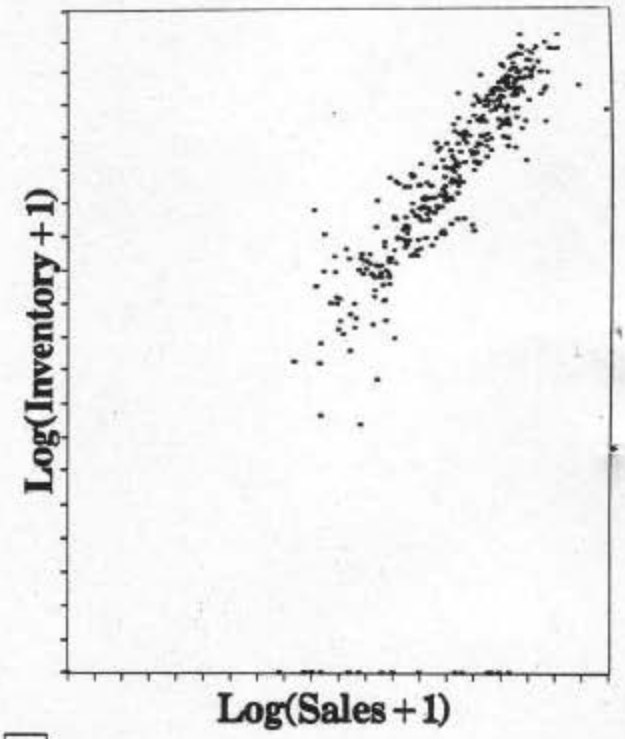
Figure 3


Log(Current Period Revenue)
Log(Prior Period Revenue)

b

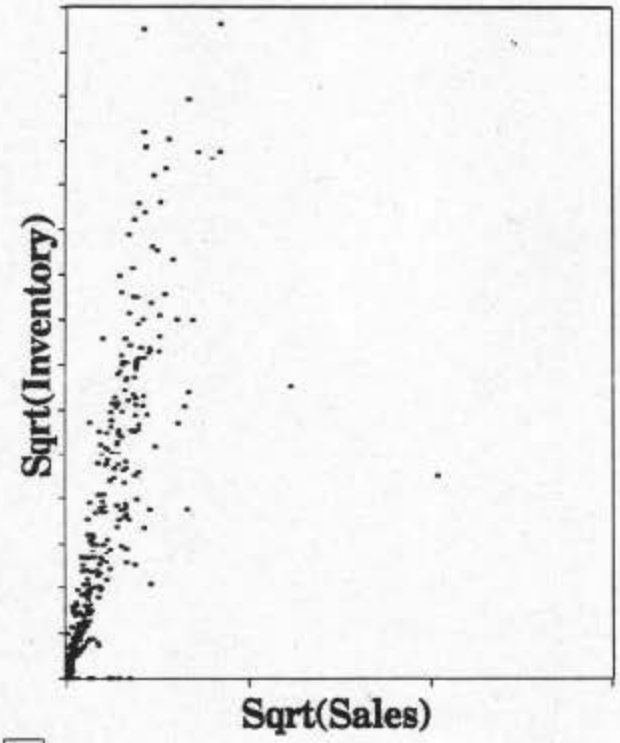


a

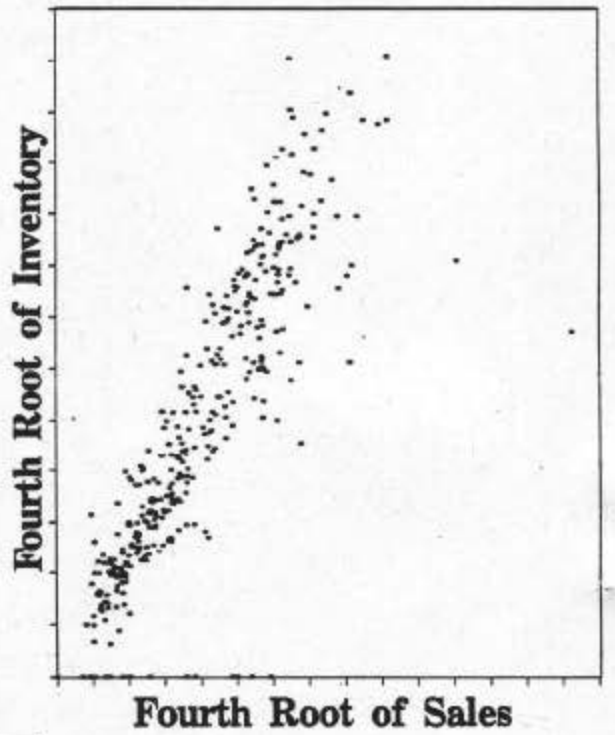


b

Figure 4



c



d

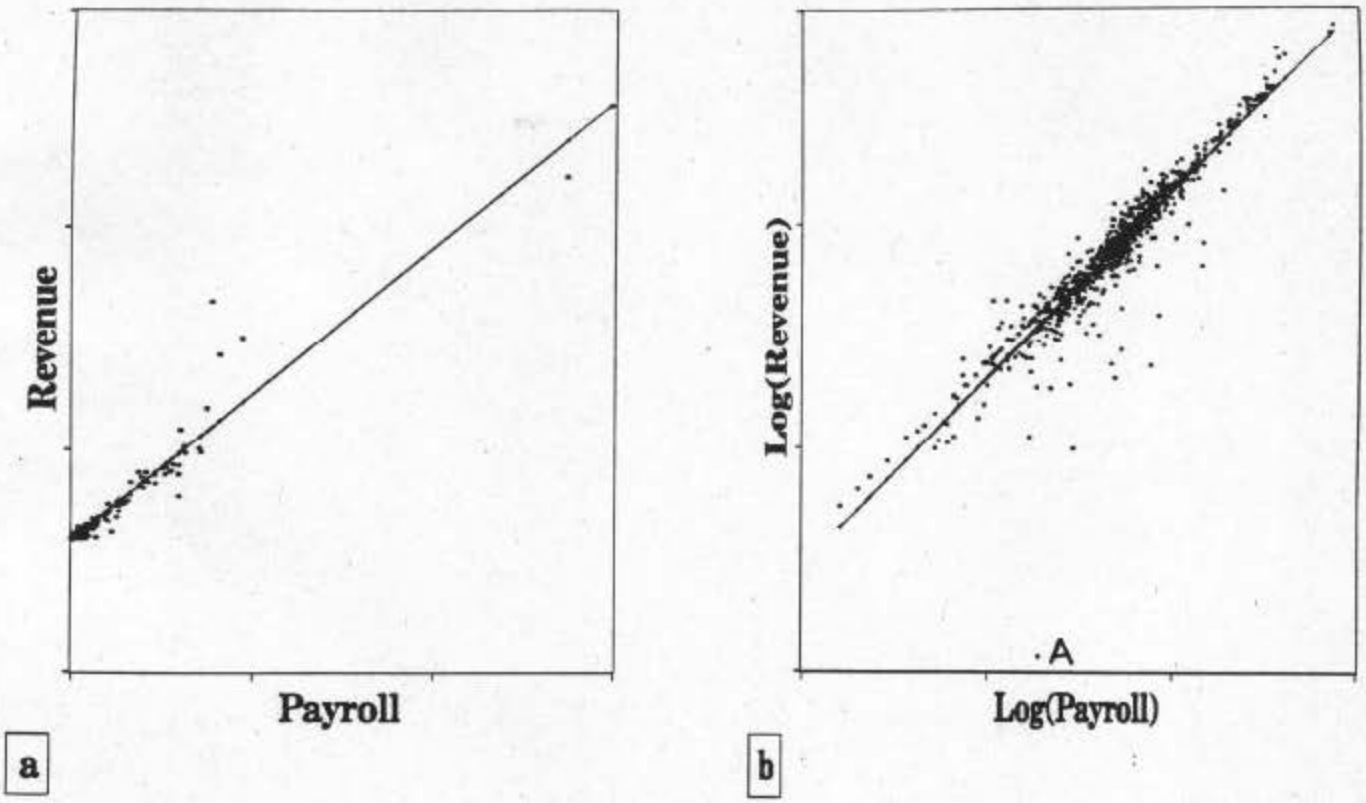
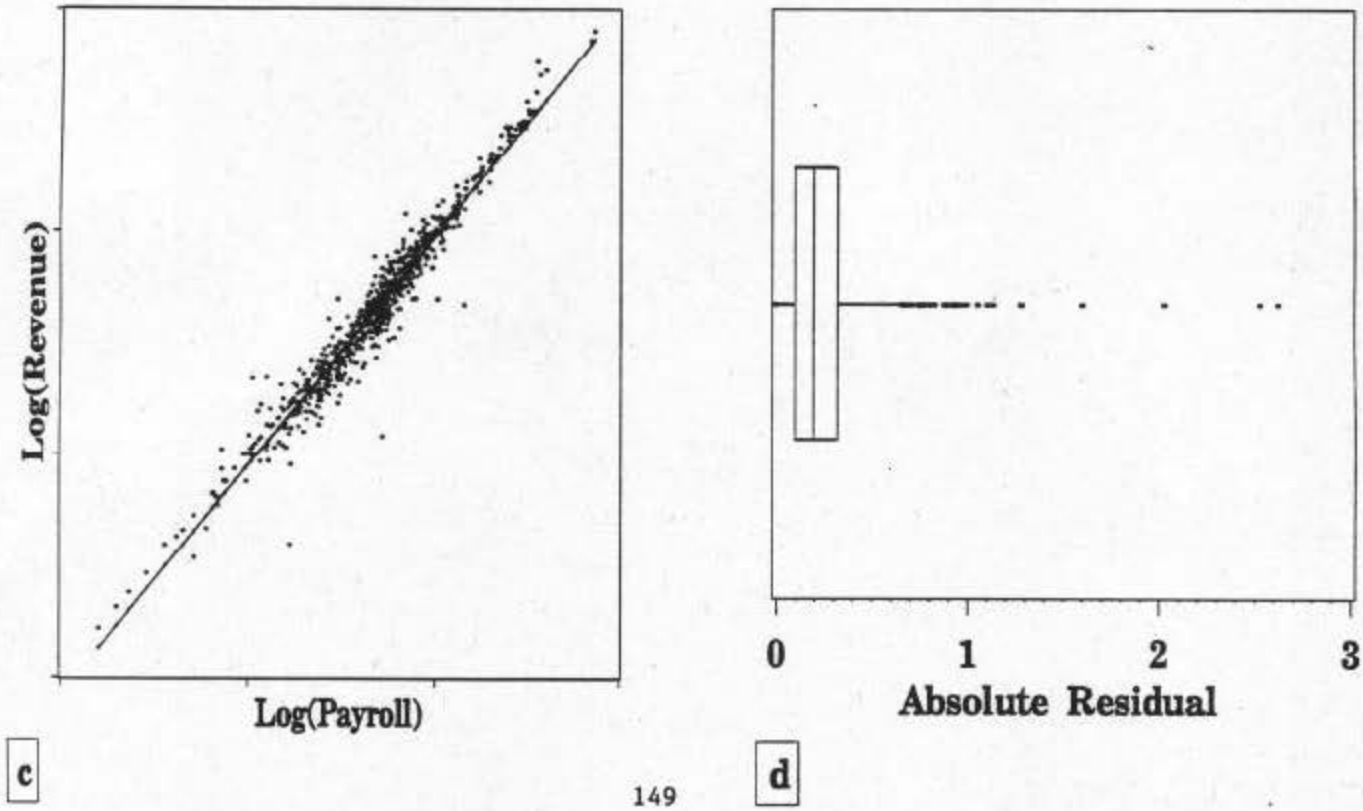
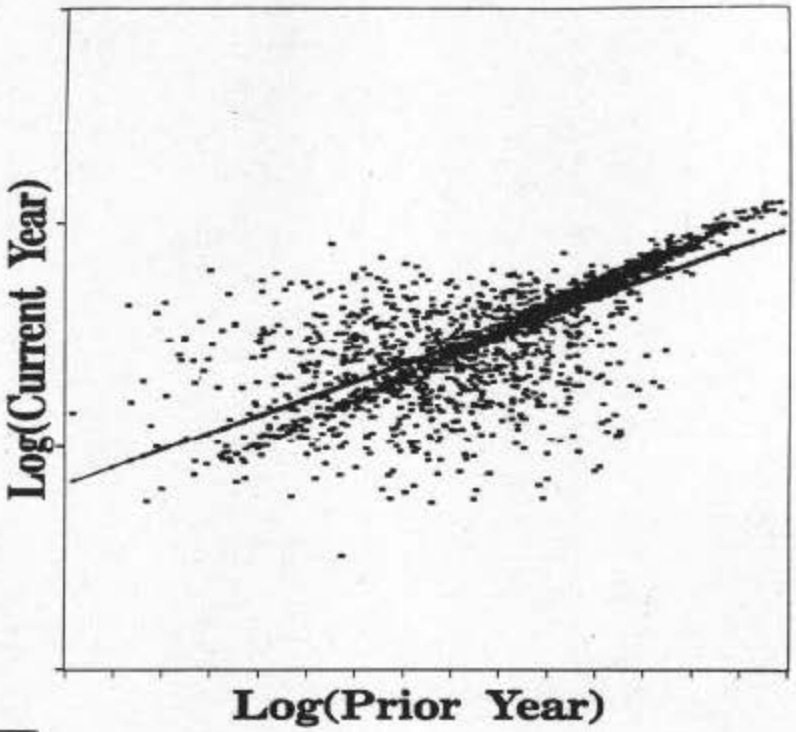

Revenue
Payroll
a
Log(Revenue)
A
Log(Payroll)
b

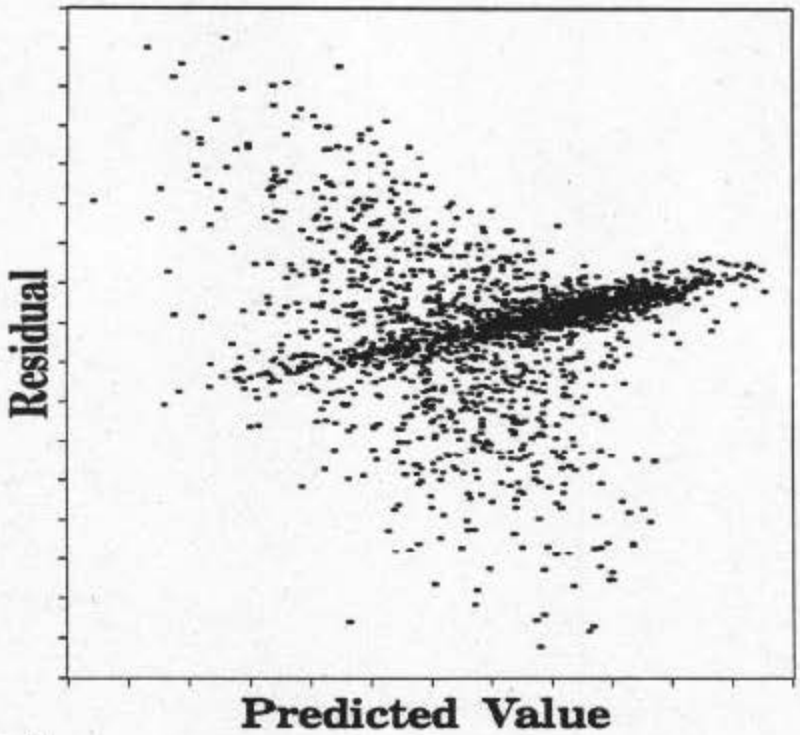
Figure 5


Log(Revenue)
Log(Payroll)
c
0
1
2
3
Absolute Residual
d

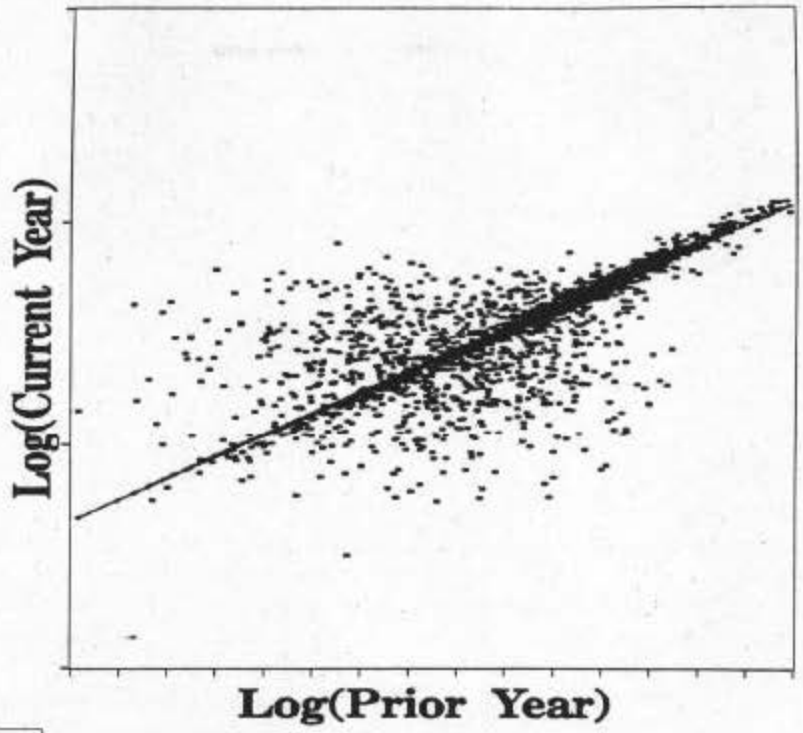


a

Figure 6

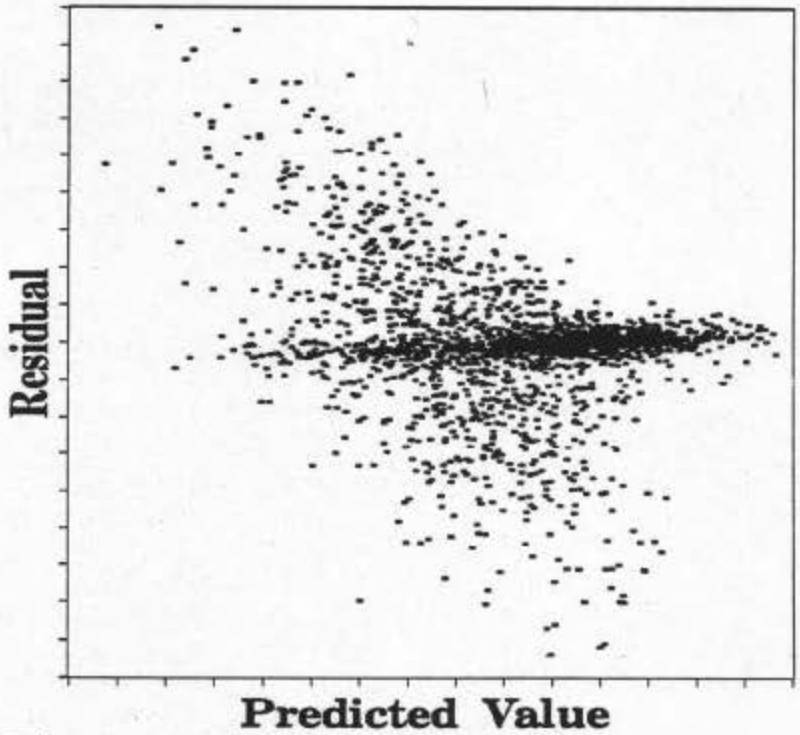


b



a

Figure 7



b

TIME SERIES AND CROSS SECTION EDITS WITH APPLICATIONS TO FEDERAL RESERVE DEPOSIT REPORTS

David A. Pierce and Laura Bauer Gillis[1]
Federal Reserve Board

ABSTRACT

Currently data from the major deposit reports submitted by commercial banks to the Federal Reserve System are edited by comparing the incoming value for a variable to that variable's value for the previous week, using a set of published *tolerances*. The previous value represents an estimate or forecast of what the current value would be in the absence of error or unusual circumstance. This paper investigates two generalizations of this editing method, which both involve incorporating information beyond that contained in the previous week's value. One of these is to base this estimate on the item values from a *cross section* of similar institutions in the current time period which have already reported, and the other is to calculate a forecast based on the *time series* of past values of the item. A composite estimate combining these two methods is also presented. Edit simulations are performed to measure the improvement from this approach (in terms of fewer edit exceptions which are correct and/or increased detection of errors), which is found to be substantial for some items and size groups. Efforts thus far to implement these enhancements are described, and possible further generalizations are mentioned.

1. INTRODUCTION AND SUMMARY

Data for the U.S. Money Supply are regularly transmitted to the Federal Reserve System by commercial banks and other financial institutions at weekly and other intervals. A major vehicle for this transmission is the "Report of Transactions Accounts, Other Deposits and Vault Cash", or simply the "Report of Deposits", on which banks and other financial institutions report weekly data for 25 deposit categories and related items. Based on these data and on similar information contained in other reports, the money supply measures are constructed and reserve requirements are maintained.

The money and reserves figures are important both as barometers of economic activity and in enabling the Federal Reserve to perform its economic stabilization and bank regulatory functions, and it is essential that the data submitted on the Report of Deposits and other reports be reliable and of high quality. To ensure their accuracy, all such data are subjected to numerical edits to detect unusual or deviant values. These edits are to two general types, *validity* edits to ensure that adding-up and other logical constraints are satisfied, and *quality* edits based on statistical or distributional aspects of the data.

---

[1] The authors are respectively Senior Statistician and Statistician, Division of Research & Statistics, Federal Reserve Board, Washington, DC 20551. The valuable assistance of Mia Johnson is gratefully acknowledged. Any views expressed do not necessarily reflect those of the Federal Reserve System.

The most commonly used quality edit involves the comparison of an incoming weekly figure to the previous value of that variable (in both dollar and percentage terms), using a tolerance band constructed about that value. The *tolerances*, or half-widths of the tolerance bands, are determined from previous estimates of the variable's distribution, in particular measures of spread, and are published in a Technical Memorandum or "Tech Memo"[2]. An edit "exception" occurs if the incoming value falls outside this tolerance band; when this happens, the reporting bank or other institution may be contacted for verification or correction. All tolerance-table comparisons are made (and edit exceptions generated) by machine, whereas the decision to contact the respondent is made by data analysts. The editing is done at both the Federal Reserve Board and the 12 Federal Reserve Banks.

Edits are in essence hypothesis tests, and errors of both kinds can occur. A major task in setting edit tolerances is to ensure adequate sensitivity without generating unnecessarily large quantities of "false positive" edit exceptions. It is because of the large number of exceptions currently generated that editing at both the Reserve Banks and the Board is currently quite labor intensive. All exceptions are reviewed by data analysts who must decide which are to be referred to the respondent institution for verification or revision. At the same time, a large majority of the data errors are not caught by these edits, based on the historical record of revisions submitted by respondents (they may be detected by other edits at a later date). There is consequently a need both to increase the sensitivity of the edits and to streamline the data editing process.

The previous value of the variable being edited, to which the tolerances are applied, in effect represents an estimate or forecast of the current figure in the absence of error or unusual circumstance. By basing this forecast or estimate on information beyond that contained in the previous week's value for that variable or item, we obtain the generalizations of the current editing method that are investigated in this paper. One generalization is to base this estimate on the item values from a *cross section* of similar institutions in the current time period which have already reported, intending to capture economic, institutional or calendar movements which tend to affect similar respondents in a similar manner. The other is to calculate a forecast based on the *time series* of past values of the item for that respondent, including possibly last month's or last year's figures in addition to the one for last week as in the current procedure. A composite estimate combining these two methods is also investigated, the idea being that each method may incorporate information not captured by the other. (We also generated a composite of the cross section and current edits).

The paper's focus is on the data submitted on the Report of Deposits, also known as the Edited Data Deposits System (EDDS) Report. We investigated four of the more important items on this report, total transactions deposits, total savings deposits, and large and small time deposits. The study was motivated by the desire for greater automation in the Federal Reserve Board's Division of Information Resources Management, which carries out the edits. The improvements resulting from the study are being incorporated into a new software package called DEEP (Distributed EDDS Editing Project), for interactive editing on the PC.

Our results vary greatly according to item, entity type (e.g. commercial bank, credit union, etc.), and the amount of data in an institution group -- the latter being important for reliable cross-

---

[2] "Processing Procedures for the Report of Transaction Accounts, Other Deposits and Vault Cash (FR2900), Technical Memorandum No. 16, Publications Section, Federal Reserve Board (December 1993).

section estimates. In most cases we find that, with sufficient data, the cross section approach is as reliable as the current editing procedure. For total transactions deposits almost uniformly, and for total savings deposits for most commercial bank categories, time series modelling plays a significant role in the edits.

The following section of the paper discusses in greater detail the methodology underlying the different data editing approaches investigated. Section 3 then describes a set of edit simulations we performed with each of the five types of edits studied, and presents the results of these. Based on the simulation results, we provided a set of recommendations for experimental edits for DEEP, for each entity type and item, which have recently become operational.

2. METHODOLOGY

Given a variable or item of interest, many data editing procedures can be characterized as first generating a forecast (a point estimate) of the incoming value for that item, next applying a tolerance to the forecast to form a tolerance interval (an interval estimate) for the incoming value, and then flagging that value if it is outside the tolerance interval. In the current editing framework, that forecast is taken to be the previous week's item value, and the tolerance is as given by the Tech Memo (footnote 2). In this section the two generalizations to the forecast noted in Section 1 are presented, along with composite procedures, after first describing the data and framework used.

2.1 Choice of Items and Statistical Form

The current approach to editing data from financial institutions is to subdivide them into homogeneous "cells", which are combinations of an institution's size group, entity type, geographic location. There are six size groups for commercial banks and a smaller number of size groups for each of the other entity types, which are credit unions, S&Ls, savings banks, agencies and branches of foreign banks, and Edge and Agreement Corporations. The geographic locations are defined in terms of 12 Federal Reserve districts.

There are thus a great many edit cells, and to make our task manageable, and to achieve comparability with the current edits, we have simplified this study in the following ways:

1. Staying with the *same cells* of the current EDDS edits. This will facilitate assessing the effects of the cross section estimates, model forecasts, and composite procedures. We recognize that more sophisticated groupings into cells may enhance the performance of the edits and plan to work with these in the future. Also we have eliminated all acquisitions and mergers from the institutions studied and have placed "credit-card banks" in a separate group.

2. Maintaining the *same tolerance widths* as currently (applied, however, to the time series / cross section estimates that we generate, as well as to the most recent value as currently done). This may at first seem unnecessary, since standard deviations, percentiles, and other aspects of the distribution can be determined from either the cross section data or the historical model. However, such calculations can sometimes be unreliable, especially with cross sections without at least several hundred institutions

in a group, as we are working with the extremes of distributions. And as with the cells themselves, keeping the current cell tolerance-interval widths facilitates comparisons among procedures.

We have also confined our attention in this study to the smaller institutions ("Priority-3" or P-3 institutions), where there may be the greatest potential for human resource savings from this approach. (Essentially this excludes the largest three size groups for commercial banks and a portion of the largest size group for other entity types). For these institutions, we have examined the following items:

Total transactions deposits	Large time deposits
Savings deposits	Small time deposits.

Current EDDS editing is performed with both dollar and percentage changes of the item being edited, with both required to exceed tolerances ("and" condition) for an exception to occur. The modifications outlined in this report are only for percentage changes; the Tech Memo tolerances continue to be applied to the dollar changes. There are several reasons for choosing percentage changes as the focus. Since they are used in current edits, the present edit cells and tolerances can be employed, and comparisons with current procedures can be made. They (or their annualized versions, growth rates) are also used in other analyses, such as with the Small Bank Sample of early reporting institutions. They are more homogeneous than dollar changes among different sized institutions, so that fewer edit groupings should eventually be needed. Percentage changes were found to be more sensitive to reporting and other errors than ratios to other items such as total deposits, which change with the denominator as well as the numerator and moreover present difficulty when the denominator was zero.

2.2 Cross Section Edits

Period-to-period edits compare an institution's current value for an item to the previous period's value. However, useful additional information may be contained in the current values of that item for other institutions that are similar to the one being edited. For example, if most of the institutions in a group experience a surge in large time deposits in a given week, then it would probably be inaccurate to list them as exceptions simply because they were outside the EDDS tolerances. Conversely, a very small change that week in large time deposits for a particular institution in that group may be suspicious even though current period-to-period tolerances would not be exceeded.

Cross section edits are carried out by examining the distribution of values (here, of percentage changes) for institutions within a homogeneous group, and listing as exceptions any values that were unusual compared to that distribution. Ordinarily one would calculate the mean and standard deviation of the percentage changes and flag those that were farther away from the mean than (say) two or three standard deviations; but in the present study we modified this set-up in two ways. First, because extreme values (the ones we hope to detect) would themselves influence the mean to which they would be compared, we "trimmed" the mean by eliminating the largest and smallest 5% of the values before calculating the estimated mean. Second, more observations are required to form a reliable estimate of the standard deviation than of the mean, and since most of the cells or groupings of institutions were too small for this, we chose to use multiples of the current EDDS tolerances as proxies for the standard deviations. As noted earlier, an additional advantage of this practice is to facilitate comparisons with the current edits.

One difficulty in using a cross section edit is that the data for an editing group need to be available in order to calculate such quantities as the average percentage change for that group. But the data for Priority-3 institutions are not due at the Board until nine days after the as-of date; and since timely estimates of the monetary aggregates and required reserves are needed, the editing process cannot be postponed this long. Our solution to this is to wait until a large enough fraction of the institutions have reported, and to form the distributional estimates (the trimmed means in this case) from the data available at that time.

For the EDDS data, more than half of the P-3 institutions' records are received by the Federal Reserve Board on the Thursday night following the as-of date (the previous Monday, on which the statement week ends), with the majority of those outstanding arriving by Friday night and the few remaining ones by the following Wednesday. For this study it was therefore decided to start the cross section editing on Friday morning, although work in progress is comparing this with the alternative of beginning on Monday morning. In either case, the trimmed mean estimates initially formed are not modified when more institutions have reported, in order not to confuse the editing process.

Some of the editing cells contain only a small number of respondents (and an even smaller number reporting by Friday), so that the estimated mean for those cells may not be very reliable. We required a minimum of 50 available observations in order to use the cross section estimate by itself. If the number of available observations is less than 50 but at least 20, a composite (see Sec. 2.4) of that estimate and the previous week's value for the institution is employed, and with less than 20 the previous week's value alone is used.

The cross section edit is performed by comparing the deviation between the observed and the estimated percentage changes to the current EDDS edit tolerance for the item. As noted earlier, if the percentage-change condition is violated, then a second comparison of the magnitude of the dollar change versus its tolerance is performed, and the item is flagged only if both sets of tolerances are exceeded. An exception to this is that, as is done with the current edits, when the item changes from zero to a nonzero value or vice versa, the current dollar-change edit tolerances are applied without any adjustment.

2.3 Time Series Edits

These edits are based on time series models, which predict or explain an item's present value in terms of its past history. This usually involves the immediately previous value, on which the current edits are based, and often additional values as well, such as last year's. To the extent that these more distant values are important in predicting the incoming value, more sensitive edits should result from taking them into account.

Editing using a time series model for generating forecasts of percentage changes implies that a historical relationship exists between the item and its previous values. The "random walk" model is a time series model in which the best forecast of the current value is simply last week's value. Thus, the random walk model is implied by the current period-to-period change edits, which take last week's value as the current-period forecast around which the tolerances are applied. More complicated time series models yield forecasts which are weighted averages of several past values of the percentage change.

We first investigated the fitting of time series models for each institution separately. Some

institutions' data fit the models quite well, with reductions in the standard deviation of the forecast errors (a key to the effectiveness of tolerances of a given width) of 50% or more, while other institutions exhibited only weak fits, or only the random walk behavior that the current editing framework already captures. Although fitting individual models is the preferable method for forecasting, it was not feasible to maintain over 8000 models for each item edited within the DEEP framework - at least not at this time. Thus, at this stage and for the P-3 institutions, a single time series model was fit to each editing cell's aggregate, and the coefficients from that estimated model were used to obtain an individual bank's forecast using its own previous values. While the benefits of time series modelling are reduced by doing this, the method can be easily implemented, and updated when necessary. Another constraint at present is that, because of data storage limitations, we only utilized terms in the model at lags of 1, 2, 3, 52 and 53 weeks, thus capturing nearby effects and annual seasonal influences but not, say, monthly or quarterly effects.

As an example of the model-fitting results, Table 1 provides information on time series models fit to cell aggregates of Total Transactions Deposits for three of the editing cells. Notice the highly statistically significant seasonal effect (lag 52, and in some cases lag 53). The strength of the fit declines going down the page, with the third one (Edges & Agreements, a root MSE reduction of only 9.2%) being not much different from the random walk model underlying current edits. On the other hand the results suggest that model-based editing may be valuable for certain commercial bank cells, for total transactions.

As with cross section edits, the deviation between the actual percentage change and the forecasted change from the time series model is compared to the edit tolerances. A tolerance exceedance both here and on the dollar change (also using current EDDS tolerances) triggers an edit exception for the record.

2.4 Composite Edits

The cross-section and time series edits are based on different sets of information, past values of the institution being edited and present values of similar institutions. Thus a forecast which combined these two estimates, thereby utilizing both sources of information, may be more accurate than either one separately, and edits derived from such forecasts correspondingly more sensitive.

For a given institution (e.g. bank) and a given item, if T denotes a time-series estimate (forecast) for a given week, C represents a cross-section estimate, and A the actual value that is reported, then the <u>composite estimate</u> is a weighted average of T and C which is of the form

$$\omega T + (1-\omega)C.$$

The weights ω and $1-\omega$ depend on the relative sizes and the correlation between the estimation / forecast errors of T and C. If these errors are given by

$$ET = A - T \text{ and } EC = A - C,$$

then

$$\omega = [Var(EC) - Cov(ET,EC)] / Var(ET-EC).$$

A composite forecast is thus a weighted average of individual component forecasts where the relative weights are chosen to minimize the sum of the squared forecast or estimation errors, and where the

sum of the weights is one.

Using past data, we investigated a composite estimate of the cross section and the time series forecasts, denoted "CSTS", for each editing cell and each item. The composite forecast defaults to the time series forecast with fewer than 20 available observations in the cell average. (With exactly 20 and using the 90% trim, 18 observed changes would be used in the cell estimate).

The other type of composite edit we considered combines the cross section and the random walk forecasts (CSRW). We employed this edit when a CS edit was indicated but the sample size -- the number of observations available on Friday morning when the cell means are formed -- was insufficient (less than 50) to obtain an adequately reliable cross section estimate. For very small sample sizes (less than 20), our procedure is to revert to the use of only the RW edit.

3. MODELLING AND SIMULATIONS

To examine the relative performance of different types of edits, we conducted simulations of these edits over the 1991-92 time period. For each cell (choice of item, entity type, size group and geographic region), we performed five sets of simulations, corresponding to the different types of edits under consideration: current (random walk), cross section, time series, cross section/time series composite, and cross section/random walk composite.

3.1 Simulation Procedure

Data preparation was a time consuming task. First, all Priority-3 reporters' weekly average data were compiled for the period from January 1986 through December 1992. While the edits were simulated only for the most recent two years, the additional data were used for fitting time series models with potential annual patterns. To avoid distortions, we eliminated all banks involved in mergers during this period. We next partitioned the data set into the editing groups or cells. We found that not all cells had a sufficient number of reporters to fit a model or to obtain reliable cross section estimates, and so some of them were combined. For commercial banks of size group 3 (total deposits between $1B+ and $3B), there were too few P-3 reporters to employ any of the new approaches. In addition, we added an editing category for known credit card banks. In total there were 40 edit cells, 37 of which were involved in the simulations.

Once the data were prepared, time series models were fit to the percentage changes in each cell's aggregate, as described in Section 2.3. Using the fitted model for a cell, predicted values for the last two years were generated for each institution in the cell. (Although forecasted values of the percentage change were generated for all periods, those in which a change of zero to a value or a value to zero were edited using the current special tolerances). Both the model-based and the zero-valued random walk forecasts were assigned to each observation in the cell. The 10% trimmed mean of the percentage changes was also calculated for each cell and each week of the two year simulation period, for use in the cross section edits. (Since the cross section simulation employed all the data within a cell to calculate the current-period forecast, rather than the available data as of Friday morning when editing begins, the simulated results will differ from those in practice). In order to generate the two composite forecasts, the prediction errors from the original three forecasts were computed and the formulas in Section 2.4 applied by institution. A cell root mean square prediction

error (RMSE) was also computed.

Since the composite forecast combines the component forecasts in such a way as to minimize the sum of the squared prediction errors, we chose to estimate the appropriate weights for each bank in a cell and then to average those weights over the cell in order to obtain the composite for editing. Since the composite is a weighted average of the individual forecasts, the sum of the weights must equal one. For some institutions, where the prediction errors were very highly correlated between methods, we obtained pairs of weights with one value less than zero and the other greater than one. Evidently it only requires a small number of observations away from that correlation structure to cause such disproportionate weights. In calculating the average pair of composite weights for each cell, therefore, we first screened out those sets of weights not within the (0,1) range. After the two composite weighting schemes were determined for each cell, the mean square prediction errors were computed for these two forecasts as well.

For each of the five edit methods, Table 2 presents the root mean square prediction errors and composite weights for the commercial bank cells for total transactions and total savings, and Table 3 presents the same information for the other entity cells, for total transactions. We anticipate the method with the smallest forecasting error to have the best potential as an edit, but until our tolerances are better tailored to the actual editing method, this potential may not be realized.

To apply the edits, we first looked for percentage changes that differed from the forecasted percentage changes by more than the appropriate tolerance (whether taken from the Tech Memo or generated as described in this paragraph), and for those ascertaining whether the dollar change tolerance was also exceeded. Since total savings and large time deposits are currently edited items, their current tolerances can be used. However, for total transactions and small time deposits, current tolerances do not exist. We therefore generated tolerances in a manner similar to that used for the creation of the current ones. This involved iterative steps with the intent of flagging approximately 0.3% of the observations per cell on average (the maximum percentage of observations flagged using current editing methods for other items, for the year 1991). Using the components of total transactions and items that were related to small time, such as total and large time, we first compiled a range of feasible values for the tolerances. We then examined where these values occurred on the distribution of percentage changes over each cell for the two-year period. Given a reasonable proportion of the changes exceeding the initial values, we then examined the dollar change distribution for the subset of percentage change exceptions. Appropriate percentiles of this distribution were then determined to obtain the expected 0.3% edit failures under the current random walk model. These percentiles became the dollar change tolerances.

Once all the forecasts and tolerances were in place, the editing experience for the 1991-92 period was simulated for each of the five forecast methods. For each method we observed which observations were flagged as edit exceptions. Then based on a history of weekly revisions to the EDDS file maintained by the Federal Reserve's Statistical Services branch, we were able to determine the rate of type I and type II errors for each method. [A type I error (a "false positive") refers to an item that was flagged but not in error, or at least not revised. A type II error occurs when an item is not flagged but is erroneous (as evidenced by a later revision)].

3.2 Simulation Results

For reference in this section, Table 4 shows our recommended edits based on these simulations. As mentioned in Section 1, these are currently being implemented as part of the Federal Reserve Board's DEEP editing software. In Table 4, the left column lists the entities (with the included size groups in parentheses), followed by the chosen edit for each item.

Turning to the results on which this table is based, Table 5 summarizes the editing simulations for commercial banks; those for other entity types were similar and are given in an earlier report[3]. To assess the magnitudes and the implications of errors caught and errors missed by the editing schemes, the tables break down these errors in terms of their size (i.e. the size of the revision--we assume, however accurately, that revised data are correct and the revision is the error in the unrevised data). Each section of these tables compares the current (random walk) method with an alternative editing strategy. It is clear from these simulations that there is room for improvement, especially regarding the type II error probabilities, which range from 98% to 99%. And although the type I error probabilities appear small, the number of flagged items that are not in error is quite large (between 87% and 94%).

Wherever the fitted time series model indicated a potentially substantial payoff relative to the random walk model (as in the first model in Table 1), the time series edit tended to be the most accurate, yielding the smallest number of edit exceptions and with fewer errors missed that were captured by other methods than vice versa. The reduction in the number of edit exceptions was not as great for the CS and CSTS composite methods, but often the composite method caused less of an increase in the type II error probability. The CS and the CSRW composite often mimicked the current RW results. Where there was doubt regarding the preferable edit method, we tended to favor the CS or CSRW -- even when the reduction in RMSE and the number of edit exceptions was small relative to the current (RW) method -- since cross section edits would allow possibly large shifts in behavior for a given week to be incorporated into the editing norm, and the DEEP software is well-suited to this type of edit. Also, we gave some preference to a uniformity of editing method across related cells (e.g. adjacent size groups within an FR region, or like size groups between regions).

For commercial banks, the alternative edits on the whole did quite well. The time series edits for total transactions and total savings were effective in reducing the total number of exceptions while missing only 3 small revisions and actually finding an additional error of over $25M.[4] For the other entity types, total transactions was the only item that allowed for an alternative other than the CSRW method (CSRW was selected for these entity types in place of CS in order to accommodate smaller sample sizes in the preliminary data). Those credit unions and savings institutions which would have more activity in transactions accounts than the other entity types, do exhibit cyclical patterns which the time series model was able to capture (See Table 3.A). Agencies and branches also exhibited

---

[3] "Editing in DEEP: Utilizing Time Series and Cross Section Information", Laura Bauer Gillis and David A. Pierce, Federal Reserve Board, 1993 (preliminary report). Available from the authors.

[4] This revision was generated either by an outside source or by an edit of another report that is not being considered here. This occurrence brings to light that some errors are detected by other sources - not the Reserve Banks or the Board. What we gain from this additional edit exception an earlier detection of the error; it would not necessarily go undetected permanently.

improved editing results with the CSTS method. As mentioned, this combination of alternative strategies yielded an 11% reduction in both the type I error probability and the number of edit exceptions, with only a very slight increase in the type II error likelihood (about 0.1%).

All of these results are based on simulations using 1991 and 1992 EDDS data. Any errors caught before the data arrived at the Board are not reflected in these data, nor are errors undetected by Banks or Board that do not show up in the revision files. And as previously mentioned, the other factor to be monitored is the use of preliminary data in cross section estimates of the mean percentage change. Depending on how and where the preliminary data fall in the distribution of all percentage changes for an item, the operational results based on the CS, CSTS, or CSRW methods may differ significantly from what is expected based on the simulation results. The data availability and timing issue for cross section estimates is currently being studied.

This investigation is still in progress, and further generalizations of the work are underway or planned. Among these are examining time series models with regression components to account for such phenomena as tax dates, calendar effects or related variables, alternative groupings of the data according to size or geographic region, modelling larger banks individually, and examining additional items or variables.

Table 1. Percentage Change Models for Total Transactions Aggregates, Selected Editing Cells

----------Cell = CB, Size Group 4, Region I----------

Root MSE(orig.) = 0.0383 Root MSE(model) = 0.0211
Reduction in Root MSE = 44.9%

Variable	Parameter Estimate	Standard Error	T-stat	p-value
TRN_{t-1}	-0.4349	0.0483	-9.005	0.0001
TRN_{t-2}	-0.0341	0.0329	-1.039	0.2996
TRN_{t-3}	-0.1510	0.0338	-4.467	0.0001
TRN_{t-52}	0.6494	0.0318	20.391	0.0001
TRN_{t-53}	0.4668	0.0440	10.606	0.0001

----------Cell = CU, Size Group 2, Regions II&III----------

Root MSE(orig.) = 0.1067 Root MSE(model)=0.0809
Reduction in Root MSE = 24.2%

Variable	Parameter Estimate	Standard Error	T-stat	p-value
TRN_{t-1}	-0.2450	0.0546	-4.486	0.0001
TRN_{t-2}	-0.1160	0.0474	-2.444	0.0151
TRN_{t-3}	-0.2200	0.0486	-4.525	0.0001
TRN_{t-52}	0.4922	0.0477	10.312	0.0001
TRN_{t-53}	0.1866	0.0533	3.498	0.0005

----------Cell = EA, All----------

Root MSE(orig.) = 0.0564 Root MSE(model)=0.0512
Reduction in Root MSE = 9.2%

Variable	Parameter Estimate	Standard Error	T-stat	p-value
TRN_{t-1}	-0.3776	0.0569	-6.632	0.0001
TRN_{t-2}	-0.1547	0.0586	-2.642	0.0087
TRN_{t-3}	-0.0449	0.0553	-0.815	0.4181
TRN_{t-52}	0.2432	0.0524	4.638	0.0001
TRN_{t-53}	0.1057	0.0540	1.955	0.0514

Table 2. Root Mean Square Errors for Forecasts: Commercial Bank Cells

A. Total Transactions	Root Mean Square Error					Weight of CS in Composite	
Cell	RW	CS	TS	CSRW	CSTS	CSRW	CSTS
Region 1							
-Size 4	0.077	0.073	0.077	0.074	0.071	0.72	0.51
-Size 5	0.096	0.094	0.097	0.094	0.089	0.73	0.55
-Size 6	1.190	1.190	1.276	1.190	1.204	0.70	0.58
Region 2							
-Size 4	0.064	0.059	0.236	0.060	0.121	0.77	0.58
-Size 5	0.210	0.209	0.223	0.209	0.212	0.62	0.55
-Size 6	0.331	0.330	0.344	0.330	0.333	0.68	0.57
Region 3							
-Size 4	0.102	0.099	0.108	0.100	0.100	0.75	0.51
-Size 5	0.054	0.048	0.051	0.050	0.046	0.74	0.58
-Size 6	0.067	0.063	0.071	0.064	0.062	0.70	0.60

B. Total Savings	Root Mean Square Error					Weight of CS in Composite	
Cell	RW	CS	TS	CSRW	CSTS	CSRW	CSTS
Region 1							
-Size 4	0.042	0.042	0.045	0.042	0.042	0.64	0.73
-Size 5	0.054	0.054	0.056	0.054	0.054	0.64	0.67
-Size 6	0.048	0.048	0.055	0.048	0.048	0.60	0.72
Region 2							
-Size 4	0.038	0.038	0.099	0.038	0.043	0.65	0.76
-Size 5	0.235	0.234	0.244	0.234	0.236	0.64	0.64
-Size 6	0.055	0.055	0.067	0.055	0.055	0.64	0.66
Region 3							
-Size 4	0.051	0.051	0.998	0.051	0.274	0.68	0.74
-Size 5	0.041	0.040	0.041	0.040	0.040	0.63	0.66
-Size 6	0.055	0.055	0.065	0.055	0.055	0.61	0.75

Table 2. Root Mean Square Errors for Forecasts: Commercial Bank Cells (Continued)

C. Large Time

Cell	Root Mean Square Error					Weight of CS in Composite	
	RW	CS	TS	CSRW	CSTS	CSRW	CSTS
Region 1							
-Size 4	0.067	0.067	0.069	0.067	0.067	0.53	0.61
-Size 5	0.110	0.110	0.117	0.110	0.110	0.52	0.75
-Size 6	0.160	0.160	0.184	0.160	0.161	0.49	0.81
Region 2							
-Size 4	0.089	0.088	0.093	0.088	0.089	0.54	0.68
-Size 5	0.063	0.063	0.065	0.063	0.063	0.48	0.62
-Size 6	0.099	0.099	0.109	0.099	0.100	0.46	0.76
Region 3							
-Size 4	0.047	0.047	0.051	0.047	0.047	0.55	0.70
-Size 5	0.075	0.075	0.076	0.075	0.076	0.54	0.59
-Size 6	0.120	0.120	0.141	0.120	0.120	0.51	0.80

D. Small Time

Cell	Root Mean Square Error					Weight of CS in Composite	
	RW	CS	TS	CSRW	CSTS	CSRW	CSTS
Region 1							
-Size 4	0.064	0.064	0.098	0.064	0.068	0.58	0.70
-Size 5	0.143	0.143	0.156	0.143	0.144	0.52	0.70
-Size 6	0.110	0.110	2.274	0.110	0.409	0.55	0.83
Region 2							
-Size 4	0.468	0.468	0.516	0.468	0.470	0.59	0.79
-Size 5	1.363	1.363	1.420	1.363	1.369	0.61	0.68
-Size 6	0.034	0.034	0.036	0.034	0.034	0.58	0.77
Region 3							
-Size 4	0.062	0.062	0.068	0.062	0.063	0.61	0.67
-Size 5	0.017	0.017	0.018	0.017	0.017	0.58	0.65
-Size 6	0.063	0.063	0.072	0.063	0.063	0.57	0.80

Table 3. Root Mean Square Errors for Forecasts: Other Entity Types, Total Transactions

A. Agencies and Branches

Cell	Root Mean Square Error					Weight of CS in Composite	
	RW	CS	TS	CSRW	CSTS	CSRW	CSTS
-All Regions							
Size 1	1.366	1.363	1.378	1.364	1.364	0.48	0.54
-Region 1							
Size 2	2.700	2.696	2.794	2.698	2.715	0.45	0.53
Size 3	5.061	5.061	5.974	5.061	5.198	0.38	0.62
-Region 2							
Size 2	2.248	2.240	2.406	2.245	2.317	0.38	0.35
Size 3	4.158	4.154	4.965	4.156	4.248	0.43	0.67
-Region 3							
Size 2	0.250	0.247	0.250	0.248	0.243	0.34	0.44
Size 3	4.300	4.289	5.416	4.295	4.544	0.42	0.58

B. Credit Unions

Cell	Root Mean Square Error					Weight of CS in Composite	
	RW	CS	TS	CSRW	CSTS	CSRW	CSTS
-All Regions							
Size 1	0.106	0.073	0.075	0.076	0.069	0.75	0.53
-Region 1							
Size 2	0.093	0.069	0.059	0.075	0.059	0.57	0.37
Size 3	0.122	0.112	0.120	0.114	0.110	0.59	0.43
Size 4	0.084	0.075	0.078	0.077	0.073	0.57	0.45
-Regions 2 & 3							
Size 2	0.084	0.062	0.054	0.065	0.052	0.64	0.57
Size 3	0.099	0.080	0.078	0.083	0.073	0.63	0.37
Size 4	0.082	0.069	0.060	0.072	0.059	0.56	0.40

Table 3. Root Mean Square Errors for Forecasts: Other Entity Types, Total Transactions (Continued)

C. Edges and Agreements

Cell	Root Mean Square Error					Weight of CS in Composite	
	RW	CS	TS	CSRW	CSTS	CSRW	CSTS
-ALL	17.21	17.20	18.94	17.20	17.72	0.44	0.47

D. Savings Institutions

Cell	Root Mean Square Error					Weight of CS in Composite	
	RW	CS	TS	CSRW	CSTS	CSRW	CSTS
-Region 1							
Size 1	0.057	0.048	0.049	0.049	0.047	0.79	0.67
Size 2	0.187	0.185	0.193	0.185	0.185	0.74	0.57
Size 3	0.744	0.743	0.965	0.743	0.779	0.71	0.61
Size 4	0.627	0.626	0.645	0.626	0.627	0.66	0.63
-Regions 2 & 3							
Size 1	0.073	0.065	0.068	0.065	0.064	0.73	0.68
-Region 2							
Size 2	0.132	0.129	0.153	0.129	0.131	0.73	0.62
Size 3	0.077	0.072	0.079	0.073	0.072	0.78	0.66
Size 4	0.066	0.062	0.069	0.062	0.061	0.75	0.66
-Region 3							
Size 2	0.077	0.069	0.077	0.070	0.068	0.78	0.58
Size 3	0.309	0.308	0.766	0.308	0.377	0.72	0.66
Size 4	0.370	0.369	0.568	0.369	0.408	0.63	0.59
-Region 4							
Size 1	10.73	10.73	11.44	10.73	10.77	0.67	0.75

Table 4. Experimental Edits for DEEP

	Total Transactions	Total Savings	Large Time	Small Time
Commercial Banks (3,Ccd)	RW	RW	RW	RW
Commercial Banks (4,5,6)	CSTS	TS	CS	CS
Credit Unions (1,2,3,4)	TS	CSRW	CSRW	CSRW
S&Ls, Coops, Sbs (1,2,3,4)	𝔯I TS / 𝔯II-IV CSTS	CSRW	CSRW	CSRW
Agencies & Brs.(1,2,3)	CSTS	CSRW	CSRW	CSRW
Edges & Agr. (1,2)	CSRW	CSRW	CSRW	CSRW

The numbers in parentheses are the size groups, with "Ccd" denoting credit card banks. CB size groups 1 and 2 are omitted, as they are priority 1 and 2 institutions. 𝔯 denotes the FR Region, as in TM#16. The other entries in this table have the following explanations:

TS: The time-series model-based forecast, utilizing the institution's past percentage changes (of 1,2,3,52 and 53 weeks ago).

CS: The cross-section forecast, or estimate of the average percentage change over all the institutions in the editing group or cell. Uses only the data received by the Friday after the as-of date and is calculated as the 90% trimmed mean of the individual percentage changes in the cell.

CSTS: A weighted average of the TS and CS percentage-change forecasts, with statistically determined weights. When the number (n) of institutions in the group available on Friday for calculating the mean is less than 20, the weights are 1 and 0 (only the TS forecast is used).

RW: The forecast based on the "random walk" model, or the time series model giving a zero period-to-period change as the best forecast -- and is thus the implicit model underlying the current edits. This translates into a percentage-change forecast of zero.

CSRW: The forecast based on a composite of the CS and RW estimates of the percentage change, again depending on the number n of available observations in the cell. Thus:

if $n \geq 50$, use CS only;
if $20 \leq n < 50$, use weighted average of the CS and RW estimates;
if $n < 20$, use the RW estimate (zero percentage change forecast).

Table 5. Editing Simulation Results: Commercial Banks

A. Total Transactions

1. Random Walk (Current Editing)

Frequency/ Percent	Not Revised	<$5M	$5M <$10M	$10M < $25M	> $25M	Total
Not Flagged	557,166 97.76	9,732 1.71	79[illegible] 0.14	508 0.09	168 0.03	568,365 99.73
Flagged	1,444 0.25	75 0.01	17 0.00	12 0.00	6 0.00	1,554 0.27
Total	558,610 98.01	9,807 1.72	808 0.14	520 0.09	174 0.03	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.26%
Pr(type II error) = Pr(Do not Flag Item | Item in error) = 99.0%
Pr(Item not in error | Item Flagged) = 92.9%

2. Cross Section - Time Series Composite

Frequency/ Percent	Not Revised	<$5M	$5M <$10M	$10M < $25M	> $25M	Total
Not Flagged	557,326 97.78	9,743 1.71	792 0.14	509 0.09	167 0.03	568,537 99.76
Flagged	1,284 0.23	61 0.01	16 0.00	11 0.00	7 0.00	1,382 0.24
Total	558,610 98.01	9,807 1.72	808 0.14	520 0.09	174 0.03	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.23%
Pr(type II error) = Pr(Do not Flag Item | Item in error) = 99.1%
Pr(Item not in error | Item Flagged) = 92.9%

Reduction in edit exceptions = 11.1%
Reduction in type I error probability = 11.5%
Increase in type II error probability = 0.1%

Table 5. Editing Simulation Results: Commercial Banks (Continued)

B. Total Savings

1. Random Walk (Current Editing)

Frequency/ Percent	Not Revised	<$5M	$5M <$10M	$10M < $25M	> $25M	Total
Not Flagged	557,547 97.83	8,772 1.54	723 0.13	375 0.07	181 0.03	567 598 99.59
Flagged	2,176 0.38	91 0.02	22 0.00	18 0.00	14 0.00	2,321 0.41
Total	559,723 98.21	8,863 1.56	745 0.13	393 0.07	195 0.03	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.39%
Pr(type II error) = Pr(Do not Flag Item | Item in error) = 98.6%
Pr(Item not in error | Item Flagged) = 93.8%

2. Time Series

Frequency Percent	Not Revised	<$5M	$5M <$10M	$10M < $25M	> $25M	Total
Not Flagged	557,743 97.86	8,775 1.54	723 0.13	376 0.07	181 0.03	567,798 99.63
Flagged	1,980 0.35	88 0.02	22 0.00	17 0.00	14 0.00	2,121 0.37
Total	559,723 98.21	8,863 1.56	745 0.13	393 0.07	195 0.03	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.35%
Pr(type II error) = Pr(Do not Flag Item | Item in error) = 98.6%
Pr(Item not in error | Item Flagged) = 93.4%

Reduction in edit exceptions = 9.8%
Reduction in type I error probability = 10.2%
Increase in type II error probability =. 0.0%

Table 5. Editing Simulation Results: Commercial Banks (Continued)

C. Large Time

1. Random Walk (Current Editing)

Frequency/ Percent	Not Revised	<$5M	$5M <$10M	$10M < $25M	> $25M	Total
Not Flagged	558,956 98.08	8,494 1.49	601 0.10	345 0.06	179 0.03	568,575 99.76
Flagged	1,248 0.22	68 0.01	19 0.00	8 0.00	1 0.00	1,344 0.24
Total	560,204 98.30	8,562 1.50	620 0.10	353 0.06	180 0.03	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.22%
Pr(type II error) = Pr(Do not Flag Item | Item in error) = 99.0%
Pr(Item not in error | Item Flagged) = 92.8%

2. Cross Section

Frequency Percent	Not Revised	<$5M	$5M <$10M	$10M < $25M	> $25M	Total
Not Flagged	558,967 98.08	8,494 1.49	601 0.10	345 0.06	179 0.03	568,586 99.76
Flagged	1,237 0.22	68 0.01	19 0.00	8 0.00	1 0.00	1,333 0.24
Total	560,204 98.30	8,562 1.50	620 0.10	353 0.06	180 0.03	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.22%
Pr(type II error) = Pr(Do not Flag Item | Item in error) = 99.0%
Pr(Item not in error | Item Flagged) = 92.8%

Reduction in edit exceptions = 0.8%
Reduction in type I error probability =. 0.0%
Increase in type II error probability = 0.0%

Table 5. Editing Simulation Results: Commercial Banks (Continued)

D. Small Time

1. Random Walk (Current Editing)

Frequency/ Percent	Not Revised	<$5M	$5M <$10M	$10M < $25M	> $25M	Total
Not Flagged	556,637 97.67	9,869 1.73	1,007 0.18	479 0.08	215 0.04	568,210 99.70
Flagged	1,496 0.26	117 0.02	42 0.01	36 0.01	18 0.00	1,709 0.30
Total	558,138 97.93	9,986 1.75	1049 0.18	515 0.09	233 0.05	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.27%
Pr(type II error) = Pr(Do not Flag Item | Item in error) = 98.2%
Pr(Item not in error | Item Flagged) = 87.5%

2. Cross Section

Frequency Percent	Not Revised	<$5M	$5M <$10M	$10M < $25M	> $25M	Total
Not Flagged	556,637 97.67	9,869 1.73	1,008 0.18	479 0.08	215 0.04	568,211 99.70
Flagged	1,496 0.26	117 0.02	41 0.01	36 0.01	18 0.00	1,708 0.30
Total	558,138 97.93	9,986 1.75	1049 0.18	515 0.09	233 0.05	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.27%
Pr(type II error) = Pr(Do not Flag Item | Item in error) = 98.2%
Pr(Item not in error | Item Flagged) = 87.6%

Reduction in edit exceptions = 0.0%
Reduction in type I error probability = 0.0%
Increase in type II error probability = 0.0%

DISCUSSION

Sandra A. West
U.S. Bureau of Labor Statistics

Let me first commend both sets of authors on very interesting and informative papers. Let me start with the David Pierce and Laura Bauer Gillis paper, "Time Series and Cross Section Edits with Applications to Federal Reserve Deposit Reports."

I enjoyed this paper very much; it was nice to see editing formally enter the realm of statistical inference. One might think of Imputation as point estimation, and editing as interval estimation--or perhaps as multiple imputation. I'd like to focus on one of the editing techniques in terms of imputation, but first let me briefly summarize the study.

In the paper there are:

5 methods for editing percent changes

1. Assuming no change from last week--current method-random walk, **RW**,---would be called Carry Over in nonresponse literature.
2. Using a cross section, **CS**, of similar respondents in the current time period which have already reported. Underlying assumption here that the previous time period values are available. (For surveys that do have nonresponse, only those entities that have reported in both time periods would be used.)
3. Using a time series, **TS**, of the past values of the respondent.
4. Composite of 1 & 2, **CSRW**.
5. Composite of 2 & 3, **CSTS**.

Several entity types-Respondents

Commercial Banks (There were two categories of this type.)
Agencies and Branches
Credit Unions
Edges and Agreements
Savings Institutions

Although there are 25 variables collected, the following 4 were studied.

Variables Collected from each Respondent

Total Transactions
Total Saving
Large Time
Small Time

Edits are performed weekly for a span of two years

Edits are performed in homogeneous cells which are combinations respondents' size, type and geographic location.

I'd like to discuss the cross sectional estimator, CS, since I've had some experience with a similar one using BLS data in terms of imputation. First, I need some notation. In a given cell, let

$Y_{t,i}$ = level of item for entity i at time t.

$\hat{Y}_{t,i}$ = predicted level for entity i at time t.

Editing is performed with percentage changes of the item; that is,

$$D_i = \frac{Y_{t,i} - Y_{(t-1),i}}{Y_{(t-1),i}}$$

(and for the current method the actual changes are also required to be in the tolerance interval).

For the CS edits, the empirical distribution is formed for the percentage changes, and the trimmed mean is computed, where 5% of each tail is trimmed. (Later I will say something about the trimmed mean.) Multiples of the current tolerances are used for proxies for the standard deviation. We could write the trimmed mean as

$$\overline{D} = \frac{1}{m}\sum_{i \in M} \frac{Y_{t,i} - Y_{(t-1),i}}{Y_{(t-1),i}} = \frac{1}{m}\sum_{i \in M} \frac{Y_{t,i}}{Y_{(t-1),i}} - 1$$

where M denotes the set of entities for which the percentage changes are in the middle 90% of the distribution, and m is the number of elements in M.

Using this technique, I'd like to come up with an imputation method. If we let $\hat{Y}_{t,j}$ be the predicted value for the j th entity at time t, and we estimate the percentage change by the trimmed mean, we have the following formula:

$$\frac{\hat{Y}_{t,j} - Y_{(t-1),j}}{Y_{(t-1),j}} = \overline{D} = \frac{1}{m}\sum_{i \in M} \frac{Y_{t,i}}{Y_{(t-1),i}} - 1$$

which leads to $\hat{Y}_{t,j}$:

$$\hat{Y}_{t,j} = \overline{D}\, Y_{(t-1),j} + Y_{(t-1),j} = \frac{1}{m}\sum_{i \in M} \frac{Y_{t,i}}{Y_{(t-1),i}}\; Y_{(t-1),j}.$$

Thus, for the j th entity, we would use his previous time period value, adjust it by the mean ratio of those entities that have already responded, to obtain his current value.

Now looking at this from a regression point of view, consider:

$$Y_{t,j} = \beta Y_{(t-1),j} + \varepsilon_{t,j} \quad where\ \varepsilon_{t,j} \sim N(0, \sigma^2 Y^{\delta}_{(t-1),j})$$
$$for\ j \in M$$

Using a weighted least squares the predicted value is:

$$\hat{Y}_{t,j} = \hat{\beta} Y_{(t-1),j} \quad for\ j \in M$$

where

if $\delta = 1$
$$\hat{\beta} = \frac{\sum_{i \in M} Y_{t,i}}{\sum_{i \in M} Y_{(t-1),i}}$$

if $\delta = 2$
$$\hat{\beta} = \frac{1}{m} \sum_{i \in M} \frac{Y_{t,i}}{Y_{(t-1),i}} .$$

Both are unbiased estimators of β but the one that is more precise depends on the value of δ. $\delta = 2$ is what underlies the CS method.

Under many situations one can show that the sum of the ratios has better properties than the ratio of the sums. However, in a study we did at BLS considering alternative imputation methods, we found that the model with $\delta = 1$ did the best. This was a study involving employment and wage variables for establishments on the Universe Data Base. We investigated many different methods; among them was a generalized Bayesian model, which led to multiple imputation. We also considered a time series going back a year, but only the prior month in this simple model was needed.

$$Y_{t,j} = \beta Y_{(t-1),j} + \varepsilon_{t,j} \quad where\ \varepsilon_{t,j} \sim N(0, \sigma^2 Y_{(t-1),j})$$
$$for\ j \in M$$

Using a weighted least squares, the predicted value is

$$\hat{Y}_{t,j} = \hat{\beta} Y_{(t-1),j} \quad for\ j \in M$$

where

$$\hat{\beta} = \frac{\sum_{i \in M} Y_{t,i}}{\sum_{i \in M} Y_{(t-1),i}}.$$

The M in this case was a set of homogeneous establishments that had reported values in both time periods. For establishment j in time period t, $Y_{t,j}$ denoted, in various studies, the reported employment, the reported ln(wages), and the reported ln(wages/employment). I'm not sure which model would work best with Bank type data, but I think it's worth writing down the underlying models so they can be tested.

In imputation studies, we have a problem similar to one that exists with the CS method. In imputation, when modeling the respondents to predict for the nonrespondents, one hopes the nonrespondents are missing at random; that is, the nonresponse mechanism is ignorable. If this is not the case, it is a difficult problem to model the response mechanism. A similar situation arises with the CS method, in that the edit criteria are set by the early arrivals, and it is hoped that the respondents that are due late, behave in a similar fashion.

I have a couple of observations from the Tables. There were 24 edit groups, consisting of the 6 types of respondents and the 4 variables. Of these 24, for more than half (13), the recommended procedure is the composite of CS and RW. In most of the cases, the CS had the larger weight than RW. Clearly, some form of the CS technique is worth pursuing.

I note from Tables 5 and 6 that the probability of a type II error is very large, and in some situations the probability of a type I error is also large, but not as large--70's versus 90's. I would think that the type II error, not flagging a value when it's in error, is more important than the type I error, flagging a value when it's not in error. But from an analysts point of view, I can see that the type I error would be more important.

Now let me discuss the Julia Bienias, David Lassman, Scott Scheleur, and Howard Hogan Paper, "Improving Outlier Detection in Two Establishment Surveys." I also enjoyed this paper. First a brief summary of the paper.

This paper uses Exploratory Data Analysis (EDA) to improve the detection of outliers in the following two establishment surveys.

1. **The Annual Survey of Communication Services**---2000 firms
2. **The Monthly Wholesale Trade Survey**---7,000 firms, only 3,500 receive forms in a given month

Techniques discussed:

Box Plots
Scatter Plots
Transformations
Fitting: Ordinary Least Squares
Weighted Least Squares

As I mentioned earlier, in our imputation study for wages, we found that if we first transformed the data by the natural logarithm, and used a weighted least squares, the imputation improved.

In general, I believe EDA should be part of any outlier detection system. There is an extensive literature on testing for outliers. A number of popular procedures have difficulty when a sample may contain multiple outliers. Problems include **masking**, in which the presence of other outliers makes each outlier difficult to detect, and **swamping** in which the

procedure tends to declare too many outliers. By using robust and resistant methods it is possible to minimize the effects of deviate observations. An example is given in the Hoaglin, Iglewicz, and Tukey, JASA, 1986 paper, "Performance of some Resistant Rules for Outlier Labeling". Here you have inner and outer fences with hinges formed by the lower and upper fourths. That is, using the lower and upper fourths, F_L and F_U, the inner rule labels as "outside" any observations below $F_L - 1.5(F_U - F_L)$ or above $F_U + 1.5(F_U - F_L)$. For the outer rule 1.5 is replaced by 3.

In comparing the two papers, I found that I would like to combine them. For example, in the cross sectional estimates of Pierce and Gillis, additional EDA techniques could be used. As an example, instead of trimmed means one might consider "adaptive trimmed means". Some "adaptive trimmed means" determine the amount of trimming according to a sample estimate of the tail heaviness of the underlying distribution. This is especially useful if the distributions are not symmetrical, which I assume is the case with bank data.

In closing, I 'd like to compliment the authors for very fine papers.

DISCUSSION

Brian V. Greenberg
U. S. Bureau of the Census

In this discussion we attempt to relate these two fine editing papers to the broader issues in data editing and highlight what one can learn from them.

1. Introduction--Role of Editing

Broadly speaking, there are two primary reasons for editing survey and census data. First, we would like to remove erroneous values from micro-data sets. A second, and related objective, is to ensure that we can generate meaningful estimates from reported data.

For some programs, there is an emphasis on actual micro-level data. For example, when one establishes a longitudinal data file or when a public-use micro-data file will be the primary survey data product. An example of a longitudinal micro-data file is described in the Pierce and Gillis paper and their edit activities focus on the underlying data set.

On the other hand, for some surveys there is a single estimate (or small number of estimates) produced from a survey, and the underlying data file is less important then the single estimate. The Census Bureau's monthly report of wholesale trade, as discussed in the Bienias, Lassman, Scheleur, and Hogan paper, is an example of such a survey.

In any event, data editing does not exist in a vacuum, and in designing and evaluating an edit system, one should be mindful of the survey's data collection and release objectives.

2. Editing Stages in Data Collection and Tabulation

There are typically three stages of data editing for a typical survey or census: (1) data entry edit, (2) automated batch edit of individual data records, and (3) review of summary tabulations.

In the data entry stage, editing often consist of rudimentary checks that only attempt to detect keying errors and major reporting problems. There are, however, data entry programs which have sophisticated and extensive data edit capabilities. For some surveys, editing in the data entry stage, including on-the-spot follow-up with respondents, serves as the primary edit activity.

Batch editing of individual data records, referred to as micro-editing, has been the mainstay of many large-scale survey and census programs. For some surveys, the automated program alters suspicious values, while for others the automated edit only flags suspicious values for analyst review and action. In addition, automated batch edit systems often impute for missing values.

After preliminary editing, data are tabulated and estimates are edited against prior time periods, against information from other sources, or against one another. The process of editing tabulation cells is often referred to as macro-editing. If a tabulation cell looks suspicious, it is reviewed and the individual micro-level records contributing to the cell are examined. Some programs have very sophisticated macro-edit systems while other macro-edit programs are essentially manual.

It is important to note that even though potential data errors are detected at the macro-level in a tabulation cell edit, problems are typically resolved at the micro-level.

After data are processed through automated edit programs, there is typically analyst review of large and/or important cases which often include direct follow-up with respondents. For large programs, there are not sufficient resources to review all records, therefore records are ranked by importance and those most important to a program are reviewed by analysts. The ranking process is often informal, however, research at Statistics Canada to formalize this process (referred to as selective editing) seems to have met with success for their Annual Survey of Manufactures.

All three edit stages come into play in virtually all survey programs. Emphasis on one stage or another is typically embedded in the edit strategy of each program, bearing in mind the proposed uses of survey products.

3. Edit Tolerances

At each stage of the editing process, edit tolerances are required to target individual records or tabulation cells for review and, if necessary, correction. In many respects, the two papers discussed here focus on deriving edit tolerances and we discuss them from this perspective.

There are several steps to deriving meaningful edit tolerances. First, one defines edit cells; that is, groups of respondents whose behavior is fairly similar with respect to the edit criteria. One typically wants cells to be small enough so that respondents can be relatively homogeneous yet large enough so that parameters are not unduly influenced by a few nontypical responses.

One can generate (explicitly or implicitly) an anticipated value for data fields or a relation between fields. The anticipated value may be based on data from the current or prior time periods and can be modeled based on all respondents in the edit cell. Tolerance limits are applied to target records which have unacceptable deviations from anticipated values. For example, anticipated values can be based on a regression line and the tolerance limits can reflect the allowable band of values about this line. Under alternative approaches to deriving tolerances one directly determines a range of acceptable data values and designates response combinations outside that range as edit failures.

Tolerance parameters are derived and applied at each of the three stages of the edit process: data entry edit, batch edit of individual records and tabulation cell edit.

User-friendly systems to support the review of data in the development of edit tolerances can be extremely valuable. It is in this light that the work of Bienias, Lassman, Scheleur, and Hogan can be viewed. The graphics techniques which they present are particularly important because they can help users organize and systemize information and share findings with others. Such systems provide analysts access to methodology not otherwise readily available to them.

4. Striking a Balance in Edit Tolerances and Review Criteria

If edit tolerances are too tight, excessive data may be altered or sent for analyst review. In the first case, edit programs can distort estimates and force data to conform to expectations. In the second, too many referrals place a major burden on analyst resources.

If edit bounds are too loose, erroneous data gets into the system. Such errors in data limit the usefulness of micro-data and may lead to unreliable estimates. Broadly speaking, parameters which are too loose deprive us of the chance to identify a source of nonsampling error.

After automated edit programs have applied tolerances and targeted records as suspicious, one would like to select the most significant problematic records for analyst review. For each survey, one needs a reliable criterion as to what is significant and what is not. The notion of significant depends a great deal on the proposed user of the survey data.

Recently, the phrase "over-editing" has come into vogue to refer to spending too much time and money on editing and/or changing too much data. I feel somewhat uncomfortable with this phrase because it is unfocused and gives a misleading impression. It seems to imply that if we edited less--perhaps had looser edit bounds or reviewed less micro-data--we would be editing better. In fact, we want to edit more cleverly, not necessarily more or less. That is, we would like to target fields for change and/or review where change is needed and not target fields for change and/or review when not needed.

A more useful formulation of the issue can be couched in terms of Type I or Type II error for the edit process, as was done in the Pierce-Gillis paper. Namely, for their purposes:

> A Type I error (a 'false positive') refers to an item that was flagged but was not in error, or at least not revised.
>
> A Type II error occurs when an item is not flagged but is erroneous (as evidenced by a later revision).

We can broaden their definition a little to say:

> Type I error refers to an item flagged for change or review but the time spent on it did not improve the data set or estimates for the survey.
>
> Type II error occurs when an erroneous value, which adversely affect the quality of the data set or survey estimates, is not flagged for change or review.

The last conference on Statistical Policy Working Papers of the Federal Committee on Statistical Methodology was held in March, 1991. At that conference, there was a session based on Working Paper #18, "Data Editing in Federal Agencies." In that report, the focus was on development of multipurpose systems, software design, and edit methodologies. There was little discussion of parameter development. Since then, it has been increasingly clear that good parameter development is crucial in all stages of editing. It is also clear that we need to give greater attention to the interplay between subject-matter staff and automated programs in the resolution of edit failures and in the design of edit tolerances.

Even the best edit methodology embedded in the finest system will perform poorly if there are bad parameters. In fact, the choice of edit tolerances has a major influence on the Type I and Type II error for editing. We certainly need more investigation to highlight what methods and tools work well for the design of edit tolerances and we need to examine and learn from clearly presented case studies.

The two papers under discussion do an excellent job in addressing these related issues.

5. Bienias, Lassman, Scheleur, and Hogan Paper

The authors illustrate graphic techniques used in the spirit of exploratory data analysis as tools for subject-matter specialists in deriving edit parameters. They also describe how the simultaneous review of survey data can introduce advantages over a case-by-case analysis of report forms.

Box plots were used to review and summarize information and directly contributed to parameter development for the Annual Survey of Communication Services (ASCS). In particular, the box plot for parameters based on the expense/revenue ratio illustrates this use.

Graphics were also used to help uncover similarities or differences between establishments. By examining residuals in the relation between revenue and payroll in ASCS, they decided to remove tax-exempt establishments from edit cells for revenue and payroll. That is, they were able to design a more effective edit cell for subsequent analysis, which they describe.

And finally, the graphics and exploratory data analysis led to a more suitable editing model for current to prior inventory on the Monthly Wholesale Trade Survey. In this usage, the techniques they employed allowed subject-matter analysts to experiment with different models and to select the model that they felt best represented the data.

An important theme of this paper was the interplay between subject-matter analyst expertise and the use of graphical methods. These tools can provide a guide for analysts and allow them to see the impact of proposed models. They can contribute to the design process and help eliminate some of the more tenuous aspects of model description. In addition, the graphs provide a useful vehicle for improved communication and shared information among those working on a project.

By all accounts, the survey analysts and project managers who work on the surveys cited above found the contributions described in this paper extremely valuable. It will be through continued and expanded use that additional benefits and applications will arise.

6. Pierce and Gillis Paper

This paper is a superb case study for the development of effective edit cells and related tolerances. This report can be a textbook study. One of the author's primary objectives is stated clearly at the onset: "A major task in setting edit tolerances is to ensure sensitivity without generating unnecessarily large quantities of 'false positive' exceptions." They have an excellent test environment because there is an unequivocal response question and the "truth" can always be determined (by subsequent revision) so the appropriateness of the edit can be evaluated.

Note that the authors clearly have a quintessential micro-editing requirement. Namely, their intended product is a longitudinal file of individual records of bank deposits.

After describing the underlying survey environment, the authors described their step-by-step process to design effective edit tolerances. They described how they developed the definition of edit cell and how they had to combine cells to get the proper break between cell size and homogeneity. They next described the model to predict (forecast) reported deposits, discussed alternative models and provided cogent reasons for each decision along the way. Following that, the authors describe their procedure for setting cell edit tolerances. After details of the edit system were decided upon, they were able to test various options based on the 1991-92 edit experience.

They took a major step in couching their analysis in terms of Type I and Type II errors to evaluate findings. The authors provided extensive tables and descriptions of their analysis. It is interesting to note that the current system has too many Type II errors, and future work will introduce refinements to achieve a lower rate.

Although one rarely comes across such a well-suited environment to test edit procedures and evaluate performance, this report is valuable in describing how to proceed under ideal circumstances. Using this ideal as a guide, one can modify procedures and change directions based on information actually available when attempting to apply the methods described in this report to other surveys.

7. Concluding Remarks

Both of these papers have a great deal to offer the reader. The first clearly illustrates how graphics can be applied to actual editing issues. One would hope that the examples here can suggest methods which can be applied to other surveys. The second paper is an excellent case-study for developing edit tolerances and evaluating them. This paper also can serve as a guide in helping others plan their own evaluation projects.

Session 5

TIME SERIES REVISION POLICIES

Time Series Revisions: The Effects on Gross Domestic Product

Robert P. Parker and Teresa L. Weadock
U.S. Department of Commerce
Bureau of Economic Analysis

Gross domestic product (GDP) is the most widely used measure of a Nation's overall economic activity. In response to the need for timely estimates, the Bureau of Economic Analysis (BEA) releases an "advance" estimate of quarterly GDP one month after the end of the each quarter. This estimate is based largely on monthly survey data for the first two months of the quarter and BEA judgmental projections for the missing source data. In each of the next two months, revised estimates of GDP that incorporate newly available and revised monthly and quarterly source data are released. Annual and benchmark (comprehensive) revisions of GDP are released on a regular schedule as annual and less frequently collected census-type data become available.

To measure the reliability and accuracy of the quarterly GDP estimates, BEA has conducted a series of studies based primarily on revisions to the successive quarterly estimates. Such studies have been found to be very useful to both users of the GDP estimates and to BEA. For users of the estimates, the studies provide insights into the likely size of future revisions to GDP and its major components and identify components whose reliability they would like to have BEA improve. For BEA, the studies help to identify components with problems in the source data or estimating methodologies. This information helps BEA to work with the agencies who prepare the source data or to devote its own resources to developing improved procedures to reduce revisions. Information from these studies also enables BEA to analyze the impact on the reliability of the GDP estimates of its revision schedule.

In addition, the studies satisfy the requirement of the Office of Management and Budget that agencies producing the principal Federal economic indicators provide periodic evaluations of their performance. This requirement, Statistical Policy Directive Number 3, states that these evaluations will "include an analysis of the accuracy of the series, the effects of revisions, and performance relative to established benchmarks."

This paper, which is based primarily on measures of reliability for the quarterly GDP estimates for 1978-91, consists of four parts: An overview of the preparation of the quarterly GDP estimates and a discussion of the measures used by BEA to measure their reliability;

NOTE: The views expressed in this paper are those of the authors and not necessarily those of the Department of Commerce or the Bureau of Economic Analysis.

highlights of a recently completed study by Allan Young at BEA on the reliability of the first three, or "current" quarterly GDP estimates for 1978-91;[1] an extension of this study to the effects on the quarterly estimates of subsequent annual revisions; and a discussion of revision practices that agencies use for the source data used to prepare the GDP estimates and how these practices affect the accuracy of the quarterly estimates.

The studies reviewed in this paper have implications both for BEA and for the agencies that provide the source data used to prepare GDP. They indicate that BEA should review the need for three current estimates and its policy of not revising prior quarters except at the time of annual and benchmark revisions. They also indicate that other agencies should review their revision practices to provide more timely and accurate revised data.

Part 1. Overview of the Preparation of Quarterly GDP Estimates and Measures of Reliability

Estimating schedule for quarterly GDP estimates

For each quarter, GDP estimates are prepared on a schedule that consists of three successive "current" estimates--"advance," "preliminary," and "final"-- and of subsequent estimates prepared as part of annual and comprehensive NIPA revisions.

The advance estimate is prepared about 1 month after the end of the quarter. For most components, the estimate is based on source data for either 2 or 3 months of the quarter. In most cases, however, the source data are not final and are subject to revision by the issuing agencies. Where source data are not available, the estimate is based primarily on past trends and on BEA analysts' judgment.

One month later, the advance estimate is replaced by the preliminary estimate, which is typically based on source data for all 3 months of the quarter. In most instances, the source data used for the preliminary estimates, particularly the data for the third month of the quarter, are subject to further revision.

One month later, the preliminary estimate is replaced by the final estimate, the last of the current estimates, which incorporates revisions in source data for the third month of the quarter and quarterly source data for some components. For virtually all components, these source data are subject to further revisions by the issuing agencies.

---

[1] See Allan H. Young, "Reliability and Accuracy of the Quarterly Estimates of GDP" in the October 1993 issue of the Survey of Current Business.

Each quarterly estimate is subject to three successive annual revisions, usually released in July. The first annual revision incorporates further revisions in the monthly or quarterly source data and introduces some annual source data. The second and third annual revisions incorporate a broad range of annual source data. For example, the "final" estimate for the fourth quarter of 1993, which was released last month, will be revised in July 1994 (first annual revision), July 1995 (second annual revision), and July 1996 (third annual revision). Each quarterly estimate is also subject to one or more comprehensive revisions, in which information from the economic and demographic censuses is incorporated in the monthly, quarterly, and annual source data by the issuing agency or by BEA.

Source data

More complete and more accurate information is generally available on an annual basis than on a quarterly or monthly basis. In many cases, annual data are based on larger samples or represent a complete universe count. In addition, annual data often correspond more closely to the desired definitions and therefore require less adjusting, or they may contain more information for making the necessary adjustments. As a result of these factors, quarterly estimates are obtained either by interpolating between annual estimates or by extrapolating from the most recent annual estimate.

Similarly, the annual estimates in many instances represent interpolations or extrapolations of the more complete and accurate information available in economic and demographic censuses, which are conducted every 5 years and 10 years, respectively.

The quarterly and monthly indicators that are used as interpolators and extrapolators are based largely on monthly or quarterly sample surveys conducted by various Federal statistical agencies. Exceptions include budgetary data from the Treasury Department, tabulations of export and import documents filed with the Customs Service, and tabulations of several types on payroll and income tax returns. Another type of exception occurs if no monthly or quarterly data are available--for example, data for some types of consumer purchases of services and of State and local government purchases of goods and services. In such cases, the quarterly estimates are obtained by interpolation and extrapolation based on a BEA analyst's judgment or related information.

An updated summary of the source data used for the NIPA's is included each year in the Survey of Current Business article that presents the annual NIPA revision (see pages 31-42 of the August 1993 Survey). For a list of methodological papers and for additional information about the NIPA's, see "A Look at How BEA Presents the NIPA's" in the February 1994 Survey, pages 31-33.

Sources of error

The GDP estimates contain several kinds of error. The most obvious kind arises in the current estimates either from preliminary or incomplete tabulations of monthly or quarterly source data or, where source data are not yet available, from BEA's judgments. Error also arises in both the current and the latest available estimates because source data do not meet NIPA requirements in terms of timing, valuation, coverage, and definitions. For example, business firms report some types of data on a fiscal year, rather than a calendar year, basis; even though adjustments by BEA reduce the effect of fiscal year reporting, the results differ from those that would have been obtained with calendar year reporting.

Error also arises from the sampling errors and biases in the monthly, quarterly and annual surveys and from biases and other errors in the annual and periodic universe counts. Probably the most troublesome of errors are those due to the delayed recognition of births and deaths of business firms in sample surveys. (These types of errors that affect source data are discussed in part 4 of this paper.)

Seasonal adjustment is another source of error.[2] Even if the unadjusted source data were free of error, seasonal adjustment would introduce errors. Although some reduction in seasonal adjustment error appears to have been achieved over time in the current estimates through the use of concurrent seasonal adjustment and by combining ARIMA methods with the ratio-to-moving-average method of seasonal adjustment, such errors are still of considerable magnitude.

Measures of reliability

The term "reliability" used in BEA studies refers to the revisions in the estimates, which reflect the following: (1) Replacement of preliminary source data with revised or more comprehensive data, (2) replacement of judgmental projections with source data, (3) changes in definitions or estimating procedures, and (4) in the constant-dollar estimates, updating of the base year.

In its studies of the reliability of the quarterly GDP estimates, BEA uses six summary measures to describe the revisions: Dispersion, bias, relative dispersion, relative bias, upward revisions, and directional misses. (This paper focuses on the dispersion because this measure effectively summarizes the information provided by the other measures.)

---

[2] Quarterly and monthly NIPA estimates are seasonally adjusted if necessary. Seasonal adjustment removes from the time series the average impact of variations that normally occur at about the same time and in about the same magnitude each year--for example, weather, holidays, and tax payment dates.

The six measures are calculated as follows. Let P represent the percentage change in the current estimates, L the percentage change in the latest available estimates, and n the number of quarterly changes.

Dispersion is the average of the absolute values of the revisions:

$$\Sigma|P-L|/n$$

Bias is the average of the revisions:

$$\Sigma(P-L)/n$$

Relative dispersion expresses the dispersion as a percentage of the average of the absolute values of the percentage change in the latest available estimates:

$$\frac{\Sigma|P-L|/n}{\Sigma|L|/n}$$

Relative bias expresses the bias as a percentage of the average of the percentage change in the latest available estimates:

$$\frac{\Sigma(P-L)/n}{\Sigma L/n}$$

Upward revisions expresses the number of times that the current estimate of the quarterly change was revised up by the latest available estimate, as a percentage of the number of quarterly changes.

Directional misses expresses the number of times that the sign of the current estimate of the quarterly change differed from that of the latest available estimate, as a percentage of the number of quarterly changes.

In evaluating these measures, they should be viewed in light of two aspects of the estimation process. First, a change in source data or estimating procedures, which one may assume affects the accuracy of the estimates, is not necessarily reflected in the revision of estimates of a given vintage. For example, an improvement in the current estimates results in a permanent decrease in revision size. An improvement in the latest available estimates results in a permanent increase in revision size. Improvement in both the current and latest available estimates results in little change. However, improvement that is introduced retrospectively into the latest available estimates, as is often the case, results in an increase in revision size for a period of years until the improvement is also reflected in the current estimates. Second, the latest available estimates reflect different vintages. The latest estimates up to 1982 at present reflect the incorporation of the benchmark input-output (I-

O) tables, which are based on detailed information from the economic censuses; the latest estimates beginning in 1983 do not yet reflect the incorporation of the recently released benchmark 1987 I-O tables.[3] Thus, the size of revisions beginning with 1983 estimates are most likely understated.

Part 2. Reliability of the Current Quarterly Estimates[4]

Summary of Young study, 1978-91

In the most recent BEA study, Young provided an overall evaluation of the reliability of the quarterly GDP estimates by comparing the successive current estimates of real GDP to the latest estimates and asking the following questions:

o Do the current estimates provide a correct indication of the direction of the change in aggregate economic activity?

o Do the current estimates provide a correct indication of whether the change in aggregate economic activity is larger (acceleration) or smaller (deceleration) than in the previous quarter?

Table 1 provides the summary answers to these questions for each of the current estimates. The record for 1978-91 shows that all three estimates correctly indicated direction of change almost 90 percent of the time. They correctly indicate acceleration and deceleration between 75 and 80 percent of the time. (If changes between -1 percent and +1 percent are disregarded, these early estimates correctly indicate direction over 90 percent of the time and acceleration about 85 percent of the time).

Young also found that for the same period, the incorporation of additional or more accurate source data in the second (preliminary) and third (final) quarterly estimates of GDP did not improve the reliability in comparison with the first (advance) estimate. He identified two factors that contributed to this finding. First, the data for second and third months of a quarter play only a small role in determining the change from the previous quarter. Second, the advance estimate is unaffected by certain sources of error in the preliminary and final estimates. In addition, the advance estimates

---

[3] The benchmark 1987 I-O tables were published in the April 1994 issue of the Survey of Current Business. The results of these tables will be incorporated into the GDP estimates in a comprehensive revision presently scheduled for release in late 1995.

[4] Most of the sections of this part of the paper, as well as several paragraphs of the previous section, were taken verbatim from Young's article in the October 1993 Survey of Current Business.

of GDP and its major components may benefit more from offsetting errors in the detailed components than the later current estimates; that is, the revisions of the advance estimates may be more negatively (or less positively) correlated than those of the preliminary and final estimates.

In the remainder of this part of the paper, Young's findings are presented in more detail.

Reliability of the quarterly estimates

Table 2 shows dispersion for quarterly changes in current- and constant-dollar GDP and its major components for 1978-82 and 1983-91. These measures show that the incorporation of additional or more accurate source data in the preliminary and final current estimates of GDP does not substantially improve the reliability in comparison with the advance estimates. Dispersion declines only slightly over the successive current-dollar estimates of GDP. For 1978-82, it declines from 1.93 percentage points in the advance estimates to 1.82 percentage points in the preliminary and final estimates. For 1983-91, it declines from 1.17 percentage points in the advance estimates to 1.14 percentage points in the preliminary estimates and 1.15 percentage points in the final estimates.

Dispersion actually increases slightly over the successive constant-dollar estimates of GDP. For 1978-82, it increases from 1.64 percentage points in the advance estimates to 1.72 percentage points in the preliminary estimate and to 1.75 percentage points in the final estimate. The corresponding figures for 1983-91 are 1.25, 1.27, and 1.33 percentage points.

A similar picture emerges for the major components of GDP. In many cases, the advance estimates provided a smaller measure of dispersion than did the preliminary or final estimates. In 1978-82, the advance estimates provided the smallest dispersion in 4 of the 11 current-dollar components--PCE nondurables and services, residential investment, and Federal Government purchases--which accounted for almost 60 percent of GDP. In 1983-91, the advance estimates provided the smallest dispersion in 3 components--PCE services, residential investment, and State and local government purchases--which accounted for over 40 percent of GDP. The record for the advance constant-dollar estimates is about the same as that for the current-dollar estimates, though the share of GDP for which the estimates perform the best is smaller for 1983-91. These results raise the question of whether one or both of the two later current estimates might be discontinued.[5]

---

[5] Previous studies have also shown that the advance estimates perform well in comparison with the later estimates, but as shown in table 2, the results are not as clear cut in some periods as in others.

Table 2 also permits one to compare the size of the dispersion measure for GDP with that of its major components. In general, dispersion in the components was larger than that in GDP. The components with the smallest dispersion--about the same as that for GDP--were total PCE and PCE services. The components with the largest dispersion--roughly 6 to 8 times as large as that for total GDP--were gross private domestic investment and Federal Government purchases. The unusually large dispersion in these components reflected a change in the treatment of the Commodity Credit Corporation (CCC) that was introduced in the 1991 comprehensive revision, whereby the CCC was shifted from government enterprises to general government. This shift affected the timing and valuation of transactions and resulted in large, essentially offsetting revisions in Federal Government purchases and the change in business inventories. Dispersion was also quite large in current-dollar nonresidential structures in 1978-82 and in constant-dollar imports in 1983-91, reflecting statistical improvements introduced in the 1991 comprehensive revision.

Trends since 1978

Table 3 examines revisions year by year to see if reliability of the GDP estimates appears to have changed in recent years. The table shows annual averages of dispersion and bias in the quarterly revisions between the successive current estimates and between the current estimates and the third annual revision estimates. For the revisions between the current estimates, the measures are shown for 1978-92; for the revisions between the current estimates and the third annual estimates, the measures are shown for 1978-89.

In Young's study, estimates from third annual revisions were used in place of the latest available estimates in order to provide a more nearly comparable standard for the entire period against which to compare the current estimates. Use of third annual estimates abstracts from much of the effect of the economic census and other information that is used in the comprehensive revisions to revise previously prepared third annual estimates. However, it does not remove the effects of definitional changes in the comprehensive revisions, because for most quarters a comprehensive revision intervenes between the current estimates and the third annual estimates. (To more fully study the effects of the annual revisions,

---

An early study concluded that the advance estimate might be sufficient; see Rosanne Cole, "Errors in Provisional Estimates of Gross National Product," National Bureau of Economic Research Studies in Business Cycles No. 21 (1969). See also Stephen K. McNees, "Estimating GNP, The Trade-off Between Timeliness and Accuracy," New England Economic Review (January/February 1986): 3-10; and Joseph W. Duncan and Andrew C. Gross, Statistics for the 21st Century (The Dun and Bradstreet Corporation, 1993).

a different approach is used in part 3 of this paper.)

The good performance of the advance estimates

The absence of much improvement in the successive current estimates has puzzled both users and estimators for some time. Two seldom recognized factors contribute to the observed result: (1) The small role played by the data for second and third months of a quarter in determining the change from the previous quarter, and (2) certain sources of error in the preliminary and final estimates to which the advance estimates are immune. In addition, advance estimates of GDP and its major components may benefit more from offsetting errors in the detailed components than the later current estimates; that is, the revisions of the advance estimates may be more negatively (or less positively) correlated than those of the preliminary and final estimates.

This section first discusses the two factors and then addresses the problem of quantifying the total error introduced by the second factor, which embodies seasonal adjustment errors and errors related to the estimation process for certain components. The section concludes with a discussion of the implications for the future.

The role played by the data for second and third months of the quarter in determining the change from the previous quarter is small. The change from the second to the third month receives a weight of only one-ninth in the determination of quarterly change. The weight of the second and third months together is only one-third. The weight of the first month is another one-third, and the second and third months of the previous quarter receive the remaining one-third.[6] Consequently, errors in neither the preliminary source data for the second and third months of a quarter nor in the judgmental projections used in lieu of source data affect the quarterly change as much as one might intuitively expect.

---

[6] This may be demonstrated as follows: Let $Q_1=X_1+X_2+X_3$ and $Q_2=X_4+X_5+X_6$, where $X_1, X_2, \ldots X_6$ are successive months of source data.
Then, if $d_4=X_4-X_3$, $d_5=X_5-X_4$, and $d_6=X_6-X_5$, the months in Q_2 may be stated as $X_4=X_3+d_4$, $X_5=X_3+d_4+d_5$, $X_6=X_3+d_4+d_5+d_6$, and

$$Q_2=3X_3+3d_4+2d_5+d_6.$$

Therefore, the quarterly change is

$$Q_2-Q_1=(3X_3+3d_4+2d_5+d_6)-(X_1+X_2+X_3)$$
$$=[(X_3-X_2)+(X_3-X_1)]+[3d_4+2d_5+d_6].$$

Introducing the notation for monthly differences, the first bracketed term becomes $[d_3+(d_2+d_3)]$, and

$$Q_2-Q_1=d_2+2d_3+3d_4+2d_5+d_6.$$

Normalizing the coefficients on the d's provides weights of 1/9, 2/9, 3/9, 2/9, and 1/9 for the five monthly changes that determine the quarterly change.

The seasonal adjustment of source data for the final current quarterly estimate introduces errors not present in the judgmental projections, which are developed on a seasonally adjusted basis. The seasonal adjustment factors for the current year are derived from the seasonal patterns of recent preceding years. (The concurrent seasonal adjustment method also includes the seasonal pattern of the current year.) The factors are revised as additional data become available, and they eventually reflect the average seasonal pattern of a period of years that extends symmetrically on either side of the given year. The difference between the initial estimate of the seasonal factor and the final estimate prepared some years later is an error that becomes part of the revision in the final current estimate. To the extent that they are based on judgmental projections, the advance and preliminary quarterly estimates do not contain this error.

Future work

The difference between the revisions to the advance estimate of a detailed component and those to the latest available estimate reflects three types of error: (1) The error in the preliminary monthly source data used for the advance estimates that is corrected in the revised monthly source data used for the final current estimate; (2) the error in the judgmental projections used in lieu of source data for the advance estimate; and (3) the error in the source data used for the final current estimate (including seasonal adjustment error) that replace the judgmental projections used for the advance estimate. The total revision in the advance estimate reflects the first two types of error; the total revision in the final current estimate reflects the third type. One should note that the second type of error, like the third, is determined with respect to the data as they stand several years later.

It would be desirable to determine the size of each of the three types of error at the detailed component level. It would also be desirable to determine the extent of correlation among the component revisions. Such analyses presently are not possible, because each vintage of each estimate and the associated source data are not available in a readily usable form. However, the database used by BEA to calculate the alternative measures of real GDP might be extended so as to retain not only the latest available estimates, but all the vintages of estimates at the detailed level at which deflation is carried out.[7] Over time, this database would be useful in exploring the outcomes of the estimation process and in developing improved procedures. For example, it will facilitate the development of econometric projection techniques and their comparison with judgmental

[7] For information about the alternative measures, see Allan H. Young, "Alternative Measures of Change in Real Output and Prices, Quarterly Estimates for 1959-91," in the March 1993 Survey of Current Business.

projections. In this respect, it should be noted that a recently completed study found that judgmental projections compared favorably with econometric techniques for certain import and export components.[8] Thus, such procedures would be difficult to justify if they do not lead to smaller revisions than do judgmental projections.

The question of whether the reliability of the seasonal adjustments on which the current quarterly estimates are based can be improved merits attention. An analysis at a fairly high level of aggregation suggests that revisions in seasonal factors may be large enough to contribute significantly to the observed results. Specifically, in some of the series examined, seasonal-factor revisions are as large as the variation in the irregular component.[9] Given that the error introduced by a judgmental projection is likely to be smaller than the irregular variation, this result suggests that for some detailed components, seasonal-factor revisions may play a significant role in causing the revision in the final estimate to be as large as that in the advance estimate. In addition, the seasonal adjustments used for source data should be designed from the standpoint of accurately measuring quarterly change. Little attention has been paid to whether the currently used seasonal adjustment procedures are suitable from this standpoint. In addition, BEA should consider whether more use of concurrent seasonal adjustment, with or without ARIMA, would improve reliability.

Finally, because the third month of a quarter receives little weight in the estimate of change for that quarter, there may be instances in which efforts to reduce revisions in the quarterly GDP estimates should focus on improving the final monthly source data rather than the preliminary monthly source data. As shown earlier, for a survey with three successive monthly estimates, two-thirds of the advance quarterly change is based on three monthly final estimates, while only one-ninth is based on the initial monthly estimate for the last month of the quarter.

Part 3. Effects of Annual Revisions on Quarterly GDP Estimates

In his study on the reliability of the quarterly GDP estimates, Young studied the reliability of the three current quarterly GDP estimates. In this part of the paper, the revisions in the quarterly estimates published in the regular annual revisions are studied.

---

[8] See Albert A. Hirsch and Michael A. Mann, An Analysis of the Use of Time-Series Models to Improve Estimates of International Transactions, Bureau of Economic Analysis Working Paper 7 (Washington, DC: Bureau of Economic Analysis, April 1993).

[9] The irregular component is the residual after the systematic components--the seasonal and trend-cycle--are determined by the seasonal adjustment method.

Dispersion

Table 4 shows the dispersion for quarterly changes in current- and constant-dollar GDP and its major components for 1981-90, the years for which comparisons using annual revisions could be made.[10] These measures show that the incorporation of additional or more accurate source data in the first and third annual revisions substantially improves the reliability in comparisons with the third, or "final" current estimates. The dispersion for current-dollar GDP from the final to the first annual revisions declines from 1.34 percentage points to 0.84 percentage point; for constant-dollar GDP, the decline is from 1.45 percentage points to 0.83 percentage point. From the second to the third annual revisions, the dispersion for current-dollar GDP declines from 0.85 percentage point to 0.65 percentage point; for constant-dollar GDP the decline is from 1.08 percentage points to 0.85 percentage point. However, from the first to the second annual revisions for the current-dollar estimates, the dispersion is virtually unchanged, and for the constant-dollar estimates, there is a significant increase in the dispersion from 0.83 percentage point to 1.08 percentage points.

For the major components of GDP, except for nonresidential structures and producers' durable equipment, the first annual revision estimates recorded a smaller measure of dispersion than did the final estimates. Between the first and second annual revisions, the reliability of all major components improved. Between the second and third annual revisions, the reliability of all major components except for durable goods personal consumption expenditures (PCE) and exports of goods and services improved.

Table 4 also permits one to compare the size of the dispersion measure for GDP with that for its major components. The dispersion for the components was larger than that for GDP. The components with the smallest dispersion were PCE nondurable goods, PCE services, and State and local government purchases. The components with the largest dispersion, more than 6 times as large as that for total GDP, were gross private domestic investment, nonresidential structures, and services, and Federal Government purchases. Of these components, the largest dispersion was Federal Government purchases. As noted in the review of Young's study, the size for this component reflected a change in the treatment of the Commodity Credit Corporation that was introduced in the 1991 comprehensive revision.

---

[10] To conduct this study, it was necessary to reduce the period covered in Young's study because there were no annual revisions in 1980 and 1991, years of benchmark revisions. See footnote 1 of table 4 for additional details.

Findings

Young found that the revisions in the advance current estimates were about the same size as those in the other current estimates. Thus, he questioned the need to continue the preparation of the preliminary and/or final quarterly GDP estimates. The comparisons in table 4 show, as expected, that generally there was continued improvement in the reliability of the estimates in each successive annual revision; therefore, the need for the annual revisions is not called into question. However, the relative size of the improved reliability from the final quarterly estimates to the quarterly estimates from the first annual revisions was larger than expected. The large improvement, which is evident in both the current- and constant-dollar GDP estimates, was unexpected because very few of the annual surveys that are used to prepare the revised GDP estimates become available in time to be incorporated into the first annual revisions. For example, the Census Bureau's annual surveys of retail trade and of manufacturers, which are used for the estimates of PCE goods and of producers' durable equipment, are not available until the second annual revision. If the improved reliability does not result primarily from the incorporation of such new source data, then the improvement might result from two other sources. One source would be the replacement at the time of the annual revisions of seasonal adjustment factors that were derived from the seasonal pattern of preceding years with factors that also reflect the pattern for the most recent year. The second source would be the incorporation of corrections to monthly or quarterly source data series that become available too late to be incorporated into the "final" current estimates.

If research of the detailed components shows that these are major sources of the large revisions in the first annual revisions, then BEA should consider revising previous quarters between annual revisions. For example, when the advance estimate for the fourth quarter of a given year is released in January, newly available corrections and revisions to source data series and updated seasonal factors could be incorporated into revised first-, second-, and third-quarter estimates. This change would allow for a more effective use of concurrent seasonal adjustment -- that is, using these updated factors for all months or quarters of the year. (For many of the source data series for which the issuing agency uses concurrent seasonal adjustment, the new factors are available but are only used for the most recent two or three months.) Changing BEA's revision practice as just described would provide more reliable quarterly estimates earlier than under the present annual revision schedule. If this change were merged with Young's suggestion to eliminate one or more of the current quarterly estimates, BEA might, for example, drop the final current estimates and revise the previous quarters of the year when the advance estimates are released.

Part 4. Revision Practices in GDP Source Data

Accuracy of the GDP estimates

The term "accuracy" refers to the total measurement error. On the assumption that each successive estimate is more accurate than previous ones, revisions can be viewed as measuring part of the total error in earlier estimates. The rest of the error in these estimates, which is unknown, becomes the total error in the latest estimates. The error in the latest estimates results primarily from the following sources: (1) Errors in the most recently available underlying monthly or quarterly, annual, or periodic census source data; (2) errors in the adjustments made by BEA to convert source data to the definitions and conventions used in compiling GDP; (3) errors in BEA's judgmental estimates for components for which there are no source data; and (4) errors because the most recently available source data have not yet been incorporated into the estimates. It is important to note that these types of measurement error have two different effects on the quarterly GDP estimates. The first type of error affects the levels of the estimates of GDP and its components; the second type affects the changes in the estimates. In this paper, the focus is on the second type of errors.

The implication of the presence of measurement error in the latest estimates limits the extent to which the BEA's measures of reliability can be used as measures of accuracy. The questionable relationship between reliability and accuracy is illustrated by the component that has a very high measure of reliability -- that is, very small revisions -- because the source data are never revised by the issuing agency even though the series maybe based on a very small sample. The next section of the paper discusses how three commonly used revision practices adversely affect the key monthly and quarterly source data series used to estimate GDP. For this purpose, revision practices are defined as those that affect only "not seasonally adjusted estimates."

Revision practices that adversely affect GDP

For some surveys, the revision practice consists of a series of regular annual revisions and, if necessary, periodic benchmark revisions. For annual revisions, the monthly or quarterly series are adjusted to reflect annual survey data, which are based on larger samples, or to reflect annual census (universe) data. For periodic benchmark revisions, first the annual survey data and then the monthly or quarterly series are adjusted for all periods since the last benchmark. Examples of GDP source data prepared using this type of revision practice, the issuing agency, and the GDP components affected are as follows: Retail and wholesale trade sales and inventories from the Bureau of the Census (BOC) -- personal consumption expenditures (PCE) and change in business inventories (CBI); farm output and income from the Department of Agriculture--CBI and farm income; manufacturers' shipments and inventories from BOC--producers' durable

equipment and CBI; and establishment employment from the Bureau of Labor Statistics (BLS)--wages and salaries and PCE.

Although the revised monthly and quarterly estimates for source data series prepared using this practice are considered very accurate, they are nevertheless subject to measurement errors that are likely to introduce errors into the latest estimates of quarterly GDP changes. These errors arise because the agencies generally proportionately allocate across months (or quarters) differences (1) between the monthly or quarterly surveys and the annual surveys and (2) between the annual surveys and the census results where proportionate allocation is not appropriate. For example, the most common sources of differences are errors in the initial survey estimates, both monthly and quarterly and annually, due to delayed recognition of births and deaths, to classification errors, or to misreported data. For births and deaths, proportionate allocation of errors introduces new errors into the revised monthly or quarterly series because this type of allocation fails to recognize that changes in the business population are significantly affected by the business cycle and are not likely to have occurred proportionately over the period covered by the revision. For classification and misreporting errors, the errors also are not likely to have occurred proportionately throughout the revision period.

Another common revision practice used for monthly and quarterly series, which can be viewed as a variant of the practice described above, also is likely to introduce errors into the latest estimates of quarterly GDP changes. Under this practice, when periodic benchmark data are introduced, they are used to adjust only the monthly or quarterly and annual estimates for the year for which such data are available and to serve as the basis for the samples to be used for subsequent periods. Data for previous periods are not revised even though the data for these periods may include errors. Examples of GDP source data that are prepared using this practice, the issuing agency, and the GDP components affected are as follows: Foreign direct investment income in the United States from BEA -- corporate profits; State and local government value of construction put-in-place from BOC -- government purchases; and the Consumer Price Index from BLS -- PCE.

A third common revision practice occurs when the agency revises a series to remove selected measurement errors but not necessarily to remove the largest errors, which therefore remain in the revised estimates. The following monthly GDP source data are examples of this type of series: monthly merchandise trade from BOC -- net exports; new residential construction put in place from BOC -- residential fixed investment; and average hourly earnings from BLS -- wages and salaries.

The monthly merchandise trade data are regularly revised to record the export and import transactions based on Customs documents in the correct month but not to record reliable estimates of transactions for which Customs documents are not required to be filed

because of the value of the transaction is below some cutoff or "low value." These unreported low value transactions, which currently account for about 2.5 percent of exports and 4 percent of imports, are estimated using factors based on the amount of such transactions when the exemptions were granted. The adjustments are extrapolated by country, but not by commodity, using changes in reported transactions. This procedure is likely to introduce errors into the monthly changes in exports and imports.

The new residential construction put-in-place series, which is based largely on a sample of housing starts, is regularly revised to reflect additional monthly reports but not to reflect more accurate annual survey data. For the new single-family residential construction component of this series, the value of housing starts are "phased" in over many months based on a fixed monthly pattern of construction activity. This practice introduces errors into the monthly changes in the put-in-place series because the pattern is not updated on a regular basis (the present pattern was estimated for 1976) and there are no data to adjust the pattern for developments such as unusually bad weather. The errors created by this practice can be illustrated using the recent California earthquake. This disaster not only delayed starts, which the series does reflect, but also delayed activity on previously started houses, which the series does not reflect. For new multi-family residential construction, BOC conducts a monthly survey of construction put in place based on a sample of housing starts; this series is not benchmarked.

The final example of a "selective" revision practice is the average hourly earnings series, which is estimated from the BLS monthly establishment survey. Although the employment data collected in that survey are benchmarked annually, the earnings data are corrected only to reflect the revised employment data, which are used to weight the detailed industry earnings estimates to arrive at "all industry" totals. It is likely that a benchmarked hourly earnings series not only would change the levels of the series, but also the monthly changes.

Improving revision practices

This discussion of revision practices identifies some of the types of measurement errors that remain in BEA's latest estimates of changes in quarterly GDP. Although reducing these errors for some series would require the collection of new data, it should be possible for the issuing agencies to reduce certain types of errors with little or no additional resources. For example, errors caused by proportionate allocation of the impact of delayed incorporation of births and deaths could be significantly reduced. Instead of waiting many years until a regular benchmark revision, agencies could continuously track births and deaths and adjust their series annually, even with a one-year lag. This procedure would improve the accuracy of the annual estimates earlier and largely eliminate the proportionate allocation of the errors at the time the agency prepares

their benchmark revision of the survey. (Discussions of such changes are currently underway between BEA and BOC for their annual surveys.)

For series where benchmark revisions are not carried backwards in time, agencies should publish the amount of the sample "drift" since the last benchmark and adjust the historical series. If it is not possible for the agency to make these adjustment, then BEA and other users could make their own adjustments. The latter solution is less desirable because the agencies usually have information with which to make the adjustments that is not available to users, and when different users make the adjustments, they are likely to develop different adjustments.

Table 1.--Reliability of Current Estimates of Quarterly Changes in Real GDP, 1978-91
[Percentages providing correct indication]

	All quarters			Omitting quarters with changes/differences of 1 percent or less	
	Direction of change (56)	Larger/smaller change than in previous quarter (55)	Change between +1 percent and +4 percent (56)	Direction of change (48)	Larger/smaller change than in previous quarter (43)
Advance	88	78	75	92	86
Preliminary	89	75	70	94	81
Final	89	76	66	94	84

NOTE.--The number of comparisons is shown in parentheses.

Table 2.--Dispersion in Revisions in the Quarterly Changes in GDP and Its Components
[Percentage points/1/]

	Current dollars		Constant dollars	
	1978-82	1983-91	1978-82	1983-91
Gross domestic product:				
Advance	1.93	1.17	1.64	1.25
Preliminary	1.82	1.14	1.72	1.27
Final	1.82	1.15	1.75	1.33
Personal consumption expenditures:				
Advance	1.71	1.40	1.88	1.37
Preliminary	1.70	1.41	2.00	1.27
Final	1.80	1.35	2.12	1.30
Durable goods:				
Advance	5.72	4.20	5.00	3.96
Preliminary	5.15	3.88	5.11	3.63
Final	5.42	3.97	5.05	3.98
Nondurable goods:				
Advance	2.31	1.74	1.75	2.26
Preliminary	2.51	1.45	2.37	2.10
Final	2.50	1.37	2.39	2.03
Services:				
Advance	1.78	1.37	1.38	1.39
Preliminary	1.90	1.51	1.50	1.36
Final	1.96	1.59	1.56	1.42
Gross private domestic investment:				
Advance	13.20	9.38	10.64	9.53
Preliminary	12.67	8.62	10.24	9.30
Final	12.11	8.68	10.75	9.32
Fixed investment:				
Advance	7.01	3.03	5.59	3.74
Preliminary	4.96	2.43	4.08	3.29
Final	4.45	2.77	3.82	3.64
Nonresidential:				
Advance	8.24	3.67	6.36	4.42
Preliminary	5.63	3.19	4.15	4.07
Final	5.11	3.20	3.62	4.56
Structures:				
Advance	13.01	6.39	9.01	5.33
Preliminary	9.90	4.54	6.18	4.13
Final	9.47	4.92	6.10	4.66

Table 2.--Dispersion in Revisions in the Quarterly Changes in GDP and Its Components--Continued
[Percentage points/1/]

	Current dollars		Constant dollars	
	1978-82	1983-91	1978-82	1983-91
Producers' durable equipment:				
Advance.....	7.09	4.02	6.65	5.21
Preliminary.....	5.17	3.87	4.85	5.39
Final.....	4.20	3.99	4.42	5.77
Residential:				
Advance.....	7.17	4.84	6.91	5.27
Preliminary.....	8.56	4.91	8.67	5.11
Final.....	7.63	4.98	7.89	5.22
Change in business inventories...	....	....	....	....
Net exports of goods and services:				
Exports:				
Advance.....	8.90	5.49	7.52	5.33
Preliminary.....	8.80	4.72	7.87	4.85
Final.....	8.02	5.19	7.07	5.67
Imports:				
Advance.....	5.48	8.12	7.21	8.92
Preliminary.....	4.98	7.24	5.64	9.29
Final.....	4.71	7.55	5.71	9.61
Government purchases:				
Advance.....	4.25	3.93	3.46	4.83
Preliminary.....	4.37	4.05	3.62	4.79
Final.....	4.34	4.05	3.75	4.89
Federal:				
Advance.....	11.40	9.09	10.36	10.70
Preliminary.....	12.29	9.11	10.48	10.49
Final.....	12.81	8.92	10.99	10.58
State and local:				
Advance.....	2.51	1.53	2.15	1.41
Preliminary.....	2.61	1.63	2.17	1.62
Final.....	2.34	1.65	2.20	1.60

1. Calculated from quarterly percentage changes at seasonally adjusted annual rates.

Table 3.--Annual Averages of Dispersion and Bias in Revisions in the Quarterly Changes in GDP
[Percentage points/1/]

Year	Dispersion						Bias					
	Advance to Preliminary	Preliminary to Final	Advance to Final	Advance	Preliminary	Final	Advance to Preliminary	Preliminary to Final	Advance to Final	Advance	Preliminary	Final
				To third annual revision estimate						To third annual revision estimate		
	Current-dollar estimates											
1978...	0.5	0.6	1.0	2.1	1.9	1.9	-0.5	-0.2	-0.7	-2.0	-1.6	-1.4
1979...	.5	.3	.3	1.2	1.1	1.0	-.4	.1	-.3	0	.4	.3
1980...	1.0	.5	.8	.7	1.1	1.3	.3	-.1	.2	-.3	-.6	-.5
1981...	2.0	.4	2.3	3.4	2.0	1.9	-2.0	-.2	-2.3	-3.1	-1.0	-.8
1982...	.7	.7	.3	1.8	1.3	1.7	.7	-.5	.1	.7	0	.5
1983...	.5	.2	.7	2.4	2.0	1.8	.3	0	.3	.5	.2	.2
1984...	.7	.4	1.0	1.6	1.8	1.7	-.4	-.3	-.6	.1	.5	.7
1985...	.4	.4	.8	.8	1.2	1.6	.4	.4	.8	-.8	-1.2	-1.6
1986...	.5	.2	.4	1.3	1.5	1.6	0	.2	.2	-.4	-.3	-.5
1987...	.6	.4	.8	1.2	.6	.8	-.6	0	-.6	-1.2	-.6	-.6
1988...	.7	.2	.9	1.4	.7	.5	-.7	-.2	-.9	-1.4	-.7	-.5
1989...	.6	.4	.8	1.3	1.3	1.3	0	.3	.3	.6	.7	.4
1990...	.2	.4	.6	NA	NA	NA	.2	.4	.6	NA	NA	NA
1991...	.4	.3	.4	NA	NA	NA	0	.2	.2	NA	NA	NA
1992...	.9	.2	.9	NA	NA	NA	-.9	-.1	-.9	NA	NA	NA

Table 3.--Annual Averages of Dispersion and Bias in Revisions in the Quarterly Changes in GDP--Continued
[Percentage points/1/]

Year	Dispersion						Bias					
	Advance to Preliminary	Preliminary to Final	Advance to Final	Advance	Preliminary	Final	Advance to Preliminary	Preliminary to Final	Advance to Final	Advance	Preliminary	Final
				To third annual revision estimate						To third annual revision estimate		
	Constant-dollar estimates											
1978...	.2	.5	.7	2.4	2.3	2.0	-.2	-.2	-.4	-1.4	-1.2	-1.0
1979...	.8	.2	.6	1.0	.8	.8	-.6	.1	-.5	-.5	.1	0
1980...	.5	.7	.7	.5	.5	.8	.3	-.3	0	-.2	-.5	-.2
1981...	1.0	.3	1.2	2.0	1.3	1.1	-1.0	-.2	-1.2	-1.8	-.8	-.6
1982...	.6	.6	.2	2.4	1.8	2.4	.3	-.5	-.2	1.0	.8	1.2
1983...	.4	.4	.7	1.5	1.4	1.1	0	-.1	-.1	-.3	-.3	-.2
1984...	.7	.4	1.0	1.5	1.8	1.8	-.3	-.1	-.4	.2	.5	.6
1985...	.7	.5	.7	1.5	1.7	2.1	.2	.5	.6	-1.5	-1.7	-2.1
1986...	.4	.4	.2	2.1	2.1	2.3	0	.1	.1	0	0	-.1
1987...	.3	.3	.3	1.2	1.1	1.3	-.1	-.1	-.2	-1.2	-1.0	-.9
1988...	.5	.1	.6	.9	1.2	1.3	-.5	0	-.5	-.9	-.4	-.4
1989...	.6	.2	.7	1.7	1.6	1.6	0	0	0	1.1	1.1	1.0
1990...	.3	.3	.6	NA	NA	NA	.3	.3	.6	NA	NA	NA
1991...	.5	.3	.5	NA	NA	NA	.1	.3	.4	NA	NA	NA
1992...	.7	.3	.6	NA	NA	NA	-.7	.1	-.6	NA	NA	NA

NA-Not Available
1. Calculated from quarterly percentage changes at seasonally adjusted annual rates.

Table 4.--Dispersion in Revisions in the Quarterly Changes in GDP and Its Components, 1981-90/1/
[Percentage points/2/]

	Current dollars	Constant dollars
Gross domestic product:		
Advance.....	1.38	1.32
Preliminary.....	1.31	1.36
Final.....	1.34	1.45
First Annual.....	.84	.83
Second Annual.....	.85	1.08
Third Annual.....	.65	.85
Personal consumption expenditures:		
Advance.....	1.48	1.47
Preliminary.....	1.50	1.39
Final.....	1.51	1.49
First Annual.....	1.32	1.19
Second Annual.....	.82	.71
Third Annual.....	.50	.47
Durable goods:		
Advance.....	4.49	4.22
Preliminary.....	4.15	4.16
Final.....	4.46	4.55
First Annual.....	3.27	3.41
Second Annual.....	1.79	1.91
Third Annual.....	1.84	1.94
Nondurable goods:		
Advance.....	1.85	2.15
Preliminary.....	1.61	2.09
Final.....	1.56	2.04
First Annual.....	1.36	1.38
Second Annual.....	1.10	1.03
Third Annual.....	.63	.63
Services:		
Advance.....	1.46	1.34
Preliminary.....	1.62	1.35
Final.....	1.74	1.45
First Annual.....	1.02	1.23
Second Annual.....	1.11	.97
Third Annual.....	.70	.58

Table 4.--Dispersion in Revisions in the Quarterly Changes in GDP and Its Components, 1981-90/1/--Continued
[Percentage points/2/]

	Current dollars	Constant dollars
Gross private domestic investment:		
Advance	11.38	10.48
Preliminary	10.37	9.93
Final	10.48	10.13
First Annual	7.10	7.21
Second Annual	5.28	5.12
Third Annual	4.88	5.22
Fixed investment:		
Advance	3.97	4.22
Preliminary	3.28	3.83
Final	3.38	3.99
First Annual	3.09	3.27
Second Annual	2.59	2.93
Third Annual	2.09	1.87
Nonresidential:		
Advance	4.61	4.84
Preliminary	3.99	4.56
Final	3.73	4.71
First Annual	4.21	4.50
Second Annual	3.99	4.65
Third Annual	2.37	2.06
Structures:		
Advance	7.98	6.19
Preliminary	6.17	4.73
Final	6.48	5.16
First Annual	6.66	5.78
Second Annual	6.49	5.87
Third Annual	5.07	4.31

Table 4.--Dispersion in Revisions in the Quarterly Changes in GDP and Its Components, 1981-90/1/--Continued
[Percentage points/2/]

	Current dollars	Constant dollars
Producers' durable equipment:		
Advance.....	4.59	5.56
Preliminary.....	4.10	5.61
Final.....	3.85	5.66
First Annual.....	4.59	6.30
Second Annual.....	3.79	5.31
Third Annual.....	1.40	2.84
Residential:		
Advance.....	6.30	6.61
Preliminary.....	6.48	6.58
Final.....	6.35	6.46
First Annual.....	4.69	5.72
Second Annual.....	3.81	4.22
Third Annual.....	3.54	3.62
Change in business inventories.....	...	....
Net exports of goods and services:		
Exports:		
Advance.....	6.13	5.80
Preliminary.....	5.57	5.29
Final.....	5.63	5.61
First Annual.....	4.41	4.25
Second Annual.....	3.09	3.52
Third Annual.....	3.73	4.20
Imports:		
Advance.....	7.99	8.74
Preliminary.....	7.85	9.38
Final.....	7.95	9.46
First Annual.....	4.76	5.50
Second Annual.....	3.44	3.94
Third Annual.....	3.25	4.80

Table 4.--Dispersion in Revisions in the Quarterly Changes in GDP and Its Components, 1981-90/1/--Continued
[Percentage points/2/]

	Current dollars	Constant dollars
Government purchases:		
Advance	4.22	4.74
Preliminary	4.46	4.87
Final	4.51	5.01
First Annual	4.19	4.49
Second Annual	3.52	3.75
Third Annual	2.91	3.21
Federal:		
Advance	10.01	11.30
Preliminary	10.54	11.48
Final	10.61	11.78
First Annual	10.45	10.75
Second Annual	8.75	9.55
Third Annual	7.14	7.55
State and local:		
Advance	1.66	1.56
Preliminary	1.81	1.76
Final	1.83	1.71
First Annual	1.49	1.43
Second Annual	1.18	1.17
Third Annual	.85	.71

1. As previously indicated in the text, for the comparisons shown in this table, it was necessary to use the period 1981-90. In addition, because the annual revisions in 1985 and 1991 were replaced by comprehensive revisions, the comparisons exclude years with no annual revision. The first annual revision comparisons exclude 1984 and 1990, the second annual comparisons exclude 1983 and 1989, and the third annual comparisons exclude 1982 and 1988.

2. Calculated from quarterly percentage changes at seasonally adjusted annual rates.

RAISING THE NATION'S UNEMPLOYMENT RATE

John E. Bregger
U.S. Bureau of Labor Statistics

I. Introduction

Data released on February 4, 1994, reflected major revisions in the questionnaire and collection methodology that were introduced into the Current Population Survey (CPS), following a planning and developmental process extending over the previous 8-year period. Looking back over these 8 years, was the process worth the effort and cost? Did the data improvements exceed the losses engendered by the breaks in time series? Were public understanding and appreciation of the data negatively affected by the changes? These and many other questions will hopefully be answered in this paper.

II. What happened?

In a nutshell, the nation's overall unemployment rate was found to be somewhat higher in early 1994 than it was in late 1993, resulting directly from a wholesale, stem to stern, series of changes to the survey questionnaire and the total conversion to computer-assisted interviewing. In addition to the rate of unemployment, a number of other important data series were affected by the CPS revisions, including the estimation of discouragement and of persons working part time involuntarily.

Based on tests of the new system, the effect on the overall annual average unemployment rate for 1993 was estimated to be about half a percentage point. That is, utilizing data gathered from a totally separate, parallel, survey, the newly redesigned questions, asked by interviewers using laptop computers or calling from a centralized interviewing facility, identified more people unemployed than under the then current procedures. And if that wasn't dramatic enough, changes in definition of discouragement caused the figure to tumble some 60 percent, and measurement refinements lowered economic part-time employment by 20-25 percent. In other words, of the three most important measures of labor market slack, the one most people point to and talk about -- the unemployment rate -- has now been raised (on a statistically significant basis), while the other two have been lowered markedly. Early results (January-April 1994) from the implementation of the new survey and procedures confirmed the direction of the expected changes.

III. History

How (or why) did this come about? The answer to this question comes in several parts, involving secular changes in the economy, a Presidential commission, the advent of new technologies and surveying knowledges, and, of course, careful planning. These will all be briefly described in a whirlwind tour of the historical backdrop for changing the Current Population Survey.

First, the secular changes. The last time that the questionnaire had been changed to any degree was in 1967, resulting from a period of research in the aftermath of the President's Committee on Employment and Unemployment Statistics (the Gordon Committee). In the subsequent years, many societal changes have taken place, including the more prominent role of women, especially mothers, in the labor force; the continuing shift from a goods- to a service-producing economy; changes in the way business operates, such as opening and closing hours; and, somewhat related to the other factors just cited, shifts in the nature of employment, including more part-time work and less permanent attachment of employees to their employers.

The next Presidential commission to study the statistics, the National Commission on Employment and Unemployment Statistics (the Levitan Commission), issued its report, including a number of recommendations in the labor force area, on Labor Day 1979. For our purposes, its most significant recommendation was for major conceptual changes in the way we measure labor market discouragement. And, while this particular recommendation was accepted for implementation by Secretary of Labor Donovan two years later, it was not implemented, owing to a lack of available funding in the early 1980s for a parallel test panel of households for testing potential questionnaire changes.

At the same time that these developments were going on, there have been many innovations in the way data are collected, innovations that could be expected to improve the quality of data. Foremost among these have been the recognition of the relevance of the theories and methods of cognitive psychology in designing survey instruments and the use of the computer in the interviewing process. With respect to cognitive psychology, under the auspices of the Committee on National Statistics, National Research Council, psychologists, other behavioral scientists, and survey methodologists had come together in 1984 to discuss the contributions that each discipline could make to survey design and, in so doing, helped to launch the cognitive aspects of survey methodology movement. One of the legacies of that advanced seminar is a four-component cognitive model of the question-response process -- comprehension,

retrieval, judgment, and response -- that has provided a very useful framework for designing and evaluating survey questions.

Use of the computer for data collection has been around for some time, but perhaps not for such a large undertaking as a monthly sample survey of 60,000 households, and certainly not for use in personal decentralized interviewing. Testing had suggested that, not only was a large-scale application doable, but, more importantly, it offered incredible gains in a variety of ways. Among these were fewer constraints on the number or variations of questions that could be included in the instrument; greater accuracy of data collection, in that interviewers were more likely to ask questions as worded (some had been anticipating what question they would be asking next); and accuracy of data transcription and transmission. When coupled with the desire to change and add questions to improve overall accuracy of identifying labor force status, the potential for improvement was ever so much greater, because the computer could permit intricate skipping and the storage of earlier information for later use that no interviewer could carry out in a pencil and paper environment.

The planning process for carrying out all of this commenced with a series of conferences involving the senior staffs of the Bureaus of Labor Statistics and the Census. The two agencies held a series of meetings beginning in April 1986 and two years later had a detailed plan to redesign the CPS, essentially in its entirety, with the questionnaire-related changes being the centerpiece. Budget submissions, with extensive, year by year, spending plans, were sent forward to OMB in time for the 1990 budget cycle. And, with favorable indications of approval forthcoming, work actually began in late 1988 toward a comprehensive survey redesign, with 1994 being the principal target date.

With respect to revisions to the questionnaire, a number of BLS-Census work groups were set up to develop a new questionnaire. The questionnaire was to be designed under the following guidelines: 1) It would not be constrained by the one-page limitation then in place; 2) it would take advantage of all aspects of automated data collection; 3) it would build upon recommendations of the 1978-79 Presidential Commission (and, to a lesser degree, the 1961-62 Commission); and 4) it would utilize to the maximum extent possible the knowledges available from cognitive science. Behavioral science laboratories were established in both agencies that brought in volunteers from the outside to react to various questions and question sequences. Questions explored included: What are the meaning of terms such as "work," "last week," "layoff," and "private company?" What method of collecting information on

the "actual number of hours worked last week" produces the most accurate data? How can response options be revised to simplify reporting and improve the categorization process -- and, in doing so, reduce measurement error? How could sensitive questions, such as on earnings, be revised to minimize nonresponse and improve reporting accuracy? How might the process of verifying information from a prior month's interview, rather than asking for the same information every month affect the quality and accuracy of the data? By the time we were through, we had managed to come up with satisfactory answers, for ourselves at least, to most of these as well as many other questions.

As a new questionnaire began to take shape, field testing became the next order of business. By this time (1990), two alternative versions of a potential new set of questions were in hand, along with the CPS questions then in use. The goal was to determine the best overall question wording from the three. To do this, the two Bureaus conducted a computer-assisted telephone interview (CATI), random digit dialing (RDD) test at the Census Bureau's centralized interviewing center in Hagerstown, Maryland. The first phase of this CATI/RDD test extended from July 1990 to January 1991, involving approximately 72,000 persons. Its purpose was to compare the then current version of the CPS questionnaire with the two test versions.

The principal product of the first phase was the selection of a single alternative questionnaire, close to the official version now in place, with appropriate additions and improvements that were deemed necessary due to the results of the testing. A second test phase was conducted between July and October of 1991 with approximately 30,000 persons, again via CATI/RDD; with very limited changes, this became the final version to be used in an 18-month parallel survey. During both the phase-one and phase-two testing, as well as the parallel survey phase, researchers employed a variety of methodologies to evaluate alternative question formats. These included respondent debriefings (via follow-up probe questions and vignettes), interviewer debriefings (via focus groups and debriefing questionnaires), response-distribution analysis, item nonresponse analysis, and behavior coding.

IV. The parallel survey

As researchers have long understood and as was once again verified in the CATI/RDD testing, if one or more important questions are changed (even slightly) in a continuing survey setting, we can expect different results. In other words, changing several questions in the CPS could be expected to have an effect on major measures, such as the rate of unemployment. Since the total number of potential questions increased from about 45 to 128 (no one, of course,

ever is asked more than a few of these!) and the wording of almost every question was changed, there was a virtual guarantee that we should expect differences on most of the statistical measures emanating from the survey. Thus, it was more than prudent to plan for a parallel, or overlap, survey for an adequate period of time in order to get some handle on the pact of these changes.

Ideally, we would have liked to have had a parallel sample extending for at least 2-1/2 years, with the same number of households as the ongoing CPS. This would have guaranteed a fully seated set of sample data for a full year, in terms of the 4-8-4 rotation group pattern. But, because our funding was limited, we had to settle for a 12,000 household sample covering the 18-month period, July 1992 through December 1993. Termed the CATI-CAPI Overlap Survey (CCO) internally and the Parallel Survey (PS) externally, this survey introduced the laptop computer (the CAPI portion of the CCO) into large-scale data collection.

One of our initial concerns was how well interviewers would adapt to using laptop computers and whether respondents would react favorably as well. We need not have worried: Both groups seemed to be happier. Interviewers, while concerned that questions did not pop up on their screens fast enough, appreciated the accuracy of the computer and thought that using it made them appear more professional. Respondents who were interviewed in person appeared to be more interested in the survey -- some, for example, invited interviewers to "plug in" -- and paid closer attention to the questions.

Ideally, with changes to the ultimate CPS coming from two directions -- the questionnaire itself and computed-assisted interviewing -- it would have been desirable to isolate the data effects on differences of these changes (questionnaire and interview mode). Alas, this was not possible. As a consequence, the significant differences between the on-going CPS and the PS that were identified can only be ascribed to the overall change in the survey and not specifically to the questionnaire or collection mode. Thus, we have been unable to discern, for example, what the specific effects have been on, say, the overall unemployment rate from changes in the questionnaire wording and question sequencing.

As soon as early PS figures started becoming available to BLS and Census researchers, it became obvious that we were indeed seeing marked changes in important statistical measures. The overall unemployment rate was higher, particularly among women and older workers but essentially across all worker groups. The employed differences were especially interesting, because more women but fewer men were found to have jobs.

There were also other dramatic changes. As expected, the new, more restrictive, measurement of discouraged workers resulted in some 60 percent fewer persons being counted in that category. To be classified as discouraged under the revised scheme, persons who wanted a job but had not looked for work in the prior 4 weeks had to have searched for work during the prior 12 months and not currently looking for work because of discouragement over the job market, while also being available to take a job during the reference week. Similarly, as a result of better question specificity, there was a 20-25 percent reduction in the number of persons working part time for economic reasons (that is, working less than 35 hours during the reference week because of poor business conditions or because of an inability to find full-time work). To be so classified, a person who usually works part time must now indicate that s/he wants a full-time job and was available to take one during the reference week.

V. Communications with official Washington

As soon as the researchers were able to verify that they had accurate data from the PS and thus could estimate the differences that the new questionnaire and collection methodology were yielding, taking appropriate measures of statistical significance into account, it was time to start communicating "up the line." We were, quite naturally, concerned about what kind of reactions there would be to a significantly higher rate of unemployment. There was, after all, a new administration in office that perhaps did not need to be saddled with yet another major issue. It had enough on its plate already.

The researchers had put together a formidable package of tables and analysis, with explanations for the many diverse changes observed over the comparison periods. Initially, 6-month comparisons of the PS with official CPS, covering the period September 1992-February 1993, were utilized, and these were the first figures to be viewed at higher levels, first of all with the heads of the two agencies. Soon thereafter, the first annual average data became available, representing the period September 1992-August 1993. Using these figures, a memorandum detailing the changes that we were expecting to introduce in January 1994 and the expected data effects was sent to Secretary Reich in late October, and this memorandum was forwarded on by him to the President. An hour-long meeting was held with Secretary Reich and top Department of Labor staff on November 1, and this was followed in short order by other high-level briefing sessions with other members of the Administration and the Federal Reserve Board.

Reaction was reasonably swift. All of the changes that were contemplated for implementation in January 1994 were fully acceptable, despite a concern that, with a higher rate of unemployment, the public might fail to recognize that the economy was still gaining steam. Indeed, it was this concern that led directly to a request to sustain the collection of parallel survey data using the paper and pencil methodology beyond the year and a half that had been planned and funded. Monies were found and commitments made to sustain the parallel survey beginning in January 1994, this time with the old questionnaire and procedures. That way, after an initial period where respondents and interviewers might be affected by the previous test, we would have a continuing measurement of the differences that were identified for 1993 as the data on the new basis became available.

VI. Communications with the outside world

A detailed planning document that had been produced and constantly updated had long identified November 16, 1993, as the date of the first public announcement of the plans for introducing changes into the Current Population Survey and detailing what the expected data effects were. Armed with briefing packets and a plethora of other useful information, Commission Katharine Abraham and members of the BLS and Census Bureau staffs presented an extensive array of information to the national economic media. Articles appearing throughout the country the next morning, as well as the more immediate wire service stories, suggested that the press well understood what was transpiring. In particular, the notion of "gender bias," which had emerged from data findings, was significantly played up. With few exceptions, they got it right. Not all did, however, as suggested by the headline, "U.S. won't ask women if they cleaned house."

BLS didn't stop with a one-day media session in Washington, D.C. There was a full-day session with technical users on the next day (November 17), also in Washington, which some 150 persons attended. Interest was running high. In December and January, combined data user-media sessions were held in 13 other large cities throughout the entire country, including New York, Chicago, and Los Angeles. Many people turned out to learn what was expected to happen and how their local unemployment rates might be affected.

By this time, we had unemployment rate comparisons for the Census regions and divisions, as well as some data for seven large states. Our uncertainty was quite high as to the reliability of our sub-national comparisons, and this was carefully communicated. Fortunately, our concerns did not fall on deaf ears, and most people, including the local

media, did not play up some of the wide differentials, such as rates that were slightly more than a point higher in the Middle Atlantic Division (covering New York, New Jersey, and Pennsylvania). This was all the more fortunate when, in actual fact, this particular region did not show large jumps in early 1994. In other words, our concerns about the reliability of sub-national data were well-founded, especially since the PS, unlike the CPS, was not a State-based design and the sample size was so much smaller.

VII. The final outcome and lessons learned

By the time the data for January 1994 were released on February 4th, it seemed that everyone -- government officials at all levels, the media, financial analysts, and the public at large -- was well aware that big changes were to be expected. And thus there were seemingly no surprises. The overall rate of unemployment for January was 0.3 percentage point higher than the December 1993 figure, quite reasonable, given expectations that we could expect as much as an 0.6 increase on an annual average basis (all other things remaining equal), 0.5 from the questionnaire and methodological changes and 0.1 from the introduction of the 1990 census-based population figures (adjusted for the estimated undercount) into the estimation procedures.

Did these data results from January, which were followed by an 0.2 percentage point drop in February, imply that our expectations based on parallel survey results -- the population effect was "guaranteed" -- were too large? Or, did January and then February really show large improvements in the underlying rate of unemployment? Even with two more months in, I think that we are still waiting to answer these two questions with more data. (Isn't that always the case?) It would appear, however, that our expectations for 1994 results for the official figures are essentially accurate, that is, the new questions and methodology suggests that the old questions did a good job of measuring mainstream labor market behavior, but not as well for more marginal types of activities, such as might be typical for certain women, youth, and older persons, for whom more jobseeking **and** more jobholding were found. Now, these missed activities tend to be of a seasonal nature and thus more likely to occur in certain months of the year. January, February, and March are months for which this sort of seasonality is fairly low; it can be expected to be much higher in months like May, June, and July. This implies that we can therefore anticipate higher levels of activity, particularly jobseeking activity, in the spring and summer months. And our seasonal adjustments, which are for the moment necessarily based on experience under the former procedures through the end of 1993, are somewhat "off." So, the answer to the second question regarding the January and February (plus March and April) results would appear to be

that, yes, we were seeing some improvement in the economy, but perhaps not quite as good as implied by these figures.

These early results also suggest that it will be quite some time before we have a full, clear realization as to all of the data effects that have been brought about with the new CPS. Ideally, for example, we should have had a longer lead-time than a year and a half with the parallel panel, so that it could have settled in better and given us more direct comparisons with the official CPS figures. Budget exigencies rarely resolve the hindsight "shouldas." New seasonal patterns are not fully discerned for at least 5 years, and we therefore may have to wait that long to expect to attain a degree of accuracy in month-to-month movements in employment and unemployment that we are fully comfortable with. The seasonal adjustment process will improve over time as data based on the new procedures are gradually taken into account. It is also possible that other improvements could be made to expedite the process.

The seasonality issue just discussed and the potential breaks in series for a number of measures, particularly those of labor market slack, inevitably raise the question as to the whether the process should have been embarked on at all. From my own viewpoint, the answer is clear: Breaks in series and comparatively short losses in time-series comparisons, while never desirable with any degree of frequency, are vitally necessary to ensure that we are accurately measuring what is occurring in our economy. We must recognize that there is always a cost to bring about improvements in data collection of economic phenomena. If we focus only on data consistency and therefore take our eyes off the prize of data improvement in a constantly changing society, we will never even attempt to undertake improvements in the measurement and collection of statistical surveys in the first place. Once undertaken, it is imperative that we go all the way, that is, make all of the improvements that are discernable and viable and then carefully measure their impact through a separate parallel survey. That is precisely what we have done with the CPS, and I firmly believe that the payoff was well worth the short-term losses that we are experiencing.

Perhaps the most significant lesson we learned from all of this was one that was a major winner: full, extensive communication. By careful interaction with, firstly, our internal customers -- i.e., the Administration and the Congress -- and then our external customers -- the media and the public -- there were few, if any, surprises. Friday, February 4th, turned out to be a business as usual, ho-hum day. Everyone knew or thought they knew what was going to happen, their expectations were more or less met, and thus not a lot of news was good news. My recommendation for any statistical agency undertaking major changes in surveys or

data series, therefore, is to err heavily on the side of both extensive and continuing communication with every possible group -- not just the media, not just here in Washington, but with everyone everywhere.

Carrying out the questionnaire-related redesign of the Current Population Survey cost the taxpayers an estimated $30 million. Was it a worthwhile expense? Coming from a highly biased person, one can take my answer with a grain or two of salt, which is a resounding yes! In return, the Nation is getting better, more accurate figures on the labor force activities of the population. It is getting new kinds of important statistics, such as monthly data on labor market discouragement (on a totally revised conceptual basis) and on multiple jobholding. And it is getting the assurance that the measurements of the labor force, employment, unemployment, and those not in the labor force have been carefully studied and researched. And that, I would argue, is an incredibly great bang for our bucks.

A potential lesson that I hope we will not forget is that the total job is not as yet completed. The data comparisons for 1993, based on parallel survey and official statistics, need to be studied much closer than we have been able to thus far. The new figures for 1994 and beyond will require careful analysis. Continuing research on bridging data estimation both back in time and forward in time should continue, with the intent of assisting time series users in their analytical endeavors. We should have learned well the benefits that behavioral science has given us in terms of all future data collection; thus, for example, survey supplements, whether ongoing ones such as income and work experience or new ones such as the upcoming inquiry into contingent work, should be subjected to careful cognitive testing. Lastly, we should carry forward what we have learned into other CPS-related areas -- such as instituting improvements into the "control card," in which demographic characteristics are identified, or in improving the coverage of minority groups in the data collection process.

Finally, the last Presidential commission to examine labor force statistics issued its report 15 years ago. It took that long to implement a couple of its important recommendations. It would not be too radical to suggest that another commission ought to be established in the not too distant future to assess the viability and adequacy of the 1994 changes and then to determine appropriate future directions as we embark on the 21st century.

References

Chester E. Bowie, Lawrence S. Cahoon, and Elizabeth A. Martin, Overhauling the Current Population Survey: Evaluating changes in the estimates, Monthly Labor Review, September 1993.

John E. Bregger, The Current Population Survey: a historical view and BLS' role, Monthly Labor Review, June 1984.

John E. Bregger and Cathryn S. Dippo, Overhauling the Current Population Survey: Why is it necessary to change?, Monthly Labor Review, September 1993.

Thomas B. Jabine, Miron L. Straf, Judith M. Tanur, and Roger Tourangeau, Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines, 1984 (National Academy Press, Washington, D.C.).

Elizabeth A. Martin, "Surveys as Social Indicators: Problems in Maintaining Trends, Chapter 16 to Handbook of Survey Research, 1983 (Academia Press, Inc.).

National Commission on Employment and Unemployment Statistics, Counting the Labor Force, Labor Day, 1979 (U.S. Government Printing Office, Washington, D.C.)

Anne E. Polivka, Comparisons of Labor Force Estimates from the Parallel Survey and the CPS During 1993: Major Labor Force Estimates, CPS Overlap Analysis Team Technical Report 1, March 18, 1994.

Anne E. Polivka and Jennifer M. Rothgeb, Overhauling the Current Population Survey: Redesigning the Questionnaire, Monthly Labor Review, September 1993.

President's Committee to Appraise Employment and Unemployment Statistics, Measuring Employment and Unemployment, 1962 (U.S. Government Printing Office, Washington, D.C.)

Jennifer M. Rothgeb, Revisions to the CPS Questionnaire: 1994 Effects on Data Quality, CPS Overlap Analysis Team Technical Report 2, April 6, 1994.

Jenny Thompson, Mode Effects Analysis of Labor Force Estimates, CPS Overlap Analysis Team Technical Report 3, April 14, 1994.

COMMENTS ON PARKER AND WEADOCK, TIME SERIES REVISIONS: THE EFFECTS ON GROSS DOMESTIC PRODUCT

Murray F. Foss
American Enterprise Institute

I. Introduction

This paper by Robert Parker and Teresa Weadock is an interesting study that extends the series of evaluations of quarterly GDP estimates undertaken in BEA by Allan Young in 1974 (and most recently in the October 1993 Survey of Current Business) and initiated by George Jaszi a decade earlier.

In my comments I will discuss some differences between the current and constant dollar figures. I suggest other things the authors might have looked at, some of which would be quite easy. I then raise the question about which figures are the best ones for gauging "the true quarterly change," a point implied by the authors in their criticism of source data. And then I ask what lessons we should learn from all of this.

II. Lack of symmetry between current and constant dollar measures

The Parker-Weadock (henceforth PW) measures of reliability are typically presented in terms of current dollars and constant dollars. We are interested in both but the two sets of figures are really not symmetrical, and I wish the authors had discussed this asymmetry. While the monthly PPI data are revised once going back 4 months as a result of the incorporation of late returns and the correction of errors by respondents and by BLS in the initial reports, this is not true of the current monthly CPI. Yet if we look at the succession of dispersion measures in Table 2 from advance to preliminary to final, for GDP as a whole as well as for personal consumption the dispersion gets worse in the constant dollar series.

Aside from making very few revisions on a monthly basis it is not the practice of BLS to conduct a bigger survey after the calendar year is over--what might be an Annual Survey of Prices, analogous to, say, the Annual Survey of Manufactures. Every ten years BLS changes its market basket for the CPI to take account of changes in consumption patterns. New and different products are appearing on the market constantly, and these BLS treats in a variety of ways, depending on continuing probability sampling to pick up new products and types of outlets.

It would be good to know whether this deterioration in the reliability of successive constant dollar estimates is statistically significant. Is it simply a reflection of newer seasonal factors, which are revised by BLS each year going back several years, and to what extent does it reflect a benchmark (10-year) change?

III. Some other perspectives of reliability

PW, like Allan Young, examine the reliability of the quarterly estimates of GDP from an historical point of view. That is useful because it permits one to say something about possible long-term trends in reliability. But outlined below are other ways that I would like to see examined.

A. Business cycle perspective

The GDP statistics are the single most important indicator about what is going on in the national economy. But as we all know economists are often in the dark about whether an expansion has begun or whether the economy has slipped into a recession. So I would like to see how these reliability measures--both dispersion and bias-- behaved around turning points. Looking at the historical record we can ask if there are any patterns, for example, in the four quarters up to and including the business cycle peak and in the first four quarters of downturns (which average not quite a year in length). Do these patterns differ from one another? Would they differ from the pattern in the first four quarters of an upturn? It would not be hard to find a rationale for any differences that might turn up; for example something concerning the quality of statistics within the firm over the business cycle but any patterns would be of interest in themselves.

B. Inflation

It would be interesting to examine the data for possible differences when the rate of inflation differed. Is there a difference between 1973-80, when inflation was very high and 1983-90, when inflation was much lower? It is more difficult to capture a change in real output when inflation is high than when it is low. When buyers resort to new sources of supply or when sellers change their discounts from list prices, the Producer Price Indexes may be slow to adapt even though the current dollar figures on sales reflect these changes immediately.

C. The current data

The first three estimates of a given quarter--the advance, preliminary and final--carry a lot of weight because these are the figures that affect decisions by business and government in the short run. So one could use this criterion: given the advance and the final, how often did the preliminary move in the direction of the final (third) change? For example, if the advance change is 1.9 and the final is 1.5, we can ask if the preliminary moved down from 1.9 or exceeded it. Small misses in direction would be ignored, following the authors' approach.

D. Calendar quarters separately

It would also be a simple matter to collate the measures of dispersion and bias by calendar quarters, to see if the fourth quarter differs from the other three quarters. The reasoning behind this is that in organizing their work accountants put most emphasis on the annual report, which comes out a few months after the end of the year. (I realize that fiscal years pose some problem). Accountants do things at the end of the year that they don't do during the year. For example, they may take a physical count of inventories at year-end but use shortcut methods to estimate inventories for months and quarters. Earlier errors that show up at the end of the year are corrected in the final quarter. If the errors are in one direction the fourth quarter correction will tend to be reduced if not eliminated by seasonal adjustments. After year-end, accountants may go back and revise earlier quarterly figures;this may be a regulatory requirement but I am not certain. If accountants in fact do a lot of estimating during the year such a practice could give rise to revisions between advance-preliminary, on the one hand, and final-first annual, on the other.

E. Final sales and inventory change

Estimating the quarterly change in business inventories is an inherently difficult task and remains so even with the many improvements made by the Census Bureau and BEA over the years. The change in inventory change is ordinarily a significant part of the average quarterly change in GDP. The inventory estimates are not shown explicitly by the authors because of the particular measures they employ for GDP and all other components. It would seem from table 4 that revisions in inventories are a significant source of total revisions. It would be a good idea to examine a common measure published by BEA, namely, total final sales, which excludes inventory change. Obviously the shifting of farm inventories between the Federal Government and private business creates a problem but it would not seem to be too difficult to make allowance for this.

As a matter of fact, the change in business inventories ought to be shown explicitly with its own reliability measures because it is so difficult and involves much judgment not only by BEA but also, I would guess, at the firm level. This suggests an additional reason why the authors should show successive revisions of nonfarm CBI: the monthly CBI's are subject to far greater variation than any of the flow components. Maybe exports and imports as now calculated would be close runnersup. PW and Allan Young point out that the final month in the quarter has a weight of only one-ninth and the second and first months weights of two-ninths and three-ninths, respectively. If the expenditure components were random numbers, the fact that they have such "small" weight would not be so important. The fact is that this month's seasonally adjusted retail sales must be very close to last month's. A one percent

difference is a big seasonally adjusted change. But that is not true of CBI. One month of inventory change can be positive, the next month, negative. Such a pattern is possible because sales can be higher or lower than expected by the firm and, with production plans based on expected sales, inventories will be correspondingly lower or higher. The same is true of incoming supplies to the purchasing firm--a consequence of capacity limitations, strikes, natural disasters, etc. This is not to deny that during the expansion firms tend to build stocks and during the contraction they cut them. But I urge the authors to do the dispersion measures of nonfarm CBI (and the GDP) in constant dollars. This would be a good test.

IV. What is the "true" quarterly change?

Given the way quarterly data are revised to make them compatible with subsequent annual figures and benchmark annual totals from the quinquennial censuses, how can we be sure that the very final quarterly pattern that emerges is superior to all previously published quarterly data for a given year? Parker and Weadock criticize the Census Bureau for making proportional adjustments in originally published monthly and quarterly data. This is an old problem. For example, Morris Cohen raised the same issue at an Income and Wealth Conference 15 years ago. He said that the data were being oversmoothed and that cyclical fluctuations were being damped if not eliminated. The late Otto Eckstein agreed with this point of view but it remains a minority opinion. The dominant revision philosophy is above all to get the long-term trend correct. If that is so, cyclical fluctuations must be fitted into the trend (that is, benchmark) values for a given year.

The answer, of course, is to get more and better within-the-year data.There is no substitute for this. That was said 15 years ago and, I am sure, many times before that. People who make decisions in business and government have a big stake in the currently available quarterly numbers and after some 50 years deserve more improvements than the agencies have made. Economists studying the business cycle have an important interest in getting the record straight.

V. What lessons should we learn from these studies?

Parker and Weadock, like Allan Young, raise the possibility of dispensing with the second and third quarterly estimates for a given quarter. A single current estimate, namely, the advance, might conceivably save some money. I am not sure that the nation would be better off. As Allan Young points out, the detailed estimates might suffer. As for the total, there are so many people and firms engaged in this business today that several estimates would make their appearance to fill the void left by BEA. Unfortunately these estimates would differ from one another. I would guess that large organizations like the Federal Reserve would

make their own estimates. Estimates of GDP made by outsiders are not likely to be as good as those made by BEA. So I am inclined to stick to the present system. It is less bad than what might supplant it.

Parker and Weadock are impressed by the improvement in reliability from final current to the first July estimate. They should, of course, find out how much of the improvement is due to better seasonals. But they should not be surprised that firms send in better data after the year is over.

I think that both the source agencies and BEA should do more field work to find out why numbers submitted to the government change. The agencies ought to do more to find out how firms obtain their monthly and quarterly numbers, the extent to which they rely on within-the-firm estimates and end-of-year adjustments. Sales probably are not much of a problem. But other statistics like inventories and profits may be. Irving Rottenberg and I found that firms using LIFO accounting had great trouble estimating inventories on a monthly basis. The fact is that firms use shortcuts of all sorts; these are doubtless useful to the firm but may be subject to biases that are functions of the stage of the business cycle or the rate of inflation. I don't have in mind huge field surveys checking data quality. One can get many insights from small field trips. I believe that BEA at least does too little in this regard. And, to repeat, Census and BLS ought to do better in obtaining within the year figures to accompany more comprehensive annual and benchmark surveys.

Reports Available in the Statistical Policy Working Paper Series

1. Report on Statistics for Allocation of Funds (Available through NTIS Document Sales, PB86-211521/AS)
2. Report on Statistical Disclosure and Disclosure-Avoidance Techniques (NTIS Document sales, PB86-211539/AS)
3. An Error Profile: Employment as Measures by the Current Population Survey (NTIS Document Sales PB86-214269/AS)
4. Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics (NTIS Document Sales, PB86-211547/AS)
5. Report on Exact and Statistical Matching Techniques (NTIS Document Sales, PB86-215829/AS)
6. Report on Statistical Uses of Administrative Records (NTIS Document Sales, PB86-214285/AS)
7. An Interagency review of Time-Series Revision Policies (NTIS Document Sales, PB86-232451/AS)
8. Statistical Interagency Agreements (NTIS Documents Sales, PB86-230570/AS)
9. Contracting for Surveys (NTIS Documents Sales, PB83-233148)
10. Approaches to Developing Questionnaires (NTIS Document Sales, PB84-105055/AS)
11. A Review of Industry Coding Systems (NTIS Document Sales, PB84-135276)
12. The Role of Telephone Data Collection in Federal Statistics (NTIS Document Sales, PB85-105971)
13. Federal Longitudinal Surveys (NTIS Documents Sales, PB86-139730)
14. Workshop on Statistical Uses of Microcomputers in Federal Agencies (NTIS Document Sales, PB87-166393)
15. Quality on Establishment Surveys (NTIS Document Sales, PB88-232921)
16. A Comparative Study of Reporting Units in Selected Employer Data Systems (NTIS Document Sales, PB90-205238)
17. Survey Coverage (NTIS Document Sales, PB90-205246)
18. Data Editing in Federal Statistical Agencies (NTIS Document Sales, PB90-205253)
19. Computer Assisted Survey Information Collection (NTIS Document Sales, PB90-205261)
20. Seminar on the Quality of Federal Data (NTIS Document Sales, PB91-142414)
21. Indirect Estimators in Federal Programs (NTIS Document Sales, PB93-209294)
22. Report on Statistical Disclosure Limitation Methodology (NTIS Document Sales, PB94-165305)
23. Seminar on New Directions in Statistical Methodology (NTIS Document Sales, PB95-182978)

Copies of these working papers may be ordered from NTIS Document Sales, 5285 Port Royal Road, Springfield, VA 22161 (703)487-4650

Part 2

Session 6

INCENTIVES IN SURVEYS

TIME, DOLLARS, AND DATA: SUCCEEDING WITH REMUNERATION IN HEALTH SURVEYS

Trena M. Ezzati-Rice, Andrew A. White, William D. Mosher
National Center for Health Statistics, CDC

Maria Elena Sanchez
Agency for Health Care Policy and Research

Introduction

All surveys experience nonresponse in spite of the fact that the goal for all sample surveys is to maximize survey response, and thus minimize nonresponse bias in the survey estimates. Over the years, survey incentives (both monetary and non-monetary) have been used in conjunction with other survey methodologies to obtain complete and accurate information for the largest number of sample units.

Incentive use in surveys has spanned a wide variety of survey types, sponsors, respondents, and survey topics. As the title suggests, we present and discuss the use of remuneration (monetary incentives) in selected health surveys conducted by the National Center for Health Statistics, CDC (NCHS), the Agency for Health Care Policy and Research(AHCPR), and Project HOPE.

Experimental and field results have demonstrated that incentives can have a positive impact on survey response rates, and we will concentrate on describing selected incentive applications and experiments. Included here are new results from a field trial of remuneration in NCHS's National Survey of Family Growth, Cycle 5 Pretest.

Additional information in the area of remuneration in health surveys not reviewed in this paper can be found in the literature[1, 2, 3, 4]. In particular, the paper by Kulka

---

[1] Kulka, R.A. (1992, October). A Brief Review of the Use of Monetary Incentives in Federal Statistical Surveys, presented at the COPAFS Symposium on Incentives in Surveys, Boston, MA.

[2] Willimack, D.K., Petrella, M., Beebe, T., and Welk, M. (1992, August). The Use of Incentives in Surveys: Annotated Bibliography, Survey Research Center, Institute for Social Research, University of Michigan.

included in this volume gives a brief background on the use of monetary incentives and references numerous articles this topic across various types of surveys. Kulka also addresses sociological models proposed to describe the operation of incentives. The paper by Willimack et al. is an unpublished review of published literature however they do make the observation : "The bulk of the published literature regards the effects of incentives in mail surveys. No doubt incentives have been and/or are being used in both telephone and face-to-face surveys, but there is a lack of documentation of tests in the published literature. Perhaps incentives have been implemented in telephone and face-to-face surveys based on the mail survey results and on 'common sense,' without specific testing within mode. Based on a brief look at conference abstracts, it appears that documentation of incentive testing and/or use in survey modes other than mail may be found in non-published literature, such as professional association conference presentations."

Overview of Health Survey Issues and Remuneration

The use of monetary incentives is not new in health surveys and over the years those of us working in the field have learned a great deal about their use. Much of the material in this section was presented at an October 1992 COPAFS/OMB Symposium on Providing Incentives to Survey Respondents held at the John F. Kennedy School of Government, Harvard University, Cambridge, MA.

Although most surveys conducted by the Federal Government are based on unpaid, voluntary participation, monetary or gift incentives for participation may be justified for certain types of household and establishment based surveys in order to increase participation rates, encourage accurate record keeping, and/or keep expenses down. In addition for health surveys, remuneration may be justified for surveys which involve a physical examination and the drawing of a blood sample in order to maximize response rates.

High response rates bring the benefits of increased validity through increased precision and reduced potential for bias in survey estimates Incentives and

---

[3] Kulka, R. A. (1994, May). The use of Incentives to Survey "Hard-to-Reach" Respondents: A Brief Review of Empirical Research and Current Research Practice, presented at the COPAFS Seminar on New Directions in Statistical Methodology, Bethesda, MD.

[4] Dillman, D. A. (1991). The design and administration of mail surveys. Annual Review of Sociology, 17, p. 225-249.

remuneration can be considered appropriate whenever respondents are asked to devote time and effort (sometimes a considerable amount) to assisting the government in obtaining high quality data for research and policy related issues which will impact the entire nation.

For population based surveys, incentives can result in a higher motivation to participate, increased effort to give accurate (honest) responses, greater acceptance of government surveys, increased response rates, lower item nonresponse rates, and cost savings through better data quality and fewer call backs and/or canceled appointments. Further, cash incentives enhance the importance of the survey to respondents and provide tangible evidence of the value of their input. Incentives may stimulate otherwise reluctant respondents to participate and to more readily make themselves available to participate. Gift incentives for children could reinforce the value and importance of their participation.

The largest potential for net benefits from remuneration or incentives will be found: among surveys that experience higher refusal or item non-response rates; among surveys where persons at higher risk of being targeted respondents are less likely to participate (thus biasing the results); in situations where respondents are easy to locate, but initial or continuing cooperation is hard to gain; among those surveys requiring repeated contact in a short time period (like a month) or with other significant burden involved (like a physical examination or drawing of blood); or among those that require respondents to do something on their own, like complete a mail questionnaire or keep diaries, or participate in a followup survey of initial nonrespondents.

In particular incentives are most likely to have an effect in surveys that: require the respondent to travel; are lengthy or have a longitudinal component; are focussed at hard to reach populations (like adolescents or young black males), or that ask questions about sensitive topics (like income, drug use, risk behaviors related to HIV/AIDS). Remuneration may also gain survey participation when sample persons do not perceive an immediate benefit to themselves and/or society by participating.

Preparation and pilot testing of questionnaires (especially laboratory based testing) often requires respondent travel and takes large blocks of respondent time. The potential benefit to the survey is so large from this type of testing that remuneration is well worth it.

For institutional surveys, incentives can result in money and time savings (e.g., hospital versus government staff abstracting hospital records), removal of a barrier to participation (financial loss), greater acceptance of government surveys, cost savings

through fewer call backs and/or canceled appointments, increased response rates, and for some businesses, lower item nonresponse rates, and better data quality.

Remuneration in institutional surveys is often seen as compensation rather than as an incentive to participate in voluntary surveys. Institutions think of time as money and may consider compensation a requirement to engage in substantial continuing activities. Some may set that standard for any voluntary survey. In this sense, remuneration may make a survey possible. In addition to participation, remuneration to the institution can be cost beneficial to the government if summary data must be compiled from business records for survey purposes. The institution can often do it for less than having government field representatives comb the records. In businesses where staff energies must be diverted to complete a survey, the thoroughness and accuracy of the response may be improved if the business is being remunerated.

In institutional surveys, the continuing nature of data collection from particular respondents is a critical factor. There may be no other way to obtain the data (e.g. the is a charge for access to hospital records and state vital records already exist, why should the government be excepted?). Remuneration may actually be less expensive than providing the person hours needed to compile the data from institutional records even if the institution is willing to grant access.

A review of these issues and the use of incentive methods took place at the 1992 COPAFS/OMB Symposium[5]. This led to a recommendation, "that OMB seriously consider an agency's request to use incentives in a limited number of specific situations in which a survey violates the norm of what is considered the standard survey." The participants defined a standard survey as: a cross-sectional survey of the household population done in about an hour in a single session at the respondent's convenience; and done in the respondent's home with non-intrusive, nonsensitive questions. Symposium participants suggested that incentives be considered in the following special situations:

Surveys of hard-to-reach or special population subgroups;

Surveys involving unusual demands or respondent intrusions such as -

---

[5] Council of Professional Associations on Federal Statistics (COPAFS). (1993), September). Providing incentives to survey respondents: Final report (Contract No. GS0092AEM0914). Washington, DC: Regulatory Information Service Center, General Services Administration.

Lengthy interviews,
Keeping a diary,
Taking physical or physiological tests, or
Going somewhere special to participate;

Surveys involving sensitive questions and/or topics;

Surveys involving a commitment to participate over time such as for a panel survey; and

Surveys where respondents are not a household respondent such as a physician, hospital, or nursing home.

Specific Experiences with Remuneration in Health Surveys

This section highlights some major institutional and population based surveys which successfully use or have used remuneration to increase survey response and/or data quality. For some of the summaries we have paraphrased or reproduced language in the original references for the sake of accuracy.

NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY

The National Health and Nutrition Examination Survey (NHANES) is a periodic survey designed to assess the health and nutritional status of the noninstitutional population of the United States. The NHANES consists of a household interview with adult, youth, and family medical history questionnaires, followed by a 3-4 hour standardized physical examination in specially equipped mobile examination centers (MEC's). The NHANES is based on a stratified multistage cluster probability sample design[6]. The on-going Third NHANES or NHANES III is the seventh in a series of surveys using health examination procedures that have been conducted since 1960 by NCHS.

As for most large-scale Federal surveys, the success of the NHANES surveys depends upon voluntary participation of individuals selected in the sample.

---

[6] Ezzati, T., Massey, J., Waksberg, J., Chu, A., and Maurer, K. (1992). Sample Design: Third National Health and Nutrition Examination Survey. National Center for Health Statistics. Vital Health Statistics. Series 2, No. 113.

Remuneration has been used in all NHANES surveys and has been shown to be necessary for attaining adequate response rates. For the three early Health Examination Surveys (HES) conducted in the 1960's of adults, children, and youths, respectively, the examination response rates were excellent ranging from 87 percent to 96 percent[7]. However, with the beginning of NHANES I(1971-74), the examination response rate (64 percent) was much lower than those in the earlier HES surveys. After extensive efforts to improve the miserable response rate through interviewer re-training, increased publicity, and community outreach, the response rates remained at an unsatisfactory level. Therefore, it was proposed that a monetary incentive be considered to reduce examination nonresponse[8]. Since there was little information available from previous studies to show the effect of paying respondents to participate in health surveys, a field experiment was undertaken. First, of course, justification had to be submitted to OMB and approval obtained. It was reasoned that remuneration for NHANES was justified since participation in the survey required several hours of the respondent's time (thus lost time from work) and paid assistance for child care might also be required. It was hypothesized that the cost of the remuneration would be offset by a reduction in the number of contacts to a household to obtain respondent participation. Further, if the response rates increased significantly, the overall validity of the survey results would outweigh the remuneration costs.

The experimental design for the study was superimposed upon the within primary sampling unit (PSU) design for NHANES I[9]. The study was undertaken during 1972 in the San Antonio, Texas, PSU. The segments within the PSU were randomly paired by segment size and median family income. All of the sample persons in one segment of each pair were told about the $10 remuneration. The sample persons in the other segment of the pair were not told of the $10 remuneration. It should be noted however, that all persons who were examined received $10. The difference was that persons in the "not told" segments did not know about the remuneration until after they had been examined, while those in the "told" segments knew of the $10 remuneration before being examined.

Telling sample persons that they would be given $10 after completing the examination phase of the survey had a positive effect on the response rate in San

---

[7] Bryant, E., Kovar, MG., and Miller, H. (1975). A Study of the Effect of Remuneration Upon Response in the Health and Nutrition Examination Survey. National Center for Health Statistics. Vital Health Statistics. Series 2, No. 67.

[8] Ibid.

[9] Ibid.

Antonio. Among the 303 persons in the experimental segments (told) who were contacted, 82 percent were examined (see Table 1). On the other hand, among the 292 persons in the control segment (not told), only 70 percent were examined. Thus, the NHANES I experiment showed that the offer of $10 to sample person increased the response rate by 12 percentage points. Also, there was some evidence that sample persons were more cooperative and that less effort was required to obtain response when remuneration was offered as evidenced by the number of persons making an examination appointment at the first interview contact. Also, a larger proportion of the "told" group kept their appointments than the "not told" group. Only 2.1 contacts per examined person were required for the "told" group as compared with 2.5 such contacts per examined person for the "not told" group.

Table 1. Examination Response Rates from NHANES I Remuneration Experiment (sample sizes shown in parentheses)

	Not Told of Payment	Told of Payment
Experiment	70%	82%
	(292)	(303)
Actual Survey*	68%	78%
	(7335)	(6035)

*Represents results from the first 35 stands of NHANES I.

The findings of the NHANES I remuneration study were considered conclusive enough to include remuneration in the remainder of NHANES I. The overall response rate at the 45 survey locations where remuneration was offered in NHANES I (including San Antonio) was 78 percent as compared to 68 percent for the 19 survey locations where remuneration was not offered.

The $10 remuneration used in NHANES I was continued in NHANES II. However, about midway through NHANES II, the response rate dropped to about 70 percent. Therefore, another field experiment was used to assess the effect of increased remuneration on survey response. The study for NHANES II (1976-80), looked at the impact of increasing the $10 remuneration used in NHANES I to $20. Three survey locations were selected for the study. Each segment was paired with another segment similar to it with regard to poverty/non-poverty status and distance from the

examination center. Sample persons in one of each pair of segments were "told" that they would receive $20 for participating in the examination, while the sample persons in the other paired segment were told they would receive $10. However, all persons who were examined were given $20 no matter which monetary amount they were originally told..

There was a significant positive effect on response rates with the $20 incentive[10]. Out of 720 persons offered $20 for participation, 79 percent were examined (see Table 2), while among the 716 persons offered $10, only 74 percent were examined. An important finding from NHANES II which did not show up for NHANES I was the increase in the response rate by number of persons in a household for the $20 group versus the $10 group. There was no effect in households with only one sample person. However, in households with two and three or more sample persons, payment of $20 rather than $10 increased response rates by 8 and 16 percentage points, respectively[11].

Table 2 Examination Response Rates from NHANES II Remuneration Experiment (sample sizes shown in parentheses)

	Told of $10 Payment	Told of $20 Payment
Experiment	74%	79%
	(716)	(720)
Actual	72%*	76%**
	(NA)	(NA)

*Represents the final 44 stands of NHANES II.
**Represents the first 16 stands of NHANES II.

[10] Findlay, J. and Schaible, W.L. (1980) A Study of the Effect of Increased Remuneration on Response in a Health and Nutrition Examination Survey. Proceedings of the Section on Survey Research Methods of the American Statistical Association, pp. 590-594.

[11] Ibid.

The results indicated that there was a potential cost saving associated with the fact that the higher paid group was generally more cooperative. For example, it was found that a larger proportion of the examined respondents in the higher paid groups went to the examination center as a result of the first contact (68 percent versus 61 percent). Also, the higher paid group required, on average, fewer contacts overall (1.77 vs. 2.09)[12].

Overall, for NHANES II, for the 44 survey locations at which $10 was offered to sample persons, the response rate was 72 percent. But for the 16 stands at which $20 was given, the response rate was 76 percent.

In the Puerto Rican phase of Hispanic HANES (1984), the response rates were unsatisfactory, so an increase in remuneration from $20 to $50 was made. A cross-tabulation of response rates (see Table 3) by method of payment before and after the increased remuneration was instituted indicated a higher response rate in the $50 group (83 percent) than for the $20 group (76 percent)[13]. The results, however, should be viewed with caution since the increased payment was not randomized within the various survey locations.

Table 3. Hispanic HANES Examination Response Rates for New York City Metro Area (Puerto Rican Phase) Stands by Payment Amount*, 1984

Payment Amount	Examination Rate
$20	76% (3101)
$50	83% (576)
DK**	0% (116)

*Unpublished data from J. Findlay, NCHS.
**Cases could not be classified into either payment category due to lack of information.

[12] Ibid.

[13] Unpublished data from Jean Findlay, National Center for Health Statistics.

In the on-going NHANES III, the basic remuneration payment is $30 for all persons who come to the examination. However, for adults 20 years and older there is an additional incentive of $20 if they come for their examination at the "right time". Time of day of the examination and fasting status need to be controlled for the analyses of many of the biochemical tests including the Oral Glucose Tolerance Test (OGTT - a test for diabetes). Thus a random half-sample of adults is designated to have blood drawn in the morning, while the remaining 50-percent is selected for afternoon or evening appointments.

The NHANES III which includes an oversample of both Mexican-Americans and Blacks and includes no upper age limit (in contrast to previous NHANES) is experiencing higher examination response rates than any other previous NHANES survey. There is a six percentage difference in the overall examination response rate between NHANES II and NHANES III-Phase 2 (79% in the on-going NHANES III-Phase II versus 73% in NHANES II - see Table 4). This increase is due not only to the increased remuneration amount but also to the fact that NHANES III samples multiple persons per household and includes on average two persons per household. Clearly, there is a strong monetary incentive for a household as a whole if several members are selected into the sample.

Table 4. Examination response Rates for NHANES II, Hispanic HANES, and NHANES III

	MEC Examined	MEC + Home Examined
NHANES I	74	NA
NHANES II	73	NA
HHANES	73	NA
NHANES III - PHASE 1	77	78
NHANES III - PHASE 2*	79	80

*On-going.

NHANES III examination of nonresponse rates by household size show that the nonresponse rates decrease significantly with increasing household size. Another important finding related to participation in NHANES III is the increase in the

response rate for the OGTT. This test requires the sample person to fast 10-16 hours and requires two blood draws. The OGTT response rate in NHANES II among examined persons was only 66 percent thus requiring the results to be interpreted with extreme caution. However, for NHANES III, the OGTT response rate, again among examined persons, has increased by 11 percentage points (77 percent response rate for phase 1) and by 23 percentage points so far for phase 2 (89 percent response rate)[14]. The increased remuneration in NHANES III is felt to be responsible in large part for this important increase in the OGTT response rate for NHANES III.

NHANES Summary

Previous research associated with NHANES I and NHANES II has involved several field experiments to assess the impact of monetary incentives on achieving an adequate response rate so reliable estimates can be produced and generalized to the total population. The NHANES I remuneration experiment generally showed that paying sample persons $10 rather than nothing at all increased response rates about 12 percent.

The NHANES II study looked at the effect on the response rate of increasing the remuneration from the $10 used in NHANES I to $20. The average increase in the response rate was about 5 percent.

The results from both the NHANES I and NHANES II had two other findings pertinent to remuneration. First, the results indicated that there were potential cost savings associated with the fact that the higher paid groups were generally more cooperative. For example, it was found that a larger proportion of the examined respondents in the higher paid groups went to the MEC as a result of the first contact. They also required, on the average, fewer contacts overall. They also had fewer broken appointments.

Further, both of these experimental studies showed a marked increase in cooperation in households with more than one sample person, and the differences between the remuneration groups went up with household size. A positive relationship between household size and response rate was also observed in the Hispanic HANES. This trend is continuing in NHANES III as well. These results, presumably, are due to the fact that the monetary award increases substantially for the household as a whole when multiple persons are selected per household. Finally, the response rates in

[14] Unpublished data from Meena Khare, National Center for Health Statistics.

NHANES III are higher than those for previous NHANES surveys and the response rates for the two minority subgroups are higher than for whites and all others. In addition, the response rate for the OGTT in NHANES III is higher than for NHANES II.

The NHANES experiences demonstrate the power of remuneration to stimulate respondent participation in the face of inconvenience and financial disincentives; and in addition to motivate respondents to provide highly personal and sensitive information including physical and biological characteristics. The use of remuneration in NHANES III shows how phased payments can improve scheduling logistics as well. The remuneration serves as a motivator, and also to offset financial disincentives, and gives evidence of the importance of participation and survey goals. NHANES data have been significantly improved through incentive use.

NHANES I EPIDEMIOLOGIC FOLLOWUP STUDY

As mentioned in the previous section, examinees were paid $10 to participate in the baseline NHANES I study. They were again paid $10 to participate in the first NHANES Epidemiologic Followup Study (1982-84 NHEFS interview). Although no remuneration was paid in the 1986 followup of the elderly (conducted using computer assisted telephone interviews: CATI), subjects with reported hospital or nursing home admissions in the 1987 Followup were paid $5 as an incentive to sign and return the Medical Authorization Form (MAF). Payment was needed in the 1987 Followup because of the sharp decrease in the rates at which MAF's were returned in the 1986 Followup. Remuneration of the 1987 respondents increased the MAF return rate in the elderly group by 11 percentage points from 75.0% in the 1986 Followup to 85.7% in the 1987 Followup.

For the 1992 wave, a $5 remuneration was made subject to reported hospital or nursing home admissions since the last contact. Respondents are interviewed by telephone. However, the form which authorizes the hospital or nursing home to release patient information is mailed to the respondent and must be signed and returned. The remuneration is paid to the respondent or to the individual who signs the MAF, if this person is different from the respondent, as an incentive to return the signed authorization.

In addition, due to the difficulty of gaining the cooperation of some hospitals and nursing homes, if a nursing home or hospital requests reimbursement for work performed in abstracting or photocopying selected information from the admission and discharge records or abstracts, a small amount of money is provided to defray their

expenses. During 1987, a total cost of $2,170 or .0012 percent of the total contract cost was paid. This remuneration mostly involved nursing homes.

The NHEFS uses remuneration to gain commitment to continuing participation in a longitudinal study.

NATIONAL HEALTH INTERVIEW SURVEY YOUTH RISK BEHAVIOR SUPPLEMENT

In 1991 the Survey Research Center of the University of Michigan and the Bureau of the Census helped the National Center for Health Statistics assess the impact of financial rewards on respondent participation and motivation in a Youth Risk Behavior Supplement (YRBS) to the 1992 National Health Interview Survey (NHIS). This study[15] employed cognitive interview techniques and a traditional split sample field experiment.

The YRBS contained a number of sensitive questions requiring answers that could be perceived as being socially undesirable or threatening to self-image. Such topics included questions on the use of alcohol, drugs, sexual activity, over- or underweight status, and exercise program involvement.

Interviews were conducted both with groups of parents and youth before field trials and a pretest were conducted. The field trials were then held where youth were interviewed in their homes, paid $20, and debriefed after the interview about their views on paying respondents for participating in surveys. The pretest of the full survey procedures included a split sample in which the interviewers mentioned the $20 payment half of the respondents and did not mention the payment to the other half.

In the split sample experiment, complete interviews were obtained in 90% of the households where the $20 was mentioned, but in only 79% of the remaining households. The cognitive interviews conducted with the groups and during the field trial debriefings helped the researchers develop insight concerning reasons for this difference in response and for a potential increase in data quality when payment is

---

[15] Kalton, G., Cannell, C., Camburn, D., Oksenburg, L. and Holland, L. (1991). The Effect of Financial Incentives on Respondent Participation. Final report of the Association of Schools of Public Health Cooperative Agreement: "Applied Research on the Conduct of Adolescent Health Behaviors and Characteristics", University of Michigan, pp. 47-56.

mentioned in advance.

The cognitive interviews indicated that unwillingness to report events or behaviors is only partially caused by concerns over privacy or confidentiality. In fact, youth respondents could see no reason to be diligent in answering survey questions. They voiced skepticism about any benefits that may accrue from survey participation. Therefore techniques typically used with adult respondents, such as appealing to their obligations as citizens or emphasizing the societal benefits may be unproductive. The researchers concluded that a $20 remuneration offer, linked with a signed commitment to participate and give accurate answers, would be an effective way to motivate respondents to participate and report accurately. The debriefing interviews confirmed this conclusion.

A further interesting note from this study is that the success of respondent payments in obtaining YRBS interviews may be partly due to the reaction of interviewers to the payments. During interviewer debriefings, their comments suggested that respondent payments also have a forceful, positive influence on the attitudes and expectations of interviewers. Interviewers with such a positive outlook may feel they are likely to obtain an interview, rather than expecting a refusal, prior to contacting a potential respondent and subconsciously may convey to potential respondents a more positive view of the YRBS study. The researchers postulate that the total impact of respondent payments on participation rates is the sum of the positive direct effect on respondents and the indirect effect that payments have on the attitudes of interviewers.

YRBS incentive use focused on the motivational aspects of incentives for a non-traditional target population in a study that collected sensitive data.

NATIONAL HOUSEHOLD HIV SEROPREVALENCE SURVEY FEASIBILITY STUDY

A feasibility study for a National Household HIV Seroprevalence Survey (NHSS), based on a probability sample of households was conducted in Dallas County, Texas, in the fall of 1989. One of the major concerns of a household survey attempting to estimate the prevalence of HIV infection is that a high proportion of persons who are at the greatest risk of HIV infection may refuse to participate. This possibility of differential rates of response between those at higher risk and lower risk of HIV infection means that estimates derived from the survey have the potential to be biased. Among procedures to maximize the response rate in the NHSS, a $50 incentive payment was provided to all sample persons who provided a blood sample to be tested

for HIV antibodies and completed a self-administered risk behavior questionnaire. The $50 incentive seemed to have a positive impact on survey participation with a higher than expected response rate for a highly sensitive survey involving the collection of HIV risk behavior data and the collection of a blood specimen in the home. The response rate for the combined questionnaire and blood sample was 84 percent, and 90 percent for the questionnaire only[16]. Respondents in the NHSS were asked to check all reasons for their participation in the survey. Among all respondents, 47 percent stated "helping with AIDS research", while another 39 percent stated the "$50 payment" (see Table 5).

Table 5. Reasons for participation in the Dallas County Household HIV Survey

Reasons	Percent
Helping with AIDS research	46.8
U.S. Public Health Service Sponsorship	2.9
$50 payment	38.5
Videotape presentation	1.2
Assurance of privacy	3.3
Other factors, unspecified	7.2

In addition to estimating the prevalence of HIV infection in Dallas County, another major objective of the survey was to evaluate various methods for assessing and reducing nonresponse bias. A standard survey method for assessing bias due to nonresponse is to conduct a followup survey with a sample of initial survey nonrespondents with different incentives for participation. In the NHSS, a special followup study of a sample of nonrespondents was conducted in which half of the sample persons were offered an incentive of $100 to complete the self-administered risk behavior questionnaire only, and the other half was offered $175 to complete both the questionnaire and provide a blood sample. The followup survey of nonrespondents increased the questionnaire only response rate by 10 percentage points (80% vs. 90%)

[16] National Household Seroprevalence Survey Feasibility Study Final Report. Research Triangle Park, NC: Research Triangle Institute, April 30, 1990. Research Triangle Report RTI/4190-01F.

and the blood and questionnaire by four percentage points (80% vs. 84% - see Table 6). Of particular importance was the increased reporting of risk behavior in the follc vup study. The prevalence rates among male respondents for three major HIV risk behaviors (intravenous drug use, receptive anal intercourse, and multiple sex partners) were 3 to 5 times higher in the followup survey than in the regular survey (see Table 7).

In summary, results from the NHSS feasibility study followup survey indicated that a high proportion of persons who initially refused to participate, when recontacted and offered an increased incentive, completed the risk behavior questionnaire. A lower proportion of persons who initially refused to participate provided both a blood sample and completed the risk questionnaire, when recontacted and offered an increased incentive. Persons at higher risk for HIV infection participated at higher levels in the followup survey than in the regular survey. The followup survey effectively increased the total number of persons who participated in the Dallas HIV survey. The increase in risk reporting among the sample of regular survey nonrespondents that were followed-up allowed for a significant reduction in nonresponse bias in the HIV estimate produced for Dallas County.

Table 6. Sample persons response rates in the Dallas County Household HIV Survey, 1989

Survey component	Regular survey	Regular + followup survey	Overall*
Screening	97	98	98
Blood & Questionnaire	80	84	82
Questionnaire only	80	90	88

*Product of screening and sample person rate.

Remuneration in the NHSS feasibility study demonstrated the power of incentives even in the most sensitive topic and invasive data collection situations, however establishing the exact mechanics of the reasons for success would require further study.

Table 7. Prevalence of selected HIV risk behaviors in the regular and followup survey, Dallas County male population, 18-54 years, 1989

Risk Behavior Since 1978	Regular Survey	Followup Survey	Total Estimate
Intravenous drug use	3	12	4
Receptive anal intercourse	3	11	5
5+ male partners	2	10	3
1+ male partner	5	17	8

THE NATIONAL SURVEY OF FAMILY GROWTH PRETEST

The National Survey of Family Growth (NSFG) is done periodically by NCHS to collect national data on the factors that affect the U.S. birth rate and women's reproductive health--factors that include sexual activity, marriage and divorce, contraception, sterilization, infertility, miscarriage, and abortion. Previous cycles of the NSFG have interviewed about 8,000 women 15-44 years of age in the noninstitutional population of the United States with response rates ranging from 75 percent to 80 percent.

Interviewing for the next NSFG, called Cycle 5, will be conducted in January-July of 1995. Three of the principal challenges for Cycle 5 of the NSFG will be (1) increasing response rates to make it possible to conduct a telephone reinterview in 1997 with as many of the original respondents as possible, (2) improving reporting of HIV-related sexual behavior, and (3) improving the reporting of abortion.

Response rates.--Most recently, in 1988, the NSFG used a list sample of households interviewed in the NHIS. Using a list sample saves nearly a million dollars on sample design and selection costs, but it makes it necessary to find women who move between the NHIS interview and the NSFG interview. Some are never found, so response rates are reduced somewhat. Response rates have been between 75 and 80 percent in recent cycles, despite the intrinsic appeal of the subject matter, the use of only female interviewers, thorough interviewer training, advance letters

introducing the survey, and expensive, intensive nonresponse follow-ups[17].

Although the NCHS obtains data from the National Health Interview Survey (NHIS) on HIV-related knowledge and attitudes, the NSFG remains the NCHS's principal vehicle for collecting data on HIV-related behavior, including such sensitive topics as age at first intercourse, numbers and characteristics of sexual partners, and condom use.

Abortion reporting is critical in the NSFG Cycle 5 because 25% of all pregnancies, and half of all unintended pregnancies, end in abortion. Fertility surveys in the U.S. and other nations have obtained incomplete reporting of abortion. In the last 3 cycles of the NSFG, in 1976, 1982 and 1988, and in most other U.S. surveys, women reported less than half of the abortions they have actually had[18]. This incomplete reporting of abortion has several potential adverse effects: it makes impossible analyses of the determinants and consequences of abortion itself; it forces us to use ad hoc methods to produce estimates of pregnancy rates for the U.S.; it produces biased estimates of the failure rates of contraceptive methods[19]; and it forces us to study unintended births instead of unintended pregnancies.

The NSFG Pretest for Cycle 5, conducted in October-December 1993, was based on about 800 eligible women, of whom 500 completed interviews. The Pretest was an experiment, which was designed to test several alternative contexts for asking questions. The pretest had 3 main groups:

1) the first group was a standard Computer-Assisted Personal Interview in the respondent's home with no incentive.
2) in the second group, the interview was moved to a neutral site--a site outside the home--where spouses, children, or parents could not hear the respondent's answers. To reimburse women for the time and inconvenience of going to the neutral site, respondents were paid $40 in cash at the end of the interview.
3) when we considered these first two groups, we were concerned that we might

---

[17] Judkins, David P, W. Mosher, and S. Botman. (1991). National Survey of Family Growth: Design, Estimation, and Inference. Vital and Health Statistics, Series 2, No. 109.

[18] Jones, Elise and J. Forrest. (1992). Under-reporting of Abortion in Surveys of U.S. Women: 1976-1988. Demography 29 (1): 113-126.

[19] Jones, Elise and J. Forrest. (1992). Contraceptive Failure Rates Based on the 1988 NSFG. Family Planning Perspectives 24 (1): 12-19.

obtain higher response rates and data quality in the neutral site/$40 group than in the in-home no-incentive group, but we would not know whether the differences were due to the non-home site or to the $40 payment. We were also concerned that the costs of implementing a $40 payment and setting up non-home sites on a national scale might be prohibitive. Therefore, we added a third group--a $20 incentive for an interview in the home.

For half the respondents in group 1 and group 3, (the in-home interviews), we also tested a short questionnaire at the end of the interview, using Audio CASI (Computer Assisted Self-Interviewing with headphones, and the respondent entering her answers into the computer).

NSFG Results

Pretest response rates (as a percent of those located) were higher for incentive cases than for non-incentive cases: 81 percent for those who received a $20 incentive vs 73% for those who received no incentive. The percent who broke an appointment with the interviewer was one-third lower for those who received $20 than for those who received no incentive (24 vs 37%).

The number of hours that the average interviewer worked to get a completed case was about 2 hours less for incentive than for non-incentive cases (8.8 vs 10.9; note also that when the incentive increases to $40, hours per case drops more than 2 full hours - see Table 8). Since the time of interviewers costs more than $10 an hour for their wages plus benefits, if the interviewer can save 2 hours of effort per case by paying a $20 incentive, then the incentive pays for itself. That is precisely what happened in the NSFG Pretest: the incentives paid for themselves in the $20 group because respondents broke fewer appointments for interviews and made themselves available after fewer telephone calls and personal visits. Costs in the $40 non-home group were higher because of high costs to set up the neutral sites--obtaining permissions, renting office space, etc.

Reporting of the number of sexual partners was higher among respondents who received incentives, still lower than reports of comparable studies of men[20, 21]. In

[20] Billy, J.O.G., Tanfer, K., Grady, W.R., and Klepinger, D. (1993). The Sexual Behavior of Men in the United States. Family Planning Perspectives 25 (2): p. 52-60.

previous cycles of the NSFG, less than half of abortions were reported[22]. Thus, abortion reporting should be at least double what it was in the last Cycle. In the pretest, a $20 incentive and the use of Audio CASI doubled abortion reporting, compared to a no-incentive, no Audio CASI group. This more complete reporting of abortions is probably due to two factors: (1) women using the Audio CASI (headphone) questionnaire reported a higher percentage of their abortions, and (2) incentives produced better coverage of groups of women who have higher abortion rates--including black women and poor women.

Table 8. Response Rates, Broken Appointment Rates, and Interviewer Hours per Case from the NSFG Pretest for Cycle 5

Experimental Group	Response Rate (%)	Percent with Broken Appointment	Hours per Case
In-Home, No Incentive	73	37	10.9
In Home, $20	81	24	8.8
Non-Home, $40	80	31	6.4

Recommendations from the Pretest

The results from the NSFG pretest suggest that a $20 incentive plus Audio CASI (self-administered questionnaires over headphones) should be used in the NSFG main study. The incentive will increase response rates, particularly among minorities and low-income women, and reduce the cost of interviewer labor because respondents will cooperate more readily.

Can these results be generalized? The results on response rates, interviewer hours and costs, in the NSFG Pretest are quite similar to those in the National Adult Literacy

[21] Smith, T.W. (1991). Adult Sexual Behavior in 1989: Number of partners, frequency of intercourse and Risk of AIDS. Family Planning Perspectives 23 (3): p. 102-107.

[22] Op cit. Jones and Forrest, Demography, Feb. 1992.

Survey (NALS) Field Test, a survey with a much larger sample (n=2,000) than the NSFG Pretest. Like the NSFG, the NALS required considerable effort from respondents[23].

The NALS and NSFG Pretest results provide evidence that incentives may be most cost-effective when the interview is:

(1) either long or a great deal of effort, or both; (2) sensitive either because it deals with private behaviors or may otherwise cause embarrassment (the NALS might cause such embarrassment among the adult illiterate, the NSFG because it includes questions on abortion and sexual behavior); and (3) part of a panel survey in which the response rate is critical to maintain the size of the panel over time.

All three of those conditions were common to both the NALS and the NSFG. The NSFG experience clearly demonstrates the success of incentives with hard-to-interview populations and sensitive topics. In addition, it clearly demonstrates the cost-effectiveness of the methodology in improving survey quality.

NATIONAL MEDICAL EXPENDITURE SURVEYS

The National Medical Expenditure Surveys (NMES) are designed to produce estimates of medical use, medical expenditures, sources of payment for medical care, and health insurance coverage. The surveys are sponsored by the Agency for Health Care Policy and Research (AHCPR). The household survey (HS) component of the NMES series yields estimates for persons in the civilian non-institutionalized population, while the survey of the institutionalized population (IPC) yields estimates for persons in nursing homes. Incentives have not been a design feature of the IPC surveys; therefore, the remarks that follow will focus on the use of incentives in the NMES HS surveys.

Respondent Incentives

The NMES series of household surveys includes the 1977 National Medical Care Expenditure Survey (NMCES), the 1987 National Medical Expenditure Survey (NMES2), and the 1996 National Medical Expenditure Survey (NMES3) presently in

---

[23] Berlin, M, L. Mohadjer, J. Waksberg, et al. (1993). An Experiment in Monetary Incentives, in American Statistical Association (editor), 1992 Proceedings of the Section on Survey Research Methods, pages 393-398.

the planning phase. These NMES studies share the following design features:

- the use of an initial screening interview to identify the households to be sampled for the study;
- the oversampling of poor people, blacks, Hispanics, and the elderly;
- a longitudinal or panel design in which sampled families are interviewed several times over a period of 14-16 months, in rounds of data collection that cover a year-long observation period;
- face-to-face interviews lasting on average between 2 and 2.5 hours in each round to be completed with a family respondent that provides information about him/herself and all other family members;
- the request that respondents prepare for the interviews by keeping a study calendar and saving records such as bills and insurance statements in order to improve the accuracy of their reporting of medical use and expenditures;
- special requests for information in addition to the interview itself, such as completion of self-administered forms;
- requests for signed permission forms from specific sample persons authorizing the collection of data from medical providers, employers and other health insurance providers to supplement and validate the data obtained from households.

In the context of the core NMES design summarized above, respondent incentives have been used primarily for the following reasons:

1) to motivate respondents to participate initially and in future interviews in order to minimize initial nonresponse and panel attrition.

2) to compensate respondents fairly for the burden associated with long interviews and the completion of additional survey tasks; and

3) to motivate respondents to keep records and provide fair compensation for the effort required to maintain the study calendar and save financial records over a long period of time.

The respondent incentives most frequently used in NMES surveys consist of cash

payments made by the interviewer at the end of each interview, typically in the form of a check. Checks provide a safe and convenient mode for interviewers to handle cash payments. In NMCES respondents were paid $5 at the end of each interview, beginning with Round 1. The amount was increased to $10 per interview in NMES2, and to a proposed $15 per completed interview in NMES3.

NMES respondents are required to sign a receipt acknowledging that money was received. In NMES2, the receipt was used as a vehicle to consolidate commitment to the role of respondent in the survey. The receipt the respondent was asked to sign included a statement indicating the willingness to accept responsibility for record-keeping in preparation for the next interview.

The use of incentives has no doubt contributed to the high response rates achieved in NMES studies, in spite of the burden that long interviews represent, and notwithstanding the oversampling of groups that in many surveys yield lower than average response rates, such as poor people and minorities. The overall response rate for the NMES2 household sample was 80.1 percent after four rounds of data collection.

Incentives are an important tool used by interviewers to convince reluctant respondents to participate in NMES surveys. Recent methodological research examined the characteristics of persons who had initially refused to be interviewed in any one of the NMES2 rounds of data collection[24]. The analysis revealed that reluctant respondents differed significantly from their cooperative counterparts with regard to the proportion of overall medical expenditures that different sources of payment covered for each group. The reduction of significant differentials with respect to health expenditures and insurance coverage, two core analytic concerns of NMES surveys, provide evidence of the beneficial use of incentives to guard off against potential nonresponse bias in national estimates.

In 1985, to aid in planning for NMES2, a feasibility study[25] was conducted to investigate a broad range of methodological issues. Among the issues examined was

---

[24] Cohen, S.B. and B.L. Carlson (1992). "An Analysis of the Characteristics of Reluctant Respondents in the National Medical Expenditure Survey". Proceedings of the Section on Social Statistics, American Statistical Association, in press.

[25] Mathiowetz, N.A. and Ward, E.P. (1987). Linking the National Medical Expenditure Survey with the national Health Interview Survey: An Analysis of Field Trials, Vital and Health Statistics, Series 2, No. 102, U.S. Government Printing Office. Washington, D.C.

the effect of several incentive protocols[26]. The feasibility study consisted of two rounds of data collection: 1) a personal interview 45-60 minutes long where the respondent was paid five dollars at the end of the interview, and 2) a second interview in person or by telephone where again the respondent was paid five dollars at the completion of the interview. Approximately two weeks prior to the second interview, self-administered questionnaires were mailed to all respondents. These questionnaires were designed to take approximately 30 minutes to complete and included some moderately threatening questions on health behavior and mental health status.

Reporting groups were divided into three treatment groups: 1) "Prepayment" - persons were sent a five dollar check with the questionnaire; 2) "Promised" - persons were told that they would be paid five dollars when the completed forms were returned; and 3) "No mention" - persons were not given any information on payment, but were paid five dollars upon questionnaire completion.

The results indicated that the prepaid incentive leads to a significant improvement in response rates. Seventy-three percent of those in the prepaid group completed the survey compared to 66% among those who were not told of the incentive. The prepaid incentive also worked better than the promised incentive, which resulted in a response rate of 60%. Item nonresponse rates were calculated for each completed self-administered questionnaire and were used as a general measure of data quality. The finding was that prepayment leads to lower item nonresponse. Ninety percent of those who were prepaid answered all of the questions in the 18 page questionnaire, compared to only 74% in the promised group and 87% in the no mention group.

The study concluded that prepaid incentives can result in higher response rates and more complete data with less need for follow-ups. These were achieved in this study at a very moderate increase in cost. However, the net added costs may be far less than the value of the incentive payments, since a substantial part of the incentive costs is offset by savings in the follow-up activities. This finding was in line with results from other mail surveys, and that mode of payment was adopted in NMES2 when self-administered forms were mailed to sample households.

Based on results from the most recent NMES feasibility study carried out in 1992, the schedule of cash incentives in NMES3 will be modified relative to earlier surveys. Instead of introducing record-keeping tasks and paying respondents for the first time at

---

[26] Berk, M.L., Mathiowetz, N.A., Ward, E.P. and White, A.A. (1987). The Effect of Prepaid and Promised Incentives: Results of a Controlled Experiment. Journal of Official Statistics, Vol. 3, No. 4, pp. 449-457.

the end of Round 1, the 1992 Feasibility Study presented the family respondent with the study calendar at the end of the Screener round and prepaid these respondents in anticipation of the time and effort that would be devoted to keeping records in order to prepare for the Round 1 interview. Payments for later rounds of the Feasibility study were also described as compensation for future effort. In the last round a token gift (a commemorative tile with the U.S. Public Health Service seal) was given to respondents instead of cash payment. Under the new plan, the gift is the only net increase in incentive costs compared to previous plans, and it appears that prepayment has advantages. The rates of NMES calendar use in Round 1 of the Feasibility Study are significantly higher than the rates achieved in NMES2 in the round immediately following the round when payment and instructions to keep records were first delivered[27].

The 1992 NMES Feasibility Study also tested successfully the use of incentives to motivate respondents to complete a complex data collection task that was time dependent. This involved the procurement of health insurance printed materials that included a description of the benefits associated with the health plan offered by the employer to each policyholder in the household[28]. In NMES2 the collection of comparable information was attempted from employers in the course of the health insurance provider survey, but the time lag between the end of the household survey and the start of that provider survey frequently made it impossible to locate the necessary information about the health plan in effect at the time the household was interviewed.

In the Feasibility Study, respondents were offered $15 per household to contact employers, either by mail or in person, and secure the necessary information. Interviewers gave respondents a request form that could be presented to the employer to facilitate the task. Payment was made when the health insurance materials were delivered in the next round, and the amount was not increased in the event that the family had more than one eligible policyholder. Health insurance booklet requests in the Feasibility Study were followed up in later rounds and, by the end of the study, a policy booklet had been retrieved for 75 percent of eligible plans, at a lower cost and

---

[27] Sanchez, M. E. (1993). "Enhancing Compliance with Record-keeping in a Household Survey". 1993 Proceedings of the Section on Survey Research Methods of the American Statistical Association, Vol. 2, p 1015-1020.

[28] Emmons, C.A., Curno, M., and Smith, K. (1993). Final Report on the Outcomes of the Procedure for Obtaining Health Insurance Policy Documents from Respondents in the NMES3 Household Survey Feasibility Study. Submitted by NORC and Westat, Inc. under contract requirements.

on a more timely basis than in NMES2.

The use of remuneration to improve survey scheduling, achieve self-response and record keeping, and establish a commitment for a longitudinal data collection process are shown in the NMES experience. The NMES experience also demonstrates the potential for improved data quality with remuneration.

Interviewer Incentives

The morale and motivation of interviewers and supervisors are important factors in response rate outcomes for sample surveys. While the impact of respondent incentives has been frequently discussed in the survey research literature, there is little evidence of systematic inquiry into the use of incentives for interviewers as a means of achieving high response rates.

Typically, studies resort to interviewer incentives in a haphazard and improvised fashion when production levels have tumbled and the study response rate is deemed unacceptably low. This plan of action may not be as desirable or as economical as the notion of setting up a planned and carefully crafted incentive plan for interviewers from the very start of the project. The experience in the 1992 NMES Feasibility Study with such a plan suggests the desirability of exploring further the manner in which interviewer incentives may be manipulated to achieve gains in field response rates and efficiency within acceptable budget limits.

The Feasibility Study included a plan for interviewer incentives in order to achieve high response rates within the schedule for data collection. The project staff and the contractor collaborated to come up with a plan that was acceptable to all. The field staff wanted a plan that would promote and reward team work as opposed to individual performance; thus, the team was defined as the cluster of interviewers working in each PSU.

Realistic response rate levels for two points in time during the round (a specified midpoint in the field period, and the end of the round) were defined for each PSU and communicated to interviewers at the beginning of each round. The incentives were cash payments (about $25 for each of the time points in a round) paid equally to all PSU interviewers provided the PSU had achieved the targeted response rate by the specified date. A very modest additional amount of money was paid for increases in the response rate beyond the specified minimal rate.

The bonus plan encouraged interviewers to talk to their fellow PSU interviewers to

coordinate the steady flow of work. A weekly memo informed interviewers of the progress in other areas and the ranking of their own PSU. With very few exceptions, the staff in each of the PSUs achieved the response rate goals consistently for a very modest investment.

The benefits of the incentives included: teamwork among the staff at the local PSU level; steady production which avoided the last minute dislocations and expenses typically associated with a late push to increase response rates; and availability of qualified staff towards the end of the study to handle difficult assignments in a planned fashion.

More systematic research on the use of interviewer bonuses and the performance of different bonus plans is desirable, as the strategy appears a cost-effective way of obtaining high response rates.

Remuneration can be a valuable tool for direct interviewer management as well as helping interviewers motivate respondents to participate.

Physicians are often surveyed to obtain a wide variety of medical information. The high frequency at which physicians are surveyed coupled with the problem of "gate keepers" probably contribute to the low response rates typically achieved in physician surveys. However, prepaid incentives seem especially effective for this population group[29, 30, 31, 32]. *The following three experiences serve as examples.*

---

[29] Berry, S. H. and Kanouse, D.E. (1987). Physicians response to a mailed survey: An experiment in timing of payment. Public Opinion Quarterly, 51, p. 102-114.

[30] Lockhart, D.C. (1991). Mailed surveys to physicians: The effect of incentives and length on the return rate. Journal of Pharmaceutical Marketing & Management, 6(1), p. 107-121.

[31] Mizes, J.S., Fleece, E.L., and Roos, C. (1984). Incentives for increasing return rates: Magnitude levels, response bias, and format. Public Opinion Quarterly, 48, p. 794-800.

[32] Berk, M.L., Edwards, S.E. and Gay, N.L. (1993). The use of a prepaid incentive to convert nonresponders on a survey of physicians. Evaluation & The Health Professions, Vol. 16, No. 2, pp.239-245.

NATIONAL SURVEY OF DIAGNOSTIC ALLERGY TESTING

In 1988 a remuneration experiment was conducted[33] on a subset of physicians selected to participate in the National Survey of Diagnostic Allergy Testing (sponsored by the Health Industry Manufacturers Association).

A sample of physicians was randomly divided into three experimental groups. The first group received a $10 incentive with the first mailing. Nonresponders to the initial mailing were sent a new questionnaire as well as a letter urging them to respond and mentioning the $10 incentive they had received earlier. The second group of physicians did not receive a monetary incentive with the initial mailing. On the second mailing of the questionnaire, however, they received another letter explaining the importance of the study as well as a $10 prepaid monetary incentive. No mention of payment was made to the third group on either the first or second mailing.

The results indicate the use of a prepaid incentive has a dramatic impact on the response to the initial mailing. Fifty-five percent of those physicians receiving a prepaid incentive responded to the initial mailing, compared to less than 20% who were not told about payment on the initial mailing. Overall, a 63% response rate was obtained for Group 1 physicians (prepaid incentive with the initial mailing), compared with only a 50% response rate for Group 2 (prepaid incentive on the first prompt), and a 40% rate for Group 3 (no incentive).

This study concluded that incentives should be used in cases in which its use is considered necessary to obtain adequate response rates. Also, while delaying the decision to use an incentive until the second wave of mailing enables the researcher to decide whether an adequate response rate is likely to be obtained without payment, the incentive is not nearly as effective when used in a follow-up mailing. Because obtaining high response rates on physician surveys is difficult, few researchers will be able to conclude at study onset that a high response rate can be obtained. The use of a prepaid monetary incentive enclosed with the initial questionnaire mailing, therefore, appears to be a cost-effective method for improving response rates on physician surveys.

NATIONAL HOSPITAL DISCHARGE SURVEY

Most of this paper has dealt with remuneration for individual respondents, but

---

[33] Op cit. Berk et al., Evaluation & the Health Professions, 1993

institutions and organizations are sometimes respondents for surveys, and the National Hospital Discharge Survey (NHDS) experience suggests that remuneration is essential in some institutional surveys.

The National Hospital Discharge Survey (NHDS), conducted by NCHS, was first fielded in 1964 following the completion of a feasibility study[34]. The NHDS is a continuing study designed to provide comprehensive general-purpose statistics on morbidity in patients discharged from the Nation's general and short-stay hospitals. The principal source of information for the survey is the medical record in the hospital. The data are obtained from probability samples of medical records abstracted in a sample of general hospitals. Hospitals are compensated for participation in the survey.

Some hospitals in the sample have automated records, and contract with an abstract service. Other hospitals abstract records manually as needed. There are two manual data collection procedures: a primary manual procedure in which hospital staff complete the abstracts, and an alternate procedure in which a Census Bureau representative completes the abstracts. Hospitals using the primary manual procedure receive an average of $2.40 per abstract submitted; those using the alternate manual procedure receive about $1.00 per abstract submitted. Data tapes of uniform abstracts covering all discharges for automated hospitals are purchased directly from abstract service organizations. The cost of these data ranges from $.003 to $.055 per discharged patient. These discharges are sampled for the survey. All hospitals participating in the NHDS are reimbursed $1.00 per record biannually for approximately 40 records that are reabstracted for quality control procedures.

A large part of the success of the NHDS depends on the willingness of the hospitals to perform substantial continuing activities. Once inducted into the survey, hospitals participate for an extended period of years. A substantial amount of work is involved, including sampling the discharge lists, pulling and refiling medical records, and abstracting approximately 20 records monthly. A feasibility study conducted in 1963-1964 found that most hospitals expected compensation for their effort.

Cost analysis supports the practice of remuneration, particularly for primary manual procedure hospitals. For example in Fiscal Year 1992, it cost the NHDS an additional average of $12.15 per record to have the Census Bureau sample and abstract the data

---

[34] Brown, A.M., Altman, I. and Thompson, D.J. (1966), Participation of Hospitals in the Pilot Study of the Hospital Discharge Survey, Vital and Health Statistics, Series 2, No. 19, U.S. Government Printing Office. Washington, D.C.

in the alternate procedure hospitals. If the Bureau of the Census performed this work in all the sample hospitals, the cost of the NHDS would be substantially increased.

The feasibility study found that "While some hospitals indicated they might be willing to collaborate in the survey without reimbursement, it was clear that most would expect some compensation for their contributions, especially where it was felt that overtime work or the employment of additional personnel might be required. The ... form shown to the [hospital] administrator ... was quite detailed and implied a fairly exhaustive review of the medical record. It contained questions on the characteristics of the patient and his *{sic}* hospitalization, including final diagnoses, operations, complications, laboratory tests, therapies, and the like. There were some differences about whether payment should be made to the hospital or to the personnel doing the work, but most administrators favored payment to the hospital."

The feasibility study recommended that "... a uniform policy be adopted for the compensation of hospitals and that fair payment, based on further examination of the true cost to the hospitals be made." A subsequent pilot study was used to confirm the acceptability of the survey procedures , including remuneration, and helped to calibrate the payment amounts.

The NHDS experience illustrates the testing and use of remuneration and their use to offset a financial disincentive, improve the acceptance of a government survey, provide evidence of the value of participation, and keep expenses down in a survey requiring the continuing participation of institutions.

Summary and Conclusions

The evidence summarized in this paper shows that remuneration for respondents can be an effective technique for raising response rates and data quality when otherwise good survey practices are not sufficient.

More and more in the last decade, policy makers and program administrators are demanding data that cannot be supplied with a standard survey -- the one-hour cross-sectional interview containing nonsensitive questions is no longer the norm. The health surveys reviewed here each have one or more features that do not fit the public's perception of what a standard survey is: some require long interviews (NSFG Pretest and NMES), others require the maintenance of records such as diaries (NMES) or event histories (NSFG); some are panel surveys with repeated interviews (NSFG, NHANES Follow-up, NMES); some have sensitive questions (NSFG Pretest, NHSS, and YRBS); others use non-home sites (NHANES and the NSFG Pretest); or ask for

medical tests (NHANES); or require information or testing that could be embarrassing (NALS, NHANES, and NHSS).

As interviews get longer and questionnaire content gets more difficult or intrusive, and hard to interview sub-populations are surveyed, the need to motivate respondent participation grows. We need to provide respondents with concrete evidence of our appreciation and the importance of their participation and willingness to provide accurate and complete information. The careful use of remuneration allows us to offer people more than the promise that policy makers including Congress will use the data to improve their lives. If an advance letter or first personal contact explains that they will be compensated, all of the practical evidence reviewed here suggests that completing a quality interview at a reasonable cost is more likely.

Although most of this paper has dealt with remuneration for individual respondents in households, it is important to note that institutions, medical professionals, and other organizations are frequently respondents for health surveys. The NHDS experience suggests that remuneration is important in some institutional surveys also.

The history of remuneration in health surveys as evidenced by the experience reviewed here is a successful one. Remuneration has stood the test of time, proving successful in controlled experiments, field trials, and long-term implementation. For relatively little cost, important improvements in response and data quality have been gained using remuneration methods.

THE USE OF INCENTIVES TO SURVEY "HARD-TO-REACH" RESPONDENTS: A BRIEF REVIEW OF EMPIRICAL RESEARCH AND CURRENT RESEARCH PRACTICE

Richard A. Kulka
National Opinion Research Center
University of Chicago

1. Introduction

Incentive payments to survey respondents have been used extensively for many years as a means of improving survey response rates (cf. Shuttleworth, 1931), and there is considerable research evidence supporting the value of compensation for increasing cooperation and improving the speed and quality of response in a broad range of data collection efforts (cf. Kulka, 1992; Willimack, Petrella, Beebe, and Welk, 1992). In particular, a large number of empirical studies concerned with increasing response to mail questionnaires consistently attests to the effectiveness of monetary incentives in increasing mail survey response rates (e.g., Armstrong, 1975; Church, 1993; Duncan, 1979; Fox, Crask, and Kim, 1988; Harvey, 1987; Heberlein and Baumgartner, 1978; Hopkins and Gullickson, 1992; Kanuck and Berenson, 1975; Linsky, 1975; Yammarino, Skinner, and Childers, 1991; Yu and Cooper, 1983).

Nevertheless, until fairly recently monetary incentives and other forms of respondent remuneration have not been used extensively in general survey practice, especially in studies under government sponsorship and large scale academic surveys--as opposed to their widespread and common use in commercial or market research. In recent years, however, it has become increasingly difficult to achieve response rates high enough to provide statistically valid results, and remuneration has become more common. In the United States, Federal statistical surveys cannot employ incentives without explicit authorization from the Office of Management and Budget (OMB), and OMB has generally prohibited the use of payments to respondents, except under circumstances where "substantial need" can be demonstrated. And, interpreting this rule on a case-by-case basis--as requests for the use of incentives from Federal agencies have become increasingly common--has made it more difficult for OMB to assure consistency in the application of these guidelines.

To assist OMB in developing appropriate principles and decision rules governing the use of respondent incentives by Federal agencies, in October, 1992, the Council of Professional Associations on Federal Statistics (COPAFS) convened a symposium of representatives of government, business, academic, and research organizations to consider the current state of experience, research, knowledge, and opinion regarding the use of such incentives (COPAFS, 1993). Although it was not expected that the symposium would provide definitive answers to the multitude of

questions surrounding the use of incentives, "OMB expected to obtain information that would help in preparing guidelines to foster greater consistency in reviewing future requests by Federal agencies to use incentives when conducting surveys" (p. 1).

Toward that goal, symposium participants discussed, in part, "the kinds of survey situations in which incentives have a high probability of being effective or necessary" (p. 8), articulating "a set of circumstances in which they thought OMB should seriously consider an agency's request to use incentives" (p. 9). Among these were a number of situations or circumstances that might be broadly conceived as under the general rubric of "hard-to-reach respondents." Specifically, the list included the use of incentives:

- To encourage *hard core refusals* to respond, especially in small subpopulations of interest where response rates are often quite low--low enough to raise serious questions about the quality of survey data for these subpopulations.

- When there is a significant likelihood that a *"gatekeeper"* will prevent the respondent from ever receiving the questionnaire or otherwise make it difficult to make contact with certain segments of the population to conduct an interview.

- When there is a special target population for whom our conventional means of motivation or encouragement will have little if any chance of working--i.e, where the positive forces to cooperate are quite low (e.g., prostitutes, the homeless, the disenfranchised).

- If the target population is a small group that is *often surveyed*, such that a particular respondent is likely to be sampled frequently for one survey or another (e.g., physicians, CEO's, university deans).

- If the population is a *control group* in a program evaluation or experiment in which it is imperative to achieve and maintain an adequate response rate in the control group sample if the integrity of the study is to be maintained.

More generally, one of five potential OMB incentive policies suggested for consideration in nonstandard survey situations was that incentives "be considered if the respondent incurred out-of-pocket cost; or if the survey was too intrusive; or the survey was aimed at a *hard-to-reach population* [emphasis added]" (COPAFS, 1993, p. 12). Overall, most of those present felt that:

hard-to-reach really meant *hard-to-interview* [emphasis added]. This category could include those who are hard to encourage to cooperate, and therefore initially refuse. In such cases, incentives might be effective. [However], participants felt that incentives would not be effective for those who are hard to find.

Participants also included in the hard-to-interview category those who are difficult to reach by mail, those who must be kept in a sample (such as members of a control group), and those disenfranchised from society. (p. 12).

Based on this broad conception, the focus of this paper is explicitly on the use of incentives to survey hard-to-reach respondents, in contrast, for example, to the use of respondent incentives as a reimbursement for out-of-pocket expenses, as a payment to respondents for their time and effort in participating in a survey, or to compensate respondents for carrying out survey tasks that entail unusual demands, i.e., those which are especially burdensome or intrusive or may put the respondent at risk. In principle, to the extent that "hard-to-reach" is viewed as synonymous with "hard-to-survey," this focus is quite broad, in that encouraging those who might otherwise be reluctant to cooperate with a given survey to indeed do so is, in essence, the basic intent of the vast majority of surveys that choose to provide remuneration to survey respondents as part of their design.

However, incentives provided to stimulate survey response are rarely given only to initial or hard-core refusals, and several important questions regarding the use of incentives to encourage response may be addressed by focusing on this particular use of remuneration (cf. COPAFS, 1993). These include:

- Are there indeed specific target populations who are routinely offered remuneration to participate in surveys by most or all survey organizations because they are regarded as especially difficult to survey?

- While respondent incentives may increase cooperation among initial refusals, are they really effective with hard-core refusals and the truly difficult or impossible to interview populations?

- Are incentives effective only for certain target populations or subpopulations or more effective for certain population subgroups than for others (i.e., are the effects of incentives different for different population subgroups)?

- Are incentives really effective in getting past "gatekeepers," either for certain professionals (e.g.,

physicians) or other difficult-to-survey subgroups of the general population?

- Are incentives indeed of little use in locating and interviewing hard-to-find cases?

- Should consideration be given to paying some, but not all respondents to a given survey?

- Should all respondents be paid the same incentive, or should consideration be given to different levels or types of remuneration for different respondents?

To seek possible answers to these and some other related questions, we conducted a focused review of the current research literature on the use of incentives, with particular attention to their use with hard-to-reach populations, broadly defined.

2. Current Research Practice

However, because we anticipated that empirical evidence bearing on many of these questions would likely be quite sparse, we also sought input from individuals at most of the government, business, academic, and research organizations represented at the COPAFS symposium in October 1992, along with a few others, to ascertain the current state of survey research practice with regard to the use of incentives to survey hard-to-reach populations. In addition to providing citations or references to any papers or publications related to this topic, each organization was asked to provide information on any recent surveys they had conducted with such populations, indicating when they had or had not used incentives, and a sense of their general organizational policies or conventions regarding the use of incentives under such circumstances. Prior to describing some of the evidence available from the research literature that bears on the questions raised above, it will be useful to summarize current practice in this area as described by these organizations.

Not surprising, the vast majority of these organizations routinely conduct surveys with hard-to-reach respondents under our broad definition. As noted by one organizational respondent, virtually every survey encounters and must deal with hard core refusals, but more specific categories of respondents designated by responding organizations as hard-to-reach are:

(1) the economically disadvantaged (e.g., lower income or lower socioeconomic status (SES) populations, welfare recipients or applicants, the homeless);

(2) the educationally disadvantaged (e.g., the less educated, high school dropouts, those with low literacy levels, the mentally retarded);

(3) minority populations (e.g., African Americans, Hispanics, disadvantaged minorities, impoverished urban minorities);

(4) adolescents, youth, and young adults (e.g., youth in general, minority youth, young black males, teen mothers, the young and mobile);

(5) drug users and those with special health problems (e.g., current or former drug users, drug abusers, cocaine users, diabetics, those with asthma);

(6) frequently surveyed professional or elite populations (e.g., physicians, nurses, CEO's, teachers, college and university faculty, both very small and very large farm operations); and

(7) transients and persons who wish not to be found for legal or other reasons (e.g., highly mobile and transient populations, runaway youth, absent parents owing child support, those defaulting on student loans).

Not only do these categories overlap a great deal, but also, in almost every case, incentives have been used in surveys of these populations to increase response rates by at least one organization, and quite often by many. For example, the use of (generally substantial) incentives in surveys of physicians is a standard practice in virtually all of these organizations. It is also the case, however, that recent surveys have been conducted with most of these subpopulations in which *no incentives* were used, including a few surveys of physicians.

Although most of the organizations queried feel that respondent incentives are generally effective in increasing response rates among these hard-to-reach target populations, very few controlled or randomized experiments have been conducted to demonstrate empirically the efficacy of incentives in improving response rates under such circumstances. Not surprisingly, the use of incentives in surveys conducted by or for Federal statistical agencies--which require OMB approval for providing respondent incentives--is somewhat more likely to be based on such empirical evidence than their use in surveys conducted by commercial, academic or private research firms under other auspices.

Even in the absence of such controlled experiments, several of these survey research professionals and firms believe that respondent incentives are an important overall tool in their arsenal for dealing with hard-to-interview survey populations, and

their experiences with such incentives bear on at least three of the basic issues raised earlier regarding the use of incentives with such populations. First, although many participants at the COPAFS (1993) symposium felt that incentives might not be effective in locating those who are hard-to-find, several of these organizations report experiences that suggest that paying respondents makes contacting and locating easier, less expensive and more effective, since contact individuals are more willing to convey messages and provide new address and telephone numbers for sample members when interviewers are able to mention that they have a monetary incentive for the respondent. Similarly, others cite experiences suggesting the efficacy of respondent incentives in "opening the door" or getting past "gatekeepers,"--i.e., in helping gain access to the respondent--because nurses, receptionists, relatives, friends and other "gatekeepers" are apparently more reluctant to restrict or deny access to a potential respondent when a monetary incentive is involved.

Third, although several of those responding to our inquiry expressed some ambivalence regarding this practice, a number of surveys have been conducted which provide incentives either to some but not all respondents, or different levels of remuneration to respondents in the same survey. In some cases, these different incentive levels (including no incentive) reflect different levels of burden, effort or risk for different respondents, but it is also not uncommon to offer incentives only to sample members for whom one is having difficulty getting them to respond--i.e., to do the survey without routine remuneration, but then use monetary incentives to try to convert refusals. Alternatively, in a survey providing incentives at a given level, interviewers might be permitted to offer increasingly larger amounts to convert increasingly hard-core refusals, to persuade extremely hard-to-convince cases to indeed cooperate.

Although these conditional incentive approaches can be quite cost effective, paying uncooperative sample members when cooperative respondents are not paid, or paying especially reluctant or difficult sample members more than those who cooperate more readily, violates our sense of fairness or equity. However, the selective or strategic use of remuneration to convert hard-core refusals and achieve higher or very high response rates is--though relatively rare and practiced with some reluctance--very much a part of current research practice with regard to the use of incentives to survey hard-to-reach respondents.

3. A Brief Review of the Literature

Having briefly summarized the current state of practice in this area, let us now explore what, if anything, the current research literature can tell us regarding the use of respondent incentives with difficult-to-survey populations or sample members.

3.1 The Use of Incentives in Surveys--A Summary

As a background for that analysis, it is important to consider momentarily what we know (or think that we know) about the use of respondent incentives in general (cf. Kulka, 1992, Willimack et al. 1992). As noted earlier, few today would question the general assertion that a monetary incentive enclosed with a mail questionnaire will serve to increase response rates. Hundreds of studies have been conducted, and review after review--both qualitative and quantitative--concludes that the importance of financial incentives is "second only" (perhaps) to the use of follow-up mailings or prompts in improving response rates (cf. Dillman, 1991). Moreover, the literature rather overwhelmingly supports the predominant effectiveness of prepaid as opposed to promised incentives (e.g., Armstrong, 1975; Berk, Mathiowetz, Ward and White, 1987; Berry and Kanouse, 1987; Blumberg, Fuller, and Hare, 1974; Church, 1993; Furse and Stewart, 1982; Hopkins and Gullickson, 1992; James and Bolstein, 1992; Kanuck and Berenson, 1975; Linsky, 1975; Peck and Dresch, 1981; Skinner, Ferrell, and Pride, 1984; Wotruba; 1966; Yu and Cooper, 1983). That is, incentives appear to be most effective in inducing survey response when they are paid in advance--at the time that the respondent's cooperation is initially solicited--rather than offered conditional on and paid subsequent to respondent cooperation, even when the promised or conditional incentive is greater than the amount prepaid (cf. Linsky, 1975; James and Bolstein, 1992).

The use of monetary incentives to increase response rates for telephone and personal interview surveys has received far less research attention, although the results of several studies are consistent with those derived from mail surveys. Overall, however, the conditions under which a monetary incentive will be effective or ineffective under these survey modes appear to be less general. Based on this more limited research literature, the greatest potential effectiveness of monetary incentives appears to be in surveys that place unusual demands upon the respondent, require continued cooperation over an extended period of time, or when the positive forces on respondents to cooperate are fairly low (cf. Cannell and Fowler, 1977).

In addition to a potential beneficial impact on response rates, the research literature suggests that incentives may have a beneficial impact on *data quality* as well. At least two theories suggest the opposite--a *detrimental* effect of remuneration on data quality:

(1) a concept based on "social desirability" theory (Cannell and Henson, 1974; Weiss, 1975) that suggests monetary inducements will increase the tendency of participants to try to please the interviewer by providing what the respondent believes is the desired or "correct" answer; and

(2) a "self-perception" model that argues that the introduction of financial incentives acts as an *external* motivator, thereby reducing the degree of internal motivation (i.e., the subject's interest or desire to participate), decreasing the degree or *quality* of compliance (i.e., quality of response), while increasing the *rate* of compliance or cooperation (cf. Hansen, 1980).

While relatively little empirical evidence has been found in support of either of these two models (see, however, Hansen, 1980; James and Bolstein, 1990), the preponderance of evidence reported to date (e.g., Berk et al., 1987; Cowan, 1977; Ferber and Sudman, 1974; Godwin, 1979; Goetz, Tyler, and Cook, 1984; Houston and Ford, 1976; James and Bolstein, 1990; Kerachsky and Mallar, 1981; McDaniel and Rao, 1980; Sudman and Ferber, 1974) is more consistent with a theory based on "social exchange," which posits that the offer of monetary incentives induces a greater commitment to the survey task among respondents, which in turn results in better data quality from survey respondents.

With regard to incentive size, the research literature is significantly less helpful, since the majority of studies have investigated the effects of incentives of $1 or less (e.g., Armstrong, 1975; Fox et al., 1988; James and Bolstein, 1992; Jobber and Saunders, 1988; Kanuck and Berenson, 1975; Linsky, 1975; Yammarino et al., 1991; Yu and Cooper, 1983), and few studies have successfully demonstrated the effectiveness of very large monetary incentives (e.g., Berry and Kanouse, 1987; Gunn and Rhodes, 1981; James and Bolstein, 1992). At both extremes, there is some evidence that increasing the size of monetary incentive will result in increases in survey response and/or quality (e.g, Armstrong, 1975; Findlay and Schaible, 1980; Fox et al., 1988; Furse and Stewart, 1982; Godwin, 1979; Gunn and Rhodes, 1981; Hubbard and Little, 1988a, 1988b; James and Bolstein, 1990, 1992), but there is also rather consistent evidence that this benefit may rather quickly reach "diminishing returns," whereby larger incentives no longer result in appreciable increases in survey response (e.g., Armstrong, 1974; Fox et al., 1988; Godwin, 1979; Hubbard and Little, 1988b; James and Bolstein, 1992; Mizes, Fleece, and Roos, 1984).

Overall, why payments are effective in improving response rates is not currently very well understood. Some believe that "payment works in increasing response rates . . . through motivating and supporting the *interviewer* [emphasis added] in his [or her] approach to the respondent' (Weinberg, 1973, p. 480), while others view "incentives . . . as impressing upon the survey *respondent* [emphasis added] the importance of the task" (Goetz et al. 1984:149; Berry and Kanouse, 1987). Still others have suggested that the predominant motivating power of an incentive is not its monetary value, but rather its symbolic, or "token" value (e.g., Linsky, 1975).

In turn, two basic classes of theories have been referenced to explain why incentives may increase survey participation (cf. Willimack et al., 1992). The first is based on the principle of "reciprocation" (Groves, Cialdini, and Couper, 1992):

> Every human society abides by a *norm of reciprocity* [emphasis added] that directs individuals to provide to others the general form of behavior that they have received from others (Gouldner, 1960). . . . [Based on] the reciprocity heuristic, . . . one should be more willing to comply with a request to the extent that the compliance constitutes the repayment of a perceived gift, favor, or concession. (p.480)

Thus, by providing an incentive as an unsolicited gift (e.g., a prepaid incentive), one invokes the norm of reciprocity among respondents, who can "reciprocate" by participating in the survey.

Closely related to the concept of a "norm of reciprocity" (and in the same basic class) are theories of "cognitive dissonance" and "social exchange." Under the former, it is postulated that the inclusion of a prepaid token incentive with a request for survey participation creates psychological dissonance, which is most easily resolved by consenting to in fact participate (cf. Furse and Stewart, 1982, 1984; Hackler and Bourgette, 1973).

As articulated by Dillman (1978), "social exchange" theory, which emphasizes the perceived costs and rewards of responding to a survey, suggests that, in order to maximize survey response, one must "minimize the costs for responding, maximize the rewards for doing so, and establish trust that those rewards will be delivered" (Dillman, 1978, p. 12). Rather than serving as a reward for survey participation, the use of an incentive serves as "a symbol of trust," a major factor necessary for social exchange to successfully occur. Consistent with this notion is research evidence showing that

> increasing the size of an incentive does not always increase response, and in fact may tend to decrease it, and that including it with the appeal [a prepaid incentive] is more effective than promising to send it on return of the questionnaire. The closer the monetary incentive comes to the value of the service performed, the more the transaction tends to move into the realm of economic exchange [in which money serves as a precise measure of the worth or value of one's actions] and the easier it becomes for people to refuse it. (Dillman, 1978, p. 16)

In general, smaller, prepaid incentives appear to invoke social exchange or the norm of reciprocity, while larger promised or conditional incentives are more likely to invoke *economic exchange*, which represents the second basic class of theories on how incentives serve to increase survey participation, i.e., by

literally paying respondents for the time and effort required to provide the information requested. Under this model, remuneration represents reimbursement for survey cooperation (cf. Cohen, Walden, and Ward, 1992), i.e., compensation to respondents for their time and effort in participating in the survey, rather than an "incentive," "gift," or "gesture of goodwill" too small to be considered equitable payment or exchange for respondents' time.

In this regard, Cannell and Henson (1974) emphasize that, since respondents rarely share the goals of a survey, they do not consider participation as a means of advancing their own, personal goals, and are thereby generally unmotivated to perform the survey task. An incentive, in the form of a payment, may serve to provide a personal goal which motivates participation in a survey. To be effective, the amount of money offered must be large enough to be worth working for--i.e., the value of the incentive should be commensurate with the task and/or time sacrificed by the respondent--but not extravagant, because, if individuals perceive that they are overcompensated, the effects on participation may actually be negative (Cannell and Henson, 1974; Groves, 1989).

Overall, the preponderance of research evidence appears to favor the concepts of social exchange or reciprocity as a basis for the effectiveness of incentives, although with the advent of increasingly more complex surveys and the use of larger incentives, a significant body of evidence consistent with the tenets of economic exchange is also accumulating, much of it in relation to surveys of apparently hard-to-reach or hard-to-interview populations, to which we now turn our full attention.

3.2 Hard-to-Reach Target Populations

Although at least seven different categories of respondents were identified by survey practitioners as "hard-to-reach," research on the effects of respondent incentives has focused directly on only a few of these--frequently surveyed professional and elite populations, adolescents and young adults, and the disadvantaged--and the relative coverage of even these three broad populations in that regard is quite uneven.

By "focusing directly" I mean that the subjects for experimentation with incentives are drawn largely or entirely from one of these target populations. A number of these studies include one or more of these subgroups as a component of the population surveyed, and such subgroups may or may not have responded differently to incentives than other groups represented in the sample, but the issue of the differential effectiveness of incentives among those hard-to-survey will be dealt with in a separate section.

Studies on the use of incentives with professional and elite populations--especially physicians--are legion. Physicians are

widely believed to be an especially difficult population from which to collect survey data (Sudman, 1985), since they "are frequently approached for surveys, the demands on their time are great, and their office staffs are vigilant in protecting them" (Berry and Kanouse, 1987, pp. 102-103). Physicians' reluctance to participate in surveys is a growing problem for researchers (cf. Berk, 1985), with the American Medical Association (AMA) warning that "physicians are becoming weary and wary of surveys" (Martin, 1984), while the response rates to even their own surveys have declined precipitously (Goodman and Jensen, 1981).

Both prepaid (Berk, Edwards, and Gay, 1993; Berry and Kanouse, 1987; Lockhart, 1991; Mizes, Fleece, and Roos, 1984) and promised (Gunn and Rhodes, 1981; Weber, Wycoff, and Adamson, 1982; Tambor, Chase, Faden, Geller, Hofman, and Holtzman, 1993) monetary incentives have been shown to significantly improve response rates in surveys of physicians, whether conducted by mail, telephone or in person. For example, Gunn and Rhodes (1981) conducted an experiment to determine the effectiveness of paying monetary incentives to physicians for their participation in a 20-30 minute telephone interview on attitudes toward influenza immunization. Physicians were systematically assigned to one of three subsamples designated to receive no incentive, $25, or $50. Study findings revealed a 58 percent response rate for the group offered no incentive, 69 percent for those promised $25, and 77 percent for those offered $50. In a personal interview survey conducted by Weber and his colleagues (1982) the same range of incentive conditions resulted in response rates of 38, 67, and 73 percent, respectively.

Berry and Kanouse (1987) compared the relative effectiveness of a prepaid and a promised incentive of $20 in a mail survey of physicians, obtaining a 78 percent response rate for the prepaid incentive group and a 66 percent rate for those paid only after they completed the survey. Mizes, Fleece, and Roos (1984) demonstrated the effectiveness of even a relatively small prepaid incentive in a brief mail survey of physicians, obtaining a response rate of 74 percent with either a $1 or $5 prepayment in comparison with 53 percent when no payment was provided. In a survey of 600 physicians from three specialty groups, Lockhart (1991) achieved a 57 percent response rate using a $20 prepaid incentive, compared with only 13 percent in a no incentive control group. More recently, Berk, Edwards and Gay (1993) confirmed the effectiveness of a prepaid incentive of $10 in a mail survey of physicians, achieving a response rate of 63 percent for those receiving the incentive with an initial mailing in comparison with 40 percent for the no incentive group.

The relative effectiveness of both large and small monetary incentives in improving response rates, speed, and/or quality among

other professional and elite populations has also been demonstrated empirically, including:

(a) international elites, ranging from university professors to cabinet ministers (Godwin, 1979);

(b) nurses (e.g., Kephart and Bressler, 1958);

(c) librarians (Hopkins, Hopkins, and Schon, 1988);

(d) various professionals subscribing to a magazine dealing with alcohol and drug use problems (Goodstadt, Chung, Kronitz, and Cook, 1977);

(e) owners of small construction subcontracting companies (James and Bolstein, 1992);

(f) community elites (Paolillo and Lorenzi, 1984);

(g) business executives (Erdos and Morgan, 1983; Robin and Walters, 1976); and

(h) farmers (Willimack, 1993).

However, some of these studies were poorly designed and/or obtained very low response rates even with incentives, and there are other studies where incentives used with professionals were either ineffective (e.g., Cook, Schoeps, and Kim, 1985; Abraham and Johnson, 1993) or resulted in poorer response rates or quality than when no incentives were provided (e.g., Hansen, 1980). For example, in the 1992 field test for a national survey of college and university faculty (Abraham and Johnson, 1993), three incentive conditions were used, including one monetary (a prepaid $2 bill); only the monetary incentive approached statistical significance in improving the response rate over no incentive (87 versus 79 percent), and this was not regarded as strong evidence of the efficacy of incentives with that particular professional population. In fact, the main survey, fielded without incentives, achieved an overall response rate of 87 percent (Abraham, 1994).

A second category of respondents which many nominate as hard-to-reach are adolescents, youth, and young adults. While there is less empirical evidence available with regard to these target populations, the results that are available are consistent with the assumption that incentives can be quite effective in stimulating survey cooperation among them. For example, a recent investigation conducted for NCHS by the Survey Research Center at the University of Michigan (Cannell and Camburn, 1991) studied the effects of respondent payments of $20 on the willingness of youth 12-20 to participate in the Youth Risk Behavior Surveillance System (YRBS) and on their motivation to answer YRBS questions as accurately and truthfully as possible. The results of this research indicated that paying respondents increased participation rates (from 79 to 90 percent), reduced parental consent refusal rates (thereby assisting in getting by an important "gatekeeper), aided

interviewers in converting refusals, increased respondents' perceptions of the importance of the survey, and may have improved the accuracy and honesty of responses.

The authors suggest that these youth represent the classic hard-to-interview respondent, in that the positive forces for them to respond and respond accurately to the YRBS are in fact quite low. "Young people had a low interest in this survey, and saw no compelling reason for responding honestly or being diligent in the task of answering the survey questions" (Cannell and Camburn, 1991, p. 1). Moreover,

> respondents could see no compelling reason to be diligent in answering survey questions. Participants in group interviews voiced skepticism about any benefits that may accrue from participating in surveys. Therefore, techniques typically used with adult respondents to encourage participation and more accurate reporting, such as appealing to their obligations as citizens [civic duty] or emphasizing the societal benefits likely to come from the survey, may be unproductive with youthful respondents. Therefore, some different ways of motivating respondent participation and accurate reporting are needed. (Cannell and Camburn, 1991, p. 11)

This is consistent, of course, with Cannell and Henson's (1974) earlier suggestion that this is precisely the role that a respondent incentive might play, i.e., providing a personal goal which motivates participation in a survey.

Similarly, students and former students have traditionally been difficult to survey. For example, "postsecondary students who have little motivation to participate in . . .research and have traditionally been difficult to survey include those who fail to complete the application process, those who are admitted but do not register for classes, dropouts, and alumni" (Zusman and Duby, 1987, p. 73). Thus, in a mail survey of undergraduate transfer students who subsequently withdrew, Zusman and Duby (1987) found that a prepaid incentive of $1 increased cooperation by nearly 20 percentage points. In an earlier study of the use of incentives with a postsecondary student survey, Huck and Gleason (1974) found that the response rate could be increased from 65 to 92 percent with the provision of an incentive. Peck and Dresch (1981) found that a prepaid $3 incentive with a 1½ hour mail survey of men and women three years after they completed high school yielded a response rate of 76 percent, compared with a 68 percent response rate for those promised a similar payment, and 54 percent among those who were offered no payment at all. Similarly, in a one-year follow-up mail survey of vocational-technical school graduates, Pucel, Nelson, and Wheeler (1971) found that the use of multiple nonmonetary incentives increased response rates by over 20 percentage points relative to a no incentive control group.

In contrast, in the 1992 Postcensal pretest for the NSF's 1993 National Survey of College Graduates, a mail survey with telephone and in person follow-ups, Mooney, Giesbrecht, and Shettle (1993) found that a $5 incentive with the initial mailing significantly increased response rates after two mailings by 11 percentage points, but, after mail and telephone follow-up, this difference was reduced to only 2 percent. In addition to increasing speed of response, however, those provided an incentive were significantly more willing to provide telephone numbers and names of contact persons, thereby reducing the effort required for future locating in this longitudinal study. In yet another incentive experiment conducted in connection with the National Assessment of Educational Progress (NAEP) household survey --in which young adults 26-35 were asked to complete a series of tests--Chromy and Horvitz (1978) demonstrated the effectiveness of a variable incentive procedure (no incentive for one package, $10 for two, $15 for three, four for $20) in increasing the overall response rate from 70.5 percent (with no incentive) to 83.3 percent, a rate which was maintained in subsequent years by adopting this procedure.

Perhaps the respondent categories most commonly mentioned as hard-to-reach or hard-to-interview are the educationally or economically disadvantaged and minorities. However, there is very little experimental evidence available attesting to the efficacy of respondent incentives based specifically on these target populations. One of the classic examples in fact overlaps our previous category, dealing as it does with disadvantaged youth. In that study (Kerachsky and Mallar, 1981), a $5 payment per interview was offered to a randomly selected portion of a national probability sample panel of 5,800 economically disadvantaged youth for three waves of interviewing, conducted in conjunction with an evaluation of the economic impact of the Job Corps program. "Youth in the age range of Corpsmembers (16 to 21) and with their economically disadvantaged backgrounds are generally very mobile and hard to locate [and interview]" (Kerachsky and Mallar, 1981, p. 263).

The researchers were able to verify the effectiveness of these respondent payments for improving both the quantity (search efficacy and interview completions) and quality (e.g., item nonresponse) of responses. After one or two interviews, the study's ability to locate potential respondents and obtain data from those who were located deteriorated in the absence of monetary incentives, but not when sample members were offered the $5 payment per interview. In addition, payment influenced the willingness of sample members to return postcards from advance letters, thereby reducing the cost of locating respondents for follow-up interviews. Significant reductions in item nonresponse associated with these payments were most pronounced at baseline and declined over time. More generally, the overall effectiveness of respondent incentives in increasing response rates and quality in studies that

overrepresent the poor and minority populations has also been demonstrated (e.g, Berk et al., 1987).

While studies of the effectiveness of incentives that focus directly on low income, minority, and disadvantaged populations are quite rare, those that address the more specific question of whether respondent incentives are *more* effective in improving cooperation among such target populations than the more affluent and advantaged are considerably more common. This question is part of a more general one--are incentives more effective for certain population subgroups than for others?--to which we now turn our attention.

3.3 Differential Effects of Incentives by Target Population

Restating our original question on this issue:
Are incentives effective only for certain target populations or subpopulations or are they more effective for certain population subgroups than for others (i.e., are the effects of incentives different for different population subgroups)?

At the most general level, a few of the quantitative reviews of the research literature on techniques designed to enhance mail survey response rates have addressed this issue. Yammarino and his colleagues (1991) found, for example, that associations between incentives and response rates "were statistically significant but derived from more than one population; that is, the relationships are situation specific and there is a need to examine potential moderators" (p. 627), other than year of publication and (a crude measure of the) type of sample (consumer versus institutional groups), the two moderators included in the study, which were not statistically significant. In contrast, neither Church (1993) nor Hopkins and Gullickson (1992) found significant differences in the effectiveness of incentives by target population (general, students, technical, business, and medical) or population type (professional, general, or semiprofessional), respectively.

However, a more detailed examination of respondent populations reveals a significant amount of variation in responsiveness to respondent incentives. For example, in an experiment recently reported by Willimack (1993) in connection with the 1992 Farm Costs and Returns Survey (FCRS), a prepaid nonmonetary incentive increased the overall response rate from 58 to 63 percent, but the incentive proved to be most effective in the smallest and largest classes of farm operations, increases of 17 and 12 percentage points, respectively.

Similarly, several studies of physicians have found significant differences by specialty in the effectiveness of incentives in increasing survey cooperation. Noting that different populations are more receptive to certain incentives than others, Erdos and Morgan (1983) reported that doctors in nine specialties

of medicine responded well to a 25¢ incentive, whereas psychiatrists required a $1 incentive. In the study cited earlier by Gunn and Rhodes (1981), response rates varied considerably across specialties: pediatricians and industrial physicians had high rates regardless of payment category ($0-$25-$50), but general and family physicians were very sensitive to payment--37 percent responded with no payment and 64 percent with $50. In the Lockhart (1991) study, the impact of incentives on general and family practitioners and physicians specializing in internal medicine was substantially greater than that observed among diabetes specialists, presumably because the topic of the survey--blood glucose monitoring--was especially salient to the latter. In contrast, in Berry and Kanouse's (1987) investigation of prepaid versus promised incentives, prepayment was effective with all nine specialties (statistically significant in four), except for oncologists, whose overall response rate was exceptionally high.

While it is commonly assumed that monetary incentives are more effective in increasing response rates among less educated, lower income respondents than among more educated, middle- or upper-income persons or households, research findings are in fact somewhat mixed regarding the effect of incentives on response rates for groups differing in socioeconomic status. For example, based on the early mail survey literature, Kanuk and Berenson (1975) cite several studies that failed to show that low income people were more likely to respond to monetary incentives than were people with higher incomes. In the 1971 National Health and Nutrition Examination Survey (NHANES I; Miller, Kennedy, and Bryant, 1972; Bryant, Kovar, and Miller, 1975), a $10 incentive increased examination cooperation rates relative to no incentive from 70 to 82 percent, but, contrary to expectations, the effect of the promised incentive did not increase as income level decreased. In a parallel remuneration experiment conducted in connection with the second survey (NHANES II) in this series (Findlay and Schaible, 1980), boosting the incentive from $10 to $20 increased the overall response rate from 74 to 79 percent, and the increased remuneration was more successful among whites than blacks, but there were no differences by income level. Goetz and his associates (1984) also found no difference in the effectiveness of incentives by education, race (cf. Dohrenwend, 1970), and income.

In contrast, in a nonexperimental comparison, Benus and Ackerman (1971) found that response rates for all major segments of the population were better in a panel where sample members were paid than a comparable panel where sample members were not paid, and they were disproportionately better among low income respondents. Similarly, using nonmonetary incentives in a mail survey, Nederhof (1983) found that incentives produced a disproportionately larger percentage of respondents from lower educational and occupational strata. Though modest, James and Bolstein (1990) found a similar pattern of relationships for level of education and income with various monetary incentives in their

mail survey of cable television subscribers. Ferber and Sudman (1974) also found that monetary incentives were indeed more effective in soliciting cooperation from lower income families than high income households, but such differential effects by education or income level have not been consistently observed in consumer expenditure studies (cf. Ferber, 1974; Cowan, 1977).

In a widely cited but poorly designed study, Gelb (1975) reported a significant difference in the response of lower-class and middle-class respondents to a conditional versus an immediate incentive to return a questionnaire, with middle-class recipients responding better to a prepaid and lower-class recipients to a promised incentive. More recently, Goyder (1990) reported a similar statistical interaction between socioeconomic status (SES) and size and type of incentive, whereby higher SES sample members were more likely to respond to no incentive or a prepaid $1 incentive, while lower SES sample members were more likely to respond to a post-paid $10 incentive than high SES sample members.

A more rigorous test of the differential effectiveness of monetary incentives by socioeconomic status was conducted in connection with a field test of 2,000 adults 16 and older for the National Adult Literacy Survey (NALS), a personal interview survey designed to measure one's ability to use printed and written material (Berlin, Mohadjer, Waksberg, Kolstad, Kirsch, Rock, and Yamamoto, 1992). Assessing the impact of incentives of $0, $20, and $35, a significant increase was found in response rates in the two incentive groups (81 and 83 percent, respectively) over the "no incentive" group (73 percent), but no significant difference by incentive level. Of particular importance to the current discussion, they found that incentives were most effective in improving response rates for people with low educational attainment and minority populations. When a monetary incentive was paid, significantly more black and Hispanic adults agreed to take both the background questionnaire and literacy test than when no incentive was offered, and similar results were observed for adults with lower levels of education. Similarly, in a recent pretest for Cycle V of the National Survey of Family Growth (NSFG), a survey of women of childbearing age with an oversample of black and Hispanic women, a $20 incentive resulted in an overall increase in cooperation of 8 percentage points, but the increase was disproportionately higher for black, Hispanic, and low income women (Duffer, Lessler, Weeks, and Moser, 1994). For example, among lower income women, the $20 incentive increased response rates by 12 percentage points, three times the level of increase (i.e., 4 percentage points) observed among higher income women. Similarly, the increases observed among black and Hispanic women were 10 and 26 percentage points, respectively, compared to only 2 percentage points for white and other women.

Thus, there is indeed some appreciable evidence that incentives are more likely to influence lower income, lower

socioeconomic subgroups and minority populations to participate in surveys. Overall, however, results regarding the effect of incentives on survey participation from groups differing by race/ethnicity or socioeconomic or status are currently still quite mixed.

3.4 The Conditional Use of Respondent Incentives

The foregoing discussion raises (but clearly doesn't answer) the question of whether all respondents should be paid the same incentive, or should consideration be given to different levels or types of remuneration for different respondents (cf. Peck and Dresch, 1981; Marrett, Kreiger, Dodds, and Hilditch, 1992). A related question raised earlier in our discussion of current practice is "should consideration be given to paying some but not all respondents to a given survey?" In fact, as noted earlier, a number of survey research organizations on some surveys use respondent incentives only for refusal conversion, usually only as a last resort to convert hard-core refusals, the truly hard-to-interview sample members. For example, in the General Social Survey (GSS), conducted by NORC under a grant from the National Science Foundation, "respondent fees" have routinely been used for this purpose; and, in 1989, up front respondent fees were also offered to a select group of respondents who resided in traditionally difficult urban primary sampling units (PSUs), a strategy that resulted in a lowering of the overall cost of the survey and a reduction in the field period, while still achieving the highest completion observed for this survey since 1985 (Law, 1989).

In a paper prepared for the COPAFS symposium in October 1992, Pendleton and Ginsberg (1992) explicitly raised this as an important research question (cf. Tucker, 1992):

> [Information is needed on] the most advantageous stage in the data collection process to offer incentives to minimize the cost and time involved in repeated follow up. A comparison of the cost of completed interviews when no incentives are offered, when offering incentives at first contact, and when offering incentives only when response rates are not found to be adequate in terms of cost of completed interview would be useful. (p. 8)

And, although none of them deal with the equity or fairness issues associated with this practice, a few studies have indeed explored the issue of the timing of when respondent incentives are offered, independent of the prepaid versus promised distinction, i.e., the "conditional" use of incentives to convert nonresponders.

For example, in a survey of college undergraduates, Huck and Gleason (1974) found that a follow-up mailing sending a quarter to nonrespondents rather than to everyone on the initial mailout list

cut costs in half, without significantly decreasing the response rate. In essence, the NAEP national assessment experiment with young adults reported by Chromy and Horvitz (1978) was a nonresponse follow-up study. After obtaining a disappointing response rate of only 44 percent with no incentive, the experiment was directed only at nonrespondents from the initial study. While nonrespondents receiving no incentive, but exposed to substantially improved field procedures, were increased to a participation rate of 71 percent, the response rates for those receiving one of three incentive conditions were boosted to 80 to 85 percent. Similarly, in the experiment with young adults surveyed by mail by Peck and Dresch (1981), a prepaid incentive of $3 in the first wave mailing resulted in a final response rate of 76 percent, compared to only 54 percent among those receiving no incentive; however, a condition in which nonrespondents to the first mailing who initially received no incentive were prepaid $3 with the second mailing ultimately achieved a response rate of 70 percent. In a household health survey of young adults in Switzerland, Perneger, Etter, and Rougemont (1993) experimented with two "incentives," a promise of 10 Swiss francs ($7 US) and a red reminder postcard mailed two days after the questionnaire. Initial response rates were 65 percent for those receiving both incentives, 57 percent for those offered the money only, 54 percent for recipients of the reminder card, and 48 percent for those who received neither. Follow-up mailings with incentives sent to all nonrespondents resulted in final response rates of 83, 84, 82, and 78 percent, respectively, attesting to the efficacy of these incentives as a productive nonresponse follow-up device.

In a mail survey of residents of New Zealand with three waves of mailing, a $1 incentive achieved a final response rate of about 70 percent, regardless of whether it was sent to all respondents with the initial mailing or included in the second mailing for nonrespondents to the first mailout (Brennan, Hoek, and Astridge, 1991). Similarly, in a mail survey of San Diego residents, a monetary incentive of $5 contingent on response (i.e., a promised incentive) to a second mailing of the questionnaire increased the response rate from initial "nonresponders" by 100 percent relative to controls who received no incentive, and by 75 percent over those who received $1 not conditional on response (Spry, Hovell, Sallis, Hofstetter, Elder, and Molgaard, 1989).

In a general population telephone survey on family health insurance recently conducted in 10 states for the Robert Wood Johnson Foundation by Mathematica Policy Research, Inc., promised incentives of $5 and $10 were compared with no incentive in three states to assess their effects on both response rates and data quality (Strouse and Hall, 1994). Although the $10 payment marginally increased screener (but not interview) response rates over no incentive prior to refusal conversion, after refusal conversion--which offered a $10 payment to nonresponders in both groups--cooperation rates for the $0 and $10 groups were virtually

equivalent. As a result, payments were retained only for refusal conversion efforts for the remainder of the survey.

Not surprisingly, the potential efficacy of using incentives only for nonresponse follow-up in surveys of physicians has also been examined, with mixed results. Recall that Berry and Kanouse (1987) achieved a 78 percent response rate with a $20 incentive included with the initial mailing, compared with only 66 percent for those who received a check only after the completed questionnaire was returned. However, a subsample of the original postpayment sample, sent a special follow-up mailing with a check, had a final response rate of 77 percent, indicating that prepayment was effective even if it was used late in the contact process.

In contrast, the recent study reported by Berk, Edwards, and Gay (1993) found that, while "some beneficial impact was found, delaying the incentive until the second round of mailing did not have the same effect as including an incentive with the initial mailing" (p. 241). Overall, a 63 percent response rate was obtained for those receiving a prepaid incentive with the initial mailing, compared with only a 50 percent response rate for those sent a prepaid incentive with a second mailing, and a 40 percent rate for those receiving no incentive.

4. Summary and Conclusion

As we noted in the introduction, the use of monetary incentives and other forms of respondent remuneration has become increasingly common in survey research, spreading rather steadily from the origin of this practice in commercial and market research to the increased use of remuneration in academic and government surveys. In any discussion of this trend among survey research professionals, examples of situations under which incentives are routinely touted as either necessary or highly desirable to achieve adequate response rates always include their use with hard-to-reach or hard-to-interview populations. Although this assertion could be regarded as essentially tautological, there is, in fact, a reasonable consensus on a broad conception of what constitutes "hard-to-reach" respondents, including: (a) hard core refusals; (b) sample members protected by gatekeepers; (c) frequently surveyed groups; (d) target populations or subpopulations among whom the traditional positive forces to cooperate are quite low; and, in general, (e) those who are hard to locate, gain access to, and interview for a variety of different reasons.

In summarizing what we know or think we know about the use of incentives with such populations, it would be most useful to do so in relation to the seven basic questions that we raised at the outset:

1. *Are there indeed specific target populations who are routinely offered remuneration to participate in surveys by most or all survey organizations because they are regarded as especially difficult to survey?*

The specific types of respondents designated as hard-to-reach by organizations that we contacted in preparing this paper include: (1) the economically disadvantaged; (2) the educationally disadvantaged; (3) minority populations; (4) adolescents, youth, and young adults; (5) drug users and those with special health problems; (6) frequently surveyed professional or elite populations; and (7) transients and persons who wish not to be found for legal or other reasons.

Of the these seven different categories of respondents, the research literature on the effects of respondent incentives has focused directly on only a few--frequently surveyed professional and elite populations, adolescents and young adults, and the disadvantaged. Overall, to varying degrees, the results of these studies suggest that respondent incentives can be quite effective in stimulating survey cooperation among each of these hard-to-reach populations, especially the first two categories.

2. *While respondent incentives may increase cooperation among initial refusals, are they really effective with hard-core refusals and the truly difficult or impossible to interview populations?*

That respondent incentives are quite effective in averting initial refusals or in converting them after the fact is fairly clear from the research literature, but none of the literature reviewed specifically addresses their efficacy with hard core refusals or those who are truly difficult or impossible to interview. Possible exceptions are the effectiveness of the NALS and NSFG Cycle V field experiments in achieving higher response rates among those subpopulations whose response rates are typically quite low, in spite of extensive follow-up and refusal conversion efforts. In addition, most survey research professionals believe that respondent incentives are an important overall tool in their arsenal for dealing with these difficult-to-interview populations in particular.

3. *Are incentives effective only for certain target populations or subpopulations or more effective for certain population subgroups than for others (i.e., are the effects of incentives different for different population subgroups)?*

While studies of the effectiveness of incentives that focus directly on low income, minority, and disadvantaged populations are quite rare, those that address the more specific question of whether respondent incentives are more effective in improving

cooperation among such target populations than the more affluent and advantaged are considerably more common. Overall, these studies provide some provocative evidence suggesting that incentives are indeed more likely to influence lower income, lower socioeconomic status and minority populations to participate in surveys than those more advantaged. Similarly, several studies of physicians have found significant differences by specialty in the effectiveness of incentives in increasing survey cooperation. More generally, there is a growing body of research evidence suggesting that incentives are more effective for certain populations or population subgroups than for others.

4. *Are incentives really effective in getting past "gatekeepers," either for certain professionals (e.g., physicians) or other difficult-to-survey subgroups of the general population?*

Evidence provided both by survey practitioners and the research literature suggests that respondent incentives can indeed be effective in getting past gatekeepers and gaining access to certain difficult to reach populations.

5. *Are incentives indeed of little use in locating and interviewing hard-to-find cases?*

A few studies and the experiences reported by several research organizations suggest that respondent incentives can indeed be effective in reducing the time and costs associated with locating, tracking, and interviewing highly mobile or otherwise difficult to locate populations.

6. *Should consideration be given to paying some, but not all respondents to a given survey?*

A number of survey research organizations on some surveys use respondent incentives only for refusal conversion, i.e., in general respondents are not paid, but incentives are offered as one tool in the refusal conversion process.

Although there are important questions to be answered related to the equity or fairness of this practice, most of the research literature bearing on its feasibility suggests that quite similar rates of response and response quality can be achieved in most surveys by delaying the use incentives to later stages of contact in the survey process.

7. *Should all respondents be paid the same incentive, or should consideration be given to different levels or types of remuneration for different respondents*

The research literature suggests that incentives (or incentives of a given level) are more effective for certain

population subgroups than for others, and that different respondent groups may be more sensitive to different levels or types (e.g., prepaid versus conditional) of remuneration than others. Thus, some have suggested that "an optimal survey design would probably use different incentives for different population groups" (Peck and Dresch, 1981, p. 256). However, while the observed variation across a number of studies is quite real, the evidence to date does not yet appear sufficient to justify the use of different incentives for different categories of sample members or survey respondents.

Overall, what does this overview suggest about how we define and approach hard-to-reach or hard-to-interview respondents and the relative or special efficacy of remuneration in surveys of these target populations or subpopulations? This question is best addressed perhaps by first approaching the problem from the other direction, i.e., by asking ourselves what motivates cooperative or "easy-to-reach" respondents to participate in surveys.

Fundamentally, it is important to realize that monetary incentives represent only one of many incentives or motivating factors available to survey researchers (cf. Groves, 1992). Survey practitioners already use a number of other types of incentives to encourage survey participation--including appeals to civic duty, the eventual use of the information provided to help people, etc.--and such appeals are generally quite effective. A sense of civic duty to respond to a government survey may be viewed as the cumulative result of the provision of goods and services by the polity to its citizens (Groves, 1992), and this "social contract" to provide information for the public good clearly appeals to many people. Thus, the techniques typically used by survey researchers to encourage participation include appealing to respondents' obligations as citizens and emphasizing the societal benefits likely to accrue from the survey.

In turn, these appeals seek to activate one of two basic types of respondent motive patterns (Cannell and Henson, 1974): (1) a perception that participation in the interview will enable one to achieve certain personal goals (i.e., by emphasizing the importance and purpose of the survey and attempting to link these to achieving some personal or societal goal); and (2) a habitual mode of response toward requests made by legitimate agencies or organizations in society, based on norms of good citizenship, politeness, acquiescence to requests for information, etc. Unfortunately, as Cannell and Henson (1974) note:

> these two types of motivations are not very effective in increasing respondent [cooperation or] activity. Respondents do not share the researcher's goals, or, if they do, they fail to see the interview as an effective way of achieving that goal. . . . Similarly, the respondent role may be seen as related to citizen responsibility, but this is not usually a

> sufficiently salient or strong motive to induce a high level of activity [or commitment] . . . (p. 313)

In effect, then, an incentive may be offered to respondents to create a personal goal which motivates participation where no other goal or motivation exists.

In this sense, hard-to-reach respondents are those for whom the positive forces to cooperate are quite low, where direct connections to personal goals cannot be readily established and appeals to civic responsibility or benefits to society are quite likely to fail. For example, economically and educationally disadvantaged populations--and others disenfranchised or alienated from the mainstream society--typically have no context for valuing research or their contribution to the research process and are less likely to be persuaded to participate in surveys on the basis of "social utility." It is thus generally unrealistic to expect them to give their time simply because of the potential policy implications of the data we ask them to provide. In short, one must recognize that there are indeed some subpopulations where the "social contract" has broken down, where the dominant mode of decision making is cost-benefit analysis (Groves, 1992) and where economic rather than social exchange processes are paramount.

It is at this intersection that we encounter many of those respondents who we designate as hard-to-reach, and where the use of remuneration takes on major significance as the incentive of choice, as those based on personal goals, obligations to society, and social utility become increasingly blurred or nonexistent. Although these segments of our society may well be growing, they still undoubtedly constitute a clear minority. Thus, a major challenge for those who choose to use respondent incentives to encourage participation among those who are hard-to-survey is to strike a pragmatic and appropriate balance between providing effective levels of remuneration to induce such groups to respond while simultaneously minimizing the risk of alienating those for whom the "social contract" is very much intact, among whom a sense of civic duty is alive and well, thereby continuing to provide a more than adequate basis for their participation in surveys they perceive to be of benefit to themselves, our society, and the nation.

REFERENCES

Abraham, S. Y. (1994, May). *Using pilot surveys to increase response rates: Findings from the 1992-93 National Study of Postsecondary Faculty.* Paper presented at the 49th Annual Conference of the American Association for Public Opinion Research, Danvers, MA.

Abraham, S. Y., and Johnson, R. A. (1993, May). *Effects of incentives in a multi-modal survey of elites: Findings from the National Study of Postsecondary Faculty.* Paper presented at the 48th Annual Conference of the American Association for Public Opinion Research, St. Charles, IL.

Armstrong, J. S. (1975). Monetary incentives in mail surveys. *Public Opinion Quarterly, 39*, 111-116.

Benus, J., and Ackerman, J. C. (1971). The problem of nonresponse in sample surveys. In J. B. Lansing, S. B. Withey, and A. C. Wolfe (Eds.), *Working papers on survey research in poverty areas* (pp. 26-59). Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan.

Berk, M. L. (1985). Interviewing physicians: The effect of improved response rate. *American Journal of Public Health, 75*, 1338-1340.

Berk, M. L., Edwards, W. S., and Gay, N. L. (1993). The use of a prepaid incentive to convert nonresponders on a survey of physicians. *Evaluation and the Health Professions, 16*(2), 239-244.

Berk, M. L., Mathiowetz, N. A., Ward, E. P., and White, A. A. (1987). The effect of prepaid and promised incentives: Results of a controlled experiment. *Journal of Official Statistics, 3*, 449-457.

Berlin, M., Mohadjer, L., Waksberg, J., Kolstad, A., Kirsch, I., Rock, D., and Yamamoto, K. (1992). An experiment in monetary incentives. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 393-398.

Berry, S. H., and Kanouse, D. E. (1987). Physicians response to a mailed survey: An experiment in timing of payment. *Public Opinion Quarterly, 51*, 102-114.

Blumberg, H. H., Fuller, C., and Hare, A. P. (1974). Response rates in postal surveys. *Public Opinion Quarterly, 38*, 113-123.

Brennan, M., Hoek, J., and Astridge, C. (1991). The effect of monetary incentives on the response rate and effectiveness of a mail survey. *Journal of the Market Research Society, 33*, 229-241.

Bryant, E. E., Kovar, M. G., and Miller, H. (1975). A study of the effect of remuneration upon response in the Health and Nutrition Examination Survey, United States. *Vital and Health Statistics*, Series 2-No.67. DHEW Pub. No (HRA) 76-1341. Washington, D.C.: U.S. Government Printing Office.

Cannell, C.F., and Camburn, D. (1991, June). *The effect of financial incentives on respondent participation in the Youth Risk Behavior Supplement*. Unpublished manuscript, Survey Research Center, University of Michigan, Ann Arbor, MI.

Cannell, C. F., and Henson, R. (1974). Incentives, motives, and responses bias. *Annals of Economic and Social Measurement, 3*, 307-317.

Cannell, C. F., and Fowler, F. J. (1977). Interviewers and interviewing techniques. In U.S. National Center for Health Services Research. *Advances in health survey research methods: Proceedings of a national invitational conference, 1975*. NCHSR Research Proceedings Series, DHEW Publication No. (HRA) 77-3154. Rockville, MD: Health Resources Administration.

Chromy, J. R., and Horvitz, D. G. (1978). The use of monetary incentives in national assessment household surveys. *Journal of the American Statistical Association, 73*, 473-478.

Church, A. H. (1993). Estimating the effects of incentives on mail response rates: A meta-analysis. *Public Opinion Quarterly, 57*, 62-79.

Cohen, S. B., Walden, D. C., and Ward, E. P. (1992, October). *Providing payments to respondents: The NMES experience*. Paper prepared for the COPAFS/OMB Symposium on Providing Incentives to Survey Respondents, John F. Kennedy School of Government, Harvard University, Cambridge, MA.

Cook, J. R., Schoeps, N., and Kim, S. (1985). Program responses to mail surveys as a function of monetary incentives. *Psychological Reports, 57*, 366.

Council of Professional Associations on Federal Statistics (COPAFS). (1993, September). *Providing incentives to survey respondents: Final report* (Contract No. GS0092AEM0914). Washington, DC: Regulatory Information Service Center, General Services Administration.

Cowan, C. D. (1977). Incentive effects on amounts reported in an expenditure diary survey. *Proceedings of the Social Statistics Section*, American Statistical Association, 498-503.

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*. New York: John Wiley.

Dillman, D. A. (1991). The design and administration of mail surveys. *Annual Review of Sociology, 17*, 225-249.

Dohrenwend, B. S. (1970). An experimental study of payments to respondents. *Public Opinion Quarterly, 34*, 621-624.

Duncan, W. J. (1979). Mail questionnaires in survey research: A review of response inducement techniques. *Journal of Management, 5*, 39-55.

Duffer, A. P., Lessler, J. T., Weeks, M. F., and Mosher, W. D. (1994, May). *Effects of incentive payments on response rates and field costs in a pretest of a national CAPI survey.* Paper presented at the 49th Annual Conference of the American Association for Public Opinion Research, Danvers, MA.

Erdos, P. L., and Morgan, A. J. (1983). *Professional mail surveys*(rev. ed.). Malabar, FL: Robert E. Krieger Publishing Company.

Ferber, R. (1974). Consumer expenditure and price data: An overview. *Annals of Economic and Social Measurement, 3*, 299-305.

Ferber, R., and Sudman, S. (1974). Effects of compensation in consumer expenditure studies. *Annuals of Economic and Social Measurement, 3*, 319-331.

Findlay, J. S., and Schaible, W. L. (1980). A study of the effect of increased remuneration on response in a health and nutrition examination survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 590-594.

Fox, R. J., Crask, M. R., and Kim, J. (1988). Mail survey response rate: A meta-analysis of selected techniques for inducing response. *Public Opinion Quarterly, 52*, 467-491.

Furse, D. H., and Stewart, D. W. (1982). Monetary incentives versus promised contribution to charity: New evidence on mail survey response. *Journal of Marketing Research, 19*, 375-380.

Furse, D. H., and Stewart, D. W. (1984). Manipulating dissonance to improve mail survey response. *Psychology and Marketing, 1* (2), 79-94.

Gelb, B. D. (1975). Incentives to increase survey returns: Social class considerations. *Journal of Marketing Research, 12*, 107-109.

Godwin, R. K. (1979). The consequences of large monetary incentives in mail surveys of elites. *Public Opinion Quarterly, 43*, 378-387.

Goetz, E. G., Tyler, T. R., and Cook, F. L. (1984). Promised incentives in media research: A look at data quality, sample representativeness, and response rate. *Journal of Marketing Research, 21*, 148-154.

Goodman, L. and Jensen, L. (1981). Economic surveys of medical practice: AMA's periodic survey of physicians, 1966-1978. In S. Sudman (Ed.), *Proceedings of the third biennial conference on health survey methods*. Rockville, MD: National Center for Health Services Research.

Goodstadt, M. S., Chung, L., Kronitz, R., and Cook, J. (1977). Mail survey response rates: Their manipulation and impact. *Journal of Marketing Research, 14*, 391-395.

Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25, 161-178.

Goyder, J. C. (1990). *Cash incentives and socio-economic bias in surveys*. Unpublished manuscript, Department of Sociology, University of Waterloo, Ontario, Canada.

Groves, R. M. (1989). *Survey errors and survey costs*. New York: John Wiley and Sons.

Groves, R. M. (1992, October). *Respondent incentives, minimum mean square error per unit cost, and the resident's responsibility to the polity*. Discussion notes prepared for the COPAFS/OMB Symposium on Providing Incentives to Survey Respondents, John F. Kennedy School of Government, Harvard University, Cambridge, MA.

Groves, R. M., Cialdini, R. B., and Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly, 56*, 475-495.

Gunn, W. J., and Rhodes, I. N. (1981). Physician response rates to a telephone survey: Effects of monetary incentive level. *Public Opinion Quarterly, 45*, 109-115.

Hackler, J. C., and Bourgette, P. (1973). Dollars, dissonance and survey returns. *Public Opinion Quarterly, 37*, 276-281.

Hansen, R. A. (1980). A self-perception interpretation of the effect of monetary and nonmonetary incentives on mail survey respondent behavior. *Journal of Marketing Research, 17*, 77-83.

Harvey, L. (1987). Factors affecting response rates to mailed questionnaires: A comprehensive literature review. *Journal of the Market Research Society, 29*, 341-353.

Heberlein, T. A., and Baumgartner, R. (1978). Factors affecting response rates to mailed questionnaires: A quantitative review of the published literature. *American Sociological Review, 43*, 447-462.

Hopkins, K. D., and Gullickson, A. R. (1992). Response rates in survey research: A meta-analysis of the effects of monetary gratuities. *Journal of Experimental Education, 61*(1), 52-62.

Hopkins, K. D., Hopkins, B. R., and Schon, I. (1988). Mail surveys of professional populations: The effects of monetary gratuities on return rates. *Journal of Experimental Education, 56*(3), 173-175.

Houston, M. J., and Ford, N. M. (1976). Broadening the scope of methodological research on mail surveys. *Journal of Marketing Research, 13*, 397-403.

Hubbard, R., and Little, E. L. (1988a). Promised contributions to charity and mail survey responses: Replication with extension. *Public Opinion Quarterly, 52*, 223-230.

Hubbard, R., and Little, E. L. (1988b). Cash prizes and mail survey response rates: A threshold analysis. *Journal of the Academy of Marketing Science, 16*, 42-44.

Huck, S. W., and Gleason, E. M. (1974). Using monetary inducements to increase response rates from mail surveys. *Journal of Applied Psychology, 59*, 222-225.

James, J. M., and Bolstein, R. (1990). The effect of monetary incentives and follow-up mailings on the response rate and response quality in mail surveys. *Public Opinion Quarterly, 54*, 346-361.

James, J. M., and Bolstein, R. (1992). Large monetary incentives and their effect on mail response rates. *Public Opinion Quarterly, 56*, 442-453.

Jobber, D., and Saunders, J. (1988). Modelling the effects of prepaid monetary incentives on mail-survey response. *Journal of the Operational Research Society, 39*(4), 365-372.

Kanuk, L., and Berenson, C. (1975). Mail surveys and response rates: A literature review. *Journal of Marketing Research, 12*, 440-453.

Kephart, W. M., and Bressler, M. (1958). Increasing the response to mail questionnaires. *Public Opinion Quarterly, 22*, 123-132.

Kerachsky, S. H., and Mallar, C. D. (1981). The effects of monetary payments on survey responses: Evidence from a longitudinal study of economically disadvantaged youths. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 258-263.

Kulka, R. A. (1992, October). *A brief review of the use of monetary incentives in federal statistical surveys*. Paper prepared for the COPAFS/OMB Symposium on Providing Incentives to Survey Respondents, John F. Kennedy School of Government, Harvard University, Cambridge, MA.

Law, J. W. (1989). *The impact of respondent fee payments on the 1989 General Social Survey*. Chicago: National Opinion Research Center, University of Chicago.

Linsky, A. S. (1975). Stimulating responses to mailed questionnaires: A review. *Public Opinion Quarterly*, 1975, *39*, 82-101.

Lockhart, D. C. (1991). Mailed surveys to physicians: The effect of incentives and length on the return rate. *Journal of Pharmaceutical Marketing and Management*, *6*(1), 107-121.

Marrett, L. D., Kreiger, N., Dobbs, L., and Hilditch, S. (1992). The effect on response rates of offering a small incentive with a mailed questionnaire. *Annals of Epidemiology*, *2*(5), 745-753.

Martin, B. (1984). *Don't survey physicians*. Chicago: American Medical Association.

McDaniel, S. W., and Rao, C. P. (1980). The effect of monetary inducement on mailed questionnaire response quality. *Journal of Marketing Research*, *17*, 265-268.

Miller. H. W., Kennedy, J., and Bryant, E. E. (1972). A study of the effect of remuneration upon response in a health and nutrition examination survey. *Proceedings of the Social Statistics Section*, American Statistical Association, 1972, 370-375.

Mizes, J. S., Fleece, E. L., and Roos, C. (1984). Incentives for increasing return rates: Magnitude levels, response bias, and format. *Public Opinion Quarterly*, *48*, 794-800.

Mooney, G. , Giesbrecht, L., and Shettle, C. (1993). *To pay or not to pay; That is the question*. Paper presented at the 48th Annual Conference of the American Association for Public Opinion Research, St. Charles, Illinois, May 20-23.

Nederhof, A. J. (1983). The effects of material incentives in mail surveys: Two studies. *Public Opinion Quarterly*, *47*, 103-111.

Paolillo, J. G. P., and Lorenzi, P. (1984). Monetary incentives and mail questionnaire response rates. *Journal of Advertising, 13*(1), 46-48.

Peck, J. K., and Dresch, S. P. (1981). Financial incentives, survey response, and sample representativeness: Does money matter? *Review of Public Data Use, 9*, 245-266.

Pendleton, A., and Ginsburg, A. (1992, October). *Providing incentives to survey respondents in education program evaluations.* Paper prepared for the COPAFS/OMB Symposium on Providing Incentives to Survey Respondents, John F. Kennedy School of Government, Harvard University, Cambridge, MA.

Perneger, T. V., Etter, J-F., and Rougemont, A. (1993). Randomized trial use of a monetary incentive and a reminder card to increase the response rate to a mailed health survey. *American Journal of Epidemiology, 138*, 714-722.

Robin, D. P., and Walters, C. G. (1976). The effect on return rate of messages explaining monetary incentives in mail questionnaire studies. *Journal of Business Communications, 13*, 49-54.

Shuttleworth, F. K. (1931). A study of questionnaire technique. *Journal of Educational Psychology, 22*, 652-658.

Sudman, S. (1985). Mail surveys of reluctant professionals. *Evaluation Review, 9*, 349-360.

Sudman, S., and Ferber, R. (1974). A comparison of alternative procedures for collecting consumer expenditure data for frequently purchased products. *Journal of Marketing Research, 11*, 128-135.

Skinner, S. J., Ferrell, O. C., and Pride, W. M. (1984). Personal versus nonpersonal incentives in mail surveys: Immediate versus delayed inducements. *Journal of the Academy of Marketing Science, 12*, 106-114.

Spry, V. M., Hovell, M. F., Sallis, J. G., Hofstetter, C. R., Elder, J. P., and Molgaard, C. A. (1989). Recruiting survey respondents to mailed surveys: Controlled trials of incentives and prompts. *American Journal of Epidemiology, 130*(1), 166-172.

Strouse, R., and Hall, J. (1994, May). *Effects of payments on response rates and data quality for a general population telephone survey.* Paper presented at the 49th Annual Conference of the American Association for Public Opinion Research, Danvers, MA.

Tambor, E. S., Chase, G. A., Faden, R. R., Geller, G., Hofman, K. J., and Holtzman, N. A. (1993). Improving response rates through incentive and follow-up: The effects on a survey of physicians' knowledge of genetics. *American Journal of Public Health, 83*, 1599-1603.

Tucker, C. (1992, October). *Providing incentives to respondents in government surveys*. Paper prepared for the COPAFS/OMB Symposium on Providing Incentives to Survey Respondents, John F. Kennedy School of Government, Harvard University, Cambridge, MA.

Weber, S. J., Wycoff, M. L., and Adamson, D. R. (1982). *The impact of two clinical trials on physician knowledge and practice*. Arlington, VA: Market Facts, Inc.

Weinberg, E. (1973). The issue of payment to respondents--wise or unwise. *Public Opinion Quarterly, 37* (3), 478-480.

Weiss, C. H. (1975). Interviewing in evaluation research. In E.I. Struening and M. Guttentag (Eds.), *Handbook of evaluation research* (pp. 355-395). Beverly Hills, CA: Sage Publishing Company.

Willimack, D. K. (1993). Selected results of the incentive experiment on the 1992 Farm Costs and Returns Survey. *Proceedings of the International Conference on Establishment Surveys* (pp. 490-495). Alexandria, VA: American Statistical Association.

Willimack, D. K., Petrella, M., Beebe, T., and Welk, M. (1992, August). *The use of incentives in surveys: Annotated bibliography*. Survey Research Center, Institute for Social Research, University of Michigan.

Wotruba, T. R. (1966). Monetary inducements and mail questionnaire response. *Journal of Marketing Research, 3*, 398-400.

Yammarino, F. J., Skinner, S. J., and Childers, T. L. (1991). Understanding mail survey response behavior: A meta-analysis. *Public Opinion Quarterly, 55*, 613-639.

Yu, J., and Cooper, H. (1983). A quantitative review of research design effects on response rates to questionnaires. *Journal of Marketing Research, 20*, 36-44.

Zusman, B. J., and Duby, P. (1987). An evaluation of the use of monetary incentives in postsecondary survey research. *Journal of Research and Development in Education, 20*(4), 73-78.

Summary of Resolvable Errors in Edit Output

Benzene

chckcode codable primary occupation (qprmy_cd) - housewife (aqid=03233);
freqs qisfx2 = 0 for 1 case
rtickicd look into the operative procedure codes on page 1 - should these codes be based on residual body site?
page 61, aqid=01187: can this be coded per earlier error resolution

TCE

mainedit flag155-multibirth no, 1 preg, 2 livebirths- aqid=06011
chckcode 1 case w/emp_cd=10- do not assign- aqid=08222
phone area codes- page 1 -obs 29 and 30: 910 should 919
duplicate phone #:
area code=815; qid=009020090-check last name-it appears it may be the same person who was a proxy with same first and middle names but different last name
addresses with more than 1 phone:
251 ...- fix 6th digit in 1st phone #
4329.. - put last digit in 3rd position for first observation

street page 69: blocks 5 and 6
page 88 blocks 4 and 5
page 126 - last 2 blocks

Transaction log

1) The transaction log should be log of CHANGES MADE to database - not changes INTENDED to be made - describe how this log is created with respect to the SAS program where corrections are programmed.

2) provide report of errors - changes not made- as provided earlier.

3) local variables (i.e. AQID=2795: LREGZIP) should never be changed

4) need consistent handling of #'s(i.e. AQID=0118: "#10"), apt, NO (1st vs. 2nd line of address)

5) re-sort this report: aqid, variable name, reverse chronological date and time

6) appears to be 2 qids for aqid=2923. Please look into this and also provide proc freq/list of aqid * qid.

7) NORC needs to do very careful QA/QC of transaction log prior to finalizing files and sending final deliverable - perhaps forward this to ATSDR for review electronically (scrambled) prior to review of other deliverables

specific problems:

AOID	Problem
0118	"Court" should be standardized
0192	"Calendar, CT" appears in final edit output, but in this log seems t be an old value that was changed to "Venice, CA". How can this be?
0802	Cancer condition (2/91) appears to have occurred BEFORE last intervie date
2692	Variable Name blank
2876	PK should be PARK
2923	*2 qids?? why so many changes - is this correct? need details on case
3954	condition (D/89) appears to have occurred BEFORE last interview date

DISCUSSION

Diane K. Willimack
National Agricultural Statistics Service

I would like to begin by congratulating and commending these authors on two thorough and well-presented papers. Kulka has written a theory section that sets the stage for incentive use in surveys and raises conceptual issues for consideration by survey designers. He goes on to provide a thorough review of the literature and a thoughtful evaluation of the implications for survey designers, particularly regarding "hard-to-reach" population subgroups.

Ezzati, White, Mosher, and Sanchez -- the NCHS authors plus one -- have provided a compilation of the findings from a series of experiments in which incentive use has been systematically studied on a wide variety of survey types and target populations -- from hospitals to households, from physicians to disadvantaged population subgroups. In addition they report the value to interviewers of incentive use: how respondent incentives enhance interviewers' confidence and their ability to do their job, as well as how interviewer incentives enhance their morale and motivation. This summary challenges us all to consider how to apply their findings to our own survey situations, so that we need not reinvent the wheel.

Allow me to share what I learned from these two papers. Kulka differentiates the social norm of reciprocity and social exchange theory from the theory of economic exchange as the conceptual basis for incentive use in surveys. Typically we describe incentives by their type -- monetary vs. nonmonetary -- and timing -- prepaid vs. promised. Let us focus on timing. Use of prepaid incentives is based on the social theories, while promised incentives are believed to invoke economic exchange in respondents.

These authors tend to use the terms "incentive" and "remuneration" interchangeably. Kulka also uses the word "compensation". I decided to look up these words in my Random House Collegiate Dictionary (1988). I know that our use of words as jargon need not have any relationship with the English language, but I thought this exercise might be instructive, as well as help me clarify my thoughts.

"Incentive," in my dictionary, is defined as "something that incites to action," and lists as synonyms, "stimulus, spur, incitement, encouragement." Let me interpret: encouragement to respond. "Remuneration" is "something that remunerates; reward; pay." "Remunerate" means "to recompense," which means "to repay or reward (someone), as for aid or service." Again I offer my interpretation: to recompense (repay) for aiding us by responding to our survey. Finally, "compensation" is "something given or received as an equivalent for services, etc.," with synonyms "recompense, remuneration, payment." It seems to me that "remuneration" and "compensation" are interchangeable with each other, but not with "incentive."

Let us return to theory. One of the problems with promised incentives is that, in many studies comparing them to prepaid incentives, the amount of money or the item was token in nature. Thus while the "promise" tends to invoke economic exchange in respondents, the token nature was not sufficient to influence respondents to engage in the economic contract. But remuneration does

invoke economic exchange, since it implies an amount of money sufficiently close to the value of the service of survey response rendered by the respondent.

Thus, if an incentive is token in nature, it must be prepaid, for it can only rely upon social norms in order to be successful. If a promised incentive is to be used, then, for it to be successful, it must be substantial enough to be considered remuneration or compensation in an economic exchange. Of course the latter must take into consideration the difficulty of the burden of the task of completing the survey. Perhaps I have just exhibited a firm grasp of the obvious, but this notion did not crystalize for me until I read these two papers side by side.

However, tokens appear to work in instances when signs of appreciation and good faith are meaningful to the respondent. For example, the National Agricultural Statistics Service (NASS) conducted an incentive experiment on the 1992 Farm Costs and Returns Survey, an annual voluntary survey of U.S. farm operators collecting detailed expenditure and income data in personal interviews lasting 90 minutes on average. Response rates were increased by nearly 5.5 percentage points in the group receiving a prepaid nonmonetary incentive. Of greater interest, though, was the finding that incentive use increased response rates by 17 points among the smallest farms (those with sales less than $20,000) and by nearly 12 points among the largest farms with sales of $500,000 or more. The token pocket portfolio and calculator that we gave them cannot have indicated an economic exchange to these large farms. Instead, it likely had appeal as a symbol of appreciation consistent with the repeated survey contact to which these two groups, in particular, are subjected (Willimack, 1993).

Farms are establishments, and NASS surveys collect establishment information from them. Although farms may be a special case because they exhibit many of the same characteristics as households, application of incentives or remuneration in establishment surveys is not trivial. The NCHS paper reports a survey of hospital records, in which hospitals are paid for their service of abstracting sampled records. This is clearly remuneration with basis in economic exchange. However this may not always be so clear cut for incentive use or remuneration in establishment surveys, particularly surveys of businesses.

Incentives are meant to "incite action," to influence the decision to participate in a survey and to motivate the respondent. But in an establishment survey, it is unclear who we are attempting to influence with incentive use. In an establishment, the person making the decision about survey participation may not be the desired respondent. The desired respondent is the person who is the most knowledgeable provider of the information being sought, usually the person who has access and understanding of any records to be used as a source for responding (Edwards and Cantor, 1991).

If a prepaid incentive is provided in an establishment survey, who gets it? Is it the boss, in order to "incite" or encourage a favorable decision about responding? Or is it the employee who actually completes the survey instrument, in order to motivate careful response? Or is it the business, as income or as payment for the service of completing the survey? It does seem less difficult to make this "leap" if it is remuneration being offered rather than an incentive. Remuneration is more clearly a payment to the business for the service of completing the questionnaire. But then the dollar amount must be reasonable relative to the burden of the task.

Let us consider further the dollar amount to be used as an incentive or as an offer of remuneration. I think money is tricky, unless the amount is clearly token. Money is a sensitive topic for many people, and it has all kinds of different connotations. No matter how much is offered, there will always be someone for whom that amount is not enough. Furthermore, since respondents use all the information available to them when responding to survey questions, it seems reasonable to suggest that the same is true in those initial moments of contact during which a survey participation decision is being made. A key piece of information is the dollar amount being offered. I think the dollar amount provides an indication to respondents about the potential difficulty of the survey task, and may arouse suspicion. The amounts of $20, $30, $50, even $100 in several of the health studies described by the NCHS authors may have provided a fair indication of the difficulty of the task. On the other hand, consider the James and Bolstein (1992) study cited by Kulka in which a promise of $50 failed to increase response rates in a survey of construction subcontractors. Similarly, in the HIV survey described in the NCHS paper, $175 offered to survey refusers increased response rates by only 4 percentage points, while $100 resulted in a 10 point increase. These inordinately large dollar amounts may have indicated a difficult survey task to the respective target populations, resulting in a sense, "Well, if you're offering that much money, then what you're asking me to do must be really hard, embarrassing, or uncomfortable. It must be something that I don't want to do." Meanwhile, the choice to accept or reject the offer -- to engage in the economic contract -- remains with the respondent.

The amount of money offered as an incentive or as remuneration is an important variable to the survey designer. Here we have reached the edge of a gaping hole in the literature on incentives. What is the trade-off between the incentive or remuneration amount and the difficulty of the survey task? Indeed, under which survey circumstances will a token incentive suffice, and when is remuneration needed? These papers seem to suggest that when the survey task is particularly intrusive, burdensome, lengthy, or longitudinal, then the promise of remuneration seems appropriate, if not necessary. But these factors must be evaluated relative to the expectations and perceptions of the members of the target population: Perhaps a personal interview lasting 1½ to 3 hours and requesting detailed expense and income data is considered by farm operators to be as personal and sensitive as the Survey of Family Growth asking women aged 15-44 about abortion and sexual practices. As discussants are required to say, "more research is needed."

Furthermore, participants at the 1992 COPAFS conference on incentives listed guidelines for OMB to consider when evaluating incentive use in Federal surveys (COPAFS, 1993). These include:

"To compensate a respondent if there is risk in participating."

"When there are unusual demands or intrusions on the respondent (e.g., lengthy interviews, keeping a diary, having a blood sample drawn, ...)."

"When sensitive questions are being asked."

"If there is a lengthy field period (e.g., a commitment over time for a panel survey)."

"If there is any out-of-pocket cost to the respondent ..."

"If the respondent is a small business or a nonprofit institution in a voluntary survey and the respondent perceives some cost and burden to participating."

Although these represent only a subset of COPAF's recommendations to OMB, it appears to me that they outline survey circumstances that favor remuneration, as I have defined it in this discussion. However, there may be survey situations in which an incentive, as I have differentiated it, will suffice. If OMB policy appears to favor remuneration over incentive use, then agencies will design surveys accordingly, with consequences for survey budgets, management, and respondent burden. OMB needs to write policy that recognizes and encourages appropriate use of both incentives and remuneration.

How do we evaluate the effectiveness of incentives or remuneration in surveys? In the early days, as Kulka points out in his literature review, it was as simple as "Were response rates increased by incentive use?" A resounding "YES" is supported by a large number of studies on various types of surveys. Soon attention turned to the quality of the data, adding the question, "Does incentive use improve the quality of the data through the respondent's increased attentiveness to the survey task, particularly reflected in reduced item nonresponse?", or, at least, "Can we be sure that data quality has not been reduced?" Again, both Kulka and the NCHS authors provide a great deal of evidence supporting improved data quality related to incentive use or remuneration.

Let me next turn attention to what I call potential "dividends" to be achieved from incentive use: reduction in components of total survey error. Both sets of authors provide a number of examples of increased survey participation among "hard-to-reach" or typically under-represented population subgroups. In addition, the NCHS authors have provided results that show systematic differences on key variables among groups whose participation was increased by remuneration. Thus, these papers present evidence that nonresponse bias in survey estimates may be reduced through use of incentives or remuneration.

Moreover, let me offer an additional "dividend" found in two incentive experiments in which I have been involved: enhanced ability to identify ineligible sample units, resulting in nonsampling error reduction. In an incentive experiment on the 1991 Detroit Area Study, a statistically significant increase was found in the rate of ineligible sample addresses. This may have been due, in part, to more diligent postal returns of the small packages containing a prepaid pen incentive, along with more reliable interviewer confirmation, when an address was a vacant housing unit, a business, or an incorrect address (Willimack and others, 1994). A replication of this result is currently being tested in the 1994 Detroit Area Study.

In addition, in the incentive experiment conducted on the 1992 Farm Costs and Returns survey, a higher rate of sample units being screened out as non-farms, and therefore not eligible for the survey, was found to be statistically significant. Incentive recipients who had no agriculture may have been more attentive to the survey request and more determined to notify the interviewer of their non-farm status, rather than to refuse or be inaccessible based on a belief that the survey did not apply to them (Willimack, 1993).

Incentive studies of the 1990's appear to have added cost evaluations or cost-benefit analyses of incentive use or remuneration. The results of the pretest of the Survey of Family Growth are particularly compelling. They show that the $20 incentive/remuneration actually saved money over

no incentive, since these respondents required fewer contacts and were less likely to break appointments, resulting in reduced interviewing costs.

Cost evaluation is an area that still needs development. Achieved cost savings may not always be so clear. Instead, frequently we are left to judge whether observed benefits were sufficient to justify the additional costs of incentive/remuneration use. It is difficult to put a dollar value on increased response rates, improved data quality, and reduction in total survey error. No doubt we all consider these to be very valuable. But how much improvement is needed in order to justify the additional cost of incentives or remuneration? What production and efficiency measures should be monitored during data collection and post-survey review and processing? What comparisons with which other survey methods should be undertaken? We must be able to evaluate whether incentive use or remuneration is the best tool to pull out of our methodology tool box.

Again I congratulate the authors, and I thank them for their contributions to our understanding of incentive use in surveys.

REFERENCES

Council of Professional Associations on Federal Statistics (COPAFS) (1993), Providing Incentives to Survey Respondents: Final Report (Contract No. GS0092AEM0914), Washington, D.C.: Regulatory Information Service Center, General Services Administration.

Edwards, W. S., and Cantor, D. (1991), "Toward a Response Model in Establishment Surveys," Ch. 12 in Measurement Errors in Surveys, edited by P. P. Beimer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman, New York: John Wiley and Sons, Inc., pp. 211-233.

James, J. M., and Bolstein, R. (1992), "Large Monetary Incentives and Their Effect on Mail Survey Response Rates," Public Opinion Quarterly 56:442-453.

The Random House Collegiate Dictionary (1988), New York: Random House, Inc.

Willimack, D. K. (1993), "Selected Results of the Incentive Experiment on the 1992 Farm Costs and Returns Survey," Proceedings of the Interantional Conference on Establishment Surveys, Alexandria, VA: Amaerican Statistical Association, pp. 490-495.

Willimack, D. K., Schuman, H., Pennell, B. E., and Lepkowski, J. M. (1994),"The Effects of a Prepaid Nonmonetary Incentive on Response Rates and Response Quality in a Face-to-Face Survey," unpublished manuscript, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI.

DISCUSSION

W. Sherman Edwards
Westat, Inc.

The use of monetary incentives in surveys is an endlessly fascinating topic for those of us in the business of collecting and analyzing survey data. Regardless of the importance and policy relevance of any particular survey, regardless of the eminent history and usefulness of the statistics derived from a survey, at some point in data collection it gets down to the trenches -- getting the last reluctant respondent to cooperate, finding the last elusive subject, persuading the last indifferent citizen to answer what seem to him like senseless and repetitive questions. The use of monetary incentives to help in this unavoidable struggle raises interesting methodological questions and, even more provocatively, questions about our relationship with our respondents. As a society, we are strongly ambivalent about money; it is the most public of our life's trappings, and yet at the same time it is the most private. Every survey researcher knows that it is not questions about sexual behavior or the intimate details of health that elicit the highest item refusal rates -- it is questions about income and financial assets. Introducing money into any relationship that is not primarily an economic one complicates that relationship, as anyone who has loaned money to a relative or friend can attest.

My discussion is organized first around two specific methodological questions and then addresses the relationship issue. Although the literature on the effects of monetary incentives is extensive, I will suggest several areas that warrant considerable further study.

The first methodological question is the most basic, "Do monetary incentives improve response rates?" The two papers today provide a fairly unequivocal "Yes" response to this question for surveys where incentive experiments were conducted. The paper by Dick Kulka examines the use of incentives for hard-to-survey populations. This is a very comprehensive paper; it provides a thorough and very useful overview of the research literature on incentives, of different theoretical views of how incentives work, and of the methodological issues raised when considering monetary incentives. The paper then goes on to explore in more detail issues around incentives for hard-to-survey populations. Although this comment is not particularly relevant to my discussion, I can't help but note that I never before considered the commonalty among drug users, physicians, youth in general, CEOs, impoverished urban minorities, diabetics, unwed mothers, small farm operators, and those defaulting on student loans. (They are all considered hard to survey, and thus worthy of consideration for monetary incentives.)

The paper by Trena Ezzati and her colleagues reviews the experiences with monetary incentives of several surveys about health and the use of health care services conducted by the National Center for Health Statistics and the Agency for Health Care Policy and Research. This paper presents a more detailed look at a smaller number of studies, including some that have pushed at the edges of the envelope containing this issue.

Kulka surveyed participants in the 1993 COPAFS conference on incentives, and found their use widespread. I think it's fair to characterize the view of many survey firms towards

incentives as one of a number of tools that they can use to achieve satisfactorily high response rates, but one that is often used with some reluctance. Other tools in this kit, of course, include repeated callbacks to convert initially reluctant respondents, the use of specially designed persuasive materials that may be targeted to particular reasons for refusal, and the use of especially effective interviewers. How do we choose among these tools in designing our data collection strategies? The most sympathetic view is that we choose rationally, based on empirical evidence and careful consideration of all alternatives, choosing the most cost-effective procedures possible. A more cynical view might be that many surveys or survey organizations opt for incentives because they are easier to implement than other approaches. I suspect that both views have some truth. For mail surveys, a fair amount of thought and research have been devoted to optimal design strategies, most notably Dillman's Total Design Method (Dillman, 1978). The literature on telephone surveys is somewhat less well developed, and that on in-person still less. This is in part because the issues become more complex as the relationship between survey and respondent becomes more personal. Recent work by Groves and Couper (1994, e.g.) is notable in applying constructs and research techniques from several disciplines to the tool kit. One lesson from their work is that we still have a lot to learn about the tools, including when and how to use them.

An important part of the question of whether incentives work is whether they are cost-effective. Some of the research studies cited in the two papers included formal evaluations of cost-effectiveness. The recent NSFG evaluation, for example, found that the cost per case for the $20 incentive treatment was lower than for the non-incentive treatment, in addition to yielding higher response rates. Other studies have examined the cost of monetary incentives against other methods, such as repeated mailings, to achieve comparable response rates. Often, incentives seem to reduce the cost of surveys The cost-effectiveness of a monetary incentive is related to the mode of administration and the size of the incentive. The marginal cost of a $20 incentive for an in-person survey is clearly much lower than the cost of a similar incentive in a mail survey. Kulka mentions the possible effectiveness of an incentive in locating elusive respondents. In-person locating is one of the most expensive of data collection activities -- even a small improvement in locating effectiveness would probably be worth the cost of a monetary incentive judiciously mentioned to an informant.

The second methodological question is, "What effect do monetary incentives have on data quality?" Kulka notes two commonly expressed views from the research literature: the first that respondents view incentives as a kind of "social exchange," and thus work harder and provide better responses. The second is that incentives are an "external motivator," decreasing internal motivation and hence reducing the quality of responses. Here, the literature is somewhat less persuasive. Some studies, like the NSFG experiment and the seroprevalence survey pilot cited by Ezzati, take the "more is better" view, which is probably often appropriate, particularly when the results are as dramatic as in these studies. Few studies have combined incentive experiments with validation of data, which is often difficult or even impossible. In particular, how does data quality differ between those who would (or do) participate without an incentive and those for whom the incentive makes the difference? To examine this, we might use respondents' perceptions about the use of incentives as an explanatory variable for some measures of data quality.

The only study cited that explicitly explored respondents' attitudes was associated with the NHIS Youth Risk Behavior Supplement, as cited by Kulka. Kalton and his colleagues used follow-up cognitive interviews with youths responding to a field test of the survey to discover an apparently strong link between the incentive and the effort expended in answering questions. This seems an especially promising methodology for exploring the effects of incentives, particularly on "special" or hard-to-survey populations.

Monetary incentives are frequently used in diary surveys -- such use is even sanctioned by the Office of Management and Budget. Could a monetary incentive affect the behavior of interest by a diary keeper? At the recent AAPOR conference, Diane Woodard of Arbitron (1994) described a series of incentive experiments for respondents in their radio listenership survey. One treatment involved a sweepstakes with a $3,000 first prize in addition to the usual nominal monetary incentive. In this treatment, all radio ratings increased. The post hoc explanation of this phenomenon was that respondents had increased their radio listening (or at least their reporting) in the erroneous belief that this would increase their chances of winning the sweepstakes. Another explanation, of course, is that the sweepstakes respondents were simply reporting better, although I am inclined to accept Woodard's interpretation.

A strongly held view of many survey researchers is that all respondents should be subject to exactly the same stimulus, as nearly as this can be controlled by the survey design. Only in this way can we be confident of the reliability of survey responses. This view has come under increasing criticism from those who view the structured interview as too restrictive a vehicle for meaningful exchange of information. For the more traditional view, monetary incentives may throw a monkey wrench in the works. As Kulka notes, some surveys offer incentives only to "hard-core" refusals or to those who are expected to be difficult to locate or interview. Even when the same incentive is offered to all respondents, their reactions to being "paid" for their time undoubtedly vary considerably. In every survey I've been associated with that offered incentives, some respondents refused to accept them. Again, it would seem useful to ask respondents how they feel about the payment, and to examine how responses vary by these views. At the heart of the matter, do monetary incentives introduce more variation in respondent motivation and perception of the survey than otherwise exists?

Let me now turn to the issue of the relationship between the survey and the respondent. If we define this relationship in terms of the use of incentives, there seem to be three views of the nature of the relationship. The first, which I would associate with not using incentives, is that surveys are a social good, and that participation alone is sufficient reward. Clearly, government surveys are the most logical candidates for this high-minded view, and there is some evidence in support of it. The Census Bureau does not use incentives, and yet achieves the highest response rates in the industry, even setting aside their mandatory surveys.

Those of us in the private sector, however, tend to be fairly pragmatic, if still somewhat ambivalent, as Kulka reports, about the use of incentives. What claim does Westat or anyone else outside of the government have on respondents, even if we do often represent the government? (I note that all of the surveys described in the Ezzati et al paper were conducted by private contractors.) Even if we have ethical concerns about the fairness of paying respondents in some surveys but not others, or of paying the "hard-core" but not the compliant, are the

alternatives any more attractive? Is it ethically more appealing to ask again and again, or to fly in the charming interviewer from Indiana, than to offer a monetary incentive?

Another view of the relationship between survey and respondent is that of the "social exchange" that Kulka describes. I would suggest that many who conduct survey research are most comfortable with this view. I have heard many researchers, some in describing incentive experiments, say, "The interviewers feel better when they can give the respondent something." Interviewers in this case, as they so often are, are our proxies. Each of today's papers had one mention of the positive effect of respondent incentives on interviewer behavior. This is another area that warrants further study.

One of the groundbreaking studies described by Ezzati et al is the pilot test of the National Household HIV Seroprevalence Survey, conducted by RTI in Dallas. Here, the size of the incentives offered ($50, $100, $175) seem to go well beyond what would be called for in a social exchange view of the interview. Respondents were asked in return for a blood sample and to complete a brief sexual history questionnaire. However, the context of this survey was socially charged. Respondents might be fearful of the blood draw in general, but particularly because it was associated with an AIDS survey. One view of the size of the incentive was that it was an appropriately attractive exchange for asking respondents to overcome a whole set of fears.

The third view of monetary incentives in this relationship is that of a business transaction, or "economic exchange," as it is described in the Kulka paper -- we are paying the respondent for his or her time. The implication of this view is that we are participants in an information marketplace, subject to the principles of supply and demand. In government household surveys, even those conducted by contractors, most researchers shy away from this view, and even explicitly deny it. Even some incentive experiments wind up giving all respondents the same payment. The survey of allergists cited by Ezzati is an example -- regardless of the incentive treatment, all participating physicians were eventually paid the same amount.

One is really taking an economic view when one uses disproportionate payments for different survey respondents, including payment to initial nonresponders but not to compliers. Here, we pay according to how badly we want the information -- truly supply and demand. This practice is anathema to some researchers, who cite the ethics of fairness and their distaste for rewarding noncompliant behavior.

Where monetary incentives have sometimes become a business transaction is, not surprisingly, in surveys of businesses. The prediction about respondents coming to expect incentives has probably come to pass for physician surveys, although Kulka notes that some are still conducted successfully without incentives. Ezzati et al describe the experience of the National Hospital Discharge Survey, where hospitals are routinely paid piecework for completing survey forms. The National Medical Expenditure Survey's Medical Provider Survey also allowed payment for survey respondents, the amount to be negotiated, but only if requested by the medical provider. On the other hand, most Federal establishment surveys do not offer reimbursement on demand or incentives.

One argument put forward against monetary incentives is that respondents will come to expect payment, and response rates will be even harder to achieve at a reasonable cost. This is referred to in the summary of last year's COPAFS symposium as the "slippery slope" argument. In essence, this argument says that the use of incentives breeds the perception of a survey marketplace. It seems unlikely that this effect would occur among the general public solely from Federal surveys. Even with the current rate of surveying, most households are never or rarely selected for a government survey (excepting the decennial Census, of course). If we are in or to be in a survey marketplace in this sense, commercial surveys will have had more to do with that result, and OMB has no control over their behavior. The "slippery slope" is more likely to refer to the behavior of survey organizations if we come to rely too routinely on monetary incentives to achieve our response rate goals.

One noticeable feature of both of these well-prepared and very interesting papers is the number of experiments that have been conducted around incentives, particularly for Federal surveys. I suspect that as we in the statistical community continue to explore the complexities of our relationship with our survey respondents, we will continue to conduct incentive experiments at an unflagging pace. There is certainly much more to learn about the effects of incentives on data quality, and about alternative methods for influencing respondent compliance.

REFERENCES

Dillman, D.A. (1978), Mail and Telephone Surveys: The Total Design Method, New York: Wiley.

Groves, R. and Couper, M. (1994), "Understanding the Process of Survey Cooperation: Household-Interviewer Doorstep Interactions," paper presented at the 49th Annual Conference of the American Association of Public Opinion Research, Danvers, MA.

Woodard, D. (1994), "Raising Response Rates with Premiums, or Is $1 Worth Anything to Anyone Anymore?" paper presented at the 49th Annual Conference of the American Association of Public Opinion Research, Danvers, MA.

Session 7

COMPUTER ASSISTED SURVEY INFORMATION COLLECTION

REDESIGNING A QUESTIONNAIRE FOR COMPUTER-ASSISTED DATA COLLECTION: THE CURRENT POPULATION SURVEY EXPERIENCE

Cathryn Dippo and Anne Polivka
Bureau of Labor Statistics

Kathleen Creighton, Donna Kostanich, and Jennifer Rothgeb
Bureau of the Census

Over the last decade, there have been two new factors that have significantly influenced the design of survey data collection--the computer and the theories and methods of cognitive psychology. When, in 1986, staffs of the Bureau of Labor Statistics (BLS) and the Bureau of the Census initiated a process for redesigning the Current Population Survey (CPS), incorporating these two new factors was made a top priority. In the paper[1], the authors[2] illustrate how, by concentrating on the cognitive processes of respondents and interviewers, computer-assisted interviewing was used as a tool for reducing measurement error.

The following topics are covered in the paper: background material on questionnaire design and computer-assisted interviewing methodologies, development of the CPS questionnaire over the last 50 years and how redesigning the CPS questionnaire for the 21st century has brought together the two new methodologies, using the computer in evaluating alternative questionnaire designs, examples of the new CPS questionnaire's design features which aid the cognitive processes of the respondent and interviewer and are primarily dependent on the use of the computer, the effects of the new questionnaire and collection procedures on labor force estimates, and a discussion of issues for the future.

The Current Population Survey is a monthly survey of approximately 60,000 households. The CPS survey, conducted for BLS by the Bureau of the Census, is the primary source of information on the U. S. labor force. Each month BLS analyzes and publishes information from the CPS, such as the unemployment rate, demographic characteristics of individuals in the labor force, and the number of hours individuals work. The survey began in 1940 under the auspices of the Works Projects Administration and was called the Monthly Report of Unemployment. The current CPS questionnaire has remained essentially

---

[1] The presentation by Cathryn Dippo at the Seminar on New Directions in Statistical Methodology was based on a paper currently under review for publication in a refereed journal. Thus, only a brief synopsis is being published here, along with a detailed bibliography of papers related to the CPS redesign.

[2] The new CPS questionnaire is the result of a team effort which involved many staff members from both BLS and Census. Space does not allow us to recognize everyone. The other members of the BLS-Census Questionnaire Design and Overlap Analysis Steering Committees over the years were Chester Bowie, John Bregger, Shail Butani, Lawrence Cahoon, Kennon Copeland, Harvey Hamel, Elizabeth Martin, Michael McMahon, Thomas Scopp, Clyde Tucker, Ronald Tucker, and Alan Tupek.

unchanged since the last major revision in January 1967. With only minor exceptions, the concepts measured have remained constant since the late 1940's.

Over its 50+-year history, the CPS has continued to be a model for survey designers. It was the first national probability sample of households, and many of the statistical methods for sampling and estimation now considered common practice were originally researched and implemented in CPS. Two of the six research areas identified in 1986 related to data collection--computer-assisted interviewing and the questionnaire. A Questionnaire Design Task Force was established to identify the cognitive and conceptual problems in the existing questionnaire, to suggest possible solutions for identified problems, and to develop a research plan to design and test a new questionnaire, along with related survey procedures. A separate task force was established to investigate the potential uses of computer-assisted interviewing. When a final consolidated research plan was approved in 1988, a major premise of the plan was that all interviews would be conducted using a computer. Following a period of questionnaire development and extensive testing, Census began collecting all CPS data using a new fully-automated questionnaire in January 1994.

The data produced from the CPS are closely-watched by economic forecasters and policy analysts. Therefore, all changes had to be carefully researched prior to implementation. By concentrating on facilitating the cognitive processes used by respondents and interviewers, research on alternative measurement processes resulted in reduced nonsampling errors. By capitalizing on the power and versatility of the computer, new research tools were developed to provide the evidence needed to understand the effects of changes in data collection procedures. We hope that the approach used for developing the new measurement process for CPS will serve as a model for future survey redesign projects.

For details on the 8 years of research that went into redesigning the CPS, please consult the papers listed in the following bibliography.

1994 CPS REDESIGN BIBLIOGRAPHY

Bienias, J. L. (1988). Preliminary laboratory research on Current Population Survey questionnaire with employed persons. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 185-190.

Bienias, J. L., Dippo, C. S., and Palmisano, M. (Eds.) (1987). *Questionnaire design: Report on the 1987 BLS Advisory Conference*. Washington, DC: U.S. Bureau of Labor Statistics.

Boehm. L. E. (1989). Reliability of proxy response in the Current Population Survey. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 486-489.

Boehm, L. E. (1989, August). The relationship between confidence, knowledge, and performance in the Current Population Survey. Paper presented annual meeting of the Southeastern Psychological Association, Washington, DC.

Bowie, C., Cahoon, L., and Martin, E. (1993). Overhauling the Current Population Survey: Evaluating changes in the estimates. *Monthly Labor Review, 116(9)*, 29-33.

Bregger, J. (1977). Establishment of a new Employment Statistics Review Commission. *Monthly Labor Review, 100 (3)*, 14-20.

Bregger, J. (1984). The Current Population Survey: A historical view and BLS's role. *Monthly Labor Review, 107 (6)*, 8-14.

Bregger, J. (1994, May). *Raising the Nation's unemployment rate*. Paper presented at the Seminar on New Directions in Statistical Methodology, Bethesda, MD.

Bregger, J., and Dippo, C. S. (1993). Overhauling the Current Population Survey: Why is it necessary to change? *Monthly Labor Review, 116(9)*, 3-9.

Bushery, J. M., Weidman, L., Briley, S. M., and Ciochetto, S. (1987). Using CATI in the Current Population Survey. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 778-783.

Butani, S., Cahoon, L., Fay, R., and Kostanich, D. (1994, in press). Measurement of different design effects. *American Statistical Association Proceedings of the Section on Survey Research Methods.*

Butz, W. P., and Plewes, T. J. (1989). A Current Population Survey for the 21st century. *Bureau of the Census Proceedings of the Fifth Annual Research Conference*, 3-13.

Cohany, S. R., Polivka, A. E., and Rothgeb, J. M. (1994). Revisions in the Current Population Survey effective January 1994. *Employment and Earnings, 41-2*, 13-37.

Cahoon, L. S. (1990, 1991, 1992, 1993). *Results of the CPS CATI Phase-in analysis*. Series of internal memoranda (issued quarterly). Washington, DC: Bureau of the Census.

Cahoon, L. S., and Russell, G. (1991, October). *Effects of CATI on estimates from demographic surveys conducted by the Bureau of the Census*. Paper presented at the

Joint Census Advisory Committee Meeting of the American Marketing Association and the Committee on Population Statistics, Alexandria, VA.

Campanelli, P. C., and Esposito, J. L. (1992, May). *The use of behavioral coding to identify problematic questions on alternative questionnaires*. Paper presented at the annual American Association for Public Opinion Research Conference, St. Petersburg, FL.

Campanelli, P. C., Martin, E. A., and Creighton, K. P. (1989). Respondents' understanding of labor force concepts: Insights from debriefing studies. *Bureau of the Census Proceedings of the Fifth Annual Research Conference*, 361-374.

Campanelli, P. C., Martin, E. A., and Rothgeb, J. M. (1991). The use of respondent and interviewer debriefing studies as a way to study response error in survey data. *The Statistician, 40*, 253-264.

Campanelli, P. C., Rothgeb, J. M., and Martin, E. A. (1989). The role of respondent comprehension and interviewer knowledge in CPS labor force classification. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 425-430.

Cantwell, P. J., and Ernst, L. R. (1993). Short-term changes to the CPS composite estimator in January 1994. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 724-729.

Cantor, D., and Esposito, J. L. (1992). Evaluating interviewer style for collecting industry and occupation information. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 661-666.

Collins, C. L. (1975). *Comparison of month-to-month changes in industry and occupation codes with respondent's report of change: CPS job mobility study*. Response Research Staff Report No. 75-5. Washington, DC: U.S. Bureau of the Census.

Copeland K. R., and Rothgeb, J. M. (1990). Testing alternative questions for the Current Population Survey. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 63-71.

Couper, M. P., and Burt, G. (1993, in press). Interviewer reactions to computer-assisted personal interviewing (CAPI). *Bureau of the Census Proceedings of the Annual Research Conference*, 429-443.

Department of Labor. (1981). *Final report of the Secretary of Labor on the Recommendations of the National Commission on Employment and Unemployment Statistics*. Transmitted to Congress, October 26.

Dippo, C. S. (1993). Designing an instrument for computer-assisted data collection in the Current Population Survey. *Bulletin of the International Statistical Institute*, Contributed Papers 49th Session, 1, 373-374.

Dippo, C. S., Kostanich, D. L., and Polivka, A. E. (1994, in press). Effects of methodological change in the Current Population Survey. *American Statistical Association Proceedings of the Section on Survey Research Methods*.

Dippo, C. S., Polivka, A. E., Creighton, K. P., Kostanich, D. L., and Rothgeb, J. M. (1994). *Redesigning a questionnaire for computer-assisted data collection: The Current Population Survey experience*. Paper submitted for publication.

Edwards, W. S. (1991, May). *Measuring question comprehension and its effects on response error: An "hours worked" case study*. Paper presented at the annual meeting of the American Association for Public Opinion Research, Phoenix, AZ.

Edwards, W. S., Levine, R. and Cohany, S. (1989). Procedures for validating reports of hours worked and for classifying discrepancies between questionnaire reports and validations totals. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 496-501.

Esposito, J. L., Campanelli, P. C., Rothgeb, J. M., and Polivka, A. E. (1991). Determining which questions are best: Methodologies for evaluating survey questions. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 46-53.

Esposito, J. L., and Hess, J. C. (1992, May). *The use of interviewer debriefings to identify problematic questions on alternative questionnaires*. Paper presented at the annual American Association for Public Opinion Research Conference, St. Petersburg, FL.

Fisher, R. (1994, in press). Logistic regression analysis of CPS Overlap Survey split panel data. *American Statistical Association Proceedings of the Section on Survey Research Methods*.

Fisher, R., Thompson, J., Robison, E., and Welch, M. (1993). Variance estimation in the CPS Overlap Test. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 802-807.

Fracasso, M. P. (1989). Reliability and validity of response categories for open-ended questions in the Current Population Survey. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 481-485.

Fracasso, M. P. (1989, August). *A procedure for the development of an alternative set of category choices using the open-ended questions in the Current Population Survey*. Paper presented at the annual meeting of the Southeastern Psychological Association, Washington, DC.

Gaertner, G. H., Cantor, D., Gay, N. L., and Shank, S. E. (1989). Tests of alternative questions to elicit occupational information in the CPS. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 490-495.

Hamel, H. (1979). Two fifths of discouraged workers sought work during prior six month period. *Monthly Labor Review, 102*, 58-60.

Hess, J. (1993). *Results of the MLS Interviewer Knowledge Concepts Test*. Memorandum for the CPS Questionnaire Evaluation Steering Committee. Washington DC: Bureau of the Census.

Kojetin, B. A., and Cashman, E. R. (1994, May). *Prior interviewing and computer experience and evaluation of CAPI training*. Paper presented at the annual Field Technologies Conference, Danvers, MA.

Kominski, R., and Siegel, P. M. (1993). Measuring education in the Current Population Survey. *Monthly Labor Review, 116(9)*, 33-38.

Kostanich, D. L., and Cahoon, L. S. (1994). *Effect of design differences between the parallel survey and the new CPS*. CPS Bridge Team Technical Report 3, dated March 4.

Lent, J., Miller., S, and Cantwell, P. (1994, in press). Composite weights for the Current Population Survey. *American Statistical Association Proceedings of the Section on Survey Research Methods.*

Martin, E. A. (1987). Some conceptual problems in the Current Population Survey. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 420-424.

Martin, E. A., Campanelli, P. C., and Fay, R. E. (1991). An application of Rasch Analysis to questionnaire design: Using vignettes to study the meaning of "work" in the Current Population Survey. *The Statistician, 40*, 265-276.

Martin, E. A., and Polivka, A. E. (1992). The effect of questionnaire redesign on conceptual problems in the Current Population Survey. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 655-660.

Martin, E. A., and Polivka, A. E. (1994). *Diagnostics for redesigning survey questionnaires: Measuring work in the Current Population Survey*. Paper submitted for publication.

McKay, R. B. (1992). Cultural factors affecting within household coverage and proxy reporting in Hispanic (Salvadoran) households. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 614-618.

McKay, R. B. (1993). Undercoverage of Hispanics in household surveys. *Monthly Labor Review, 116(9), 38-42.*

McKay, R. B. (1994, in press). The Spanish translation of the redesigned CPS: Lessons learned. *American Statistical Association Proceedings of the Section on Survey Research Methods.*

McKay, R. B. (1994, in press). Touch-tone data entry for CPS sample expansion. *American Statistical Association Proceedings of the Section on Survey Research Methods.*

McKay, R. B., and LaVallee, A. (1993, May). *The Hispanic version of the redesigned CPS questionnaire: Applying sociolinguistic and survey research methods to translating survey questionnaires*. Paper presented at the annual meeting of the American Association for Public Opinion Research, Pheasant Run, IL.

McKay, R. B., Malik, A., and Tupek, A. (1992, May). *Potential use of touch-tone recognition technology in the Current Population Survey*. Paper presented at the annual Field Technology Conference, St. Petersburg, FL.

McKay, R. B., and Sangster, R. L. (1994, May). *Touch-tone data entry for household reporters*. Paper presented at the annual Field Technologies Conference, Danvers, MA.

Miller, S. M. (1994). *What would the unemployment rate have been had the redesigned Current Population Survey been in place from September 1992 to December 1993? : A measurement error analysis*. CPS Bridge Team Technical Report 1, dated March 3.

Miller, S. M. (1994, in press). Measurement error modeling of the CPS and parallel survey data. *American Statistical Association Proceedings of the Section on Survey Research Methods.*

Mullin, P. A., and Tonn, B. E. (1993). Factors associated with proxy knowledge of employment-related information. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 948-953.

National Commission on Employment and Unemployment Statistics. (1979). *Counting the Labor Force*. Washington, DC: Government Printing Office.

Palmisano, M. (1988). *Some results on the application of laboratory research techniques to survey methodology*. Paper presented at the meetings of the American Association for the Advancement of Science.

Palmisano, M. (1989, August). *Respondent understanding of key labor force concepts used in the CPS*. Paper presented at the annual meetings of the American Statistical Association, Washington, DC. (Not published in conference proceedings).

Polivka, A. E. (1994). *Comparisons of labor force estimates from the parallel survey and the CPS during 1993: Major labor force estimates*. CPS Overlap Analysis Team Technical Report 1, dated March 18.

Polivka, A. E., and Martin, E. A. (1992, May). *The use of vignettes in pretesting and selecting questions*. Paper presented at the annual meeting of the American Association for Public Opinion Research, St. Petersburg, FL.

Polivka, A. E., and Miller, S. M. (1994, December). *The CPS after the redesign: Refocusing the economic lens*. Paper to be presented at the Bureau of Economic Research conference.

Polivka, A. E., and Rothgeb, J. M. (1993). Overhauling the Current Population Survey: Redesigning the questionnaire. *Monthly Labor Review, 116(9)*, 10-28.

President's Committee to Appraise Employment and Unemployment Statistics. (1962). *Measuring employment and unemployment*. Washington, DC: U.S. Government Printing Office.

Rothgeb, J. M. (1982). *Summary Report of July Follow-up of the Unemployed*. Unpublished memorandum. Washington, DC: Bureau of the Census.

Rothgeb, J. M. (1994). *Revisions to the CPS questionnaire: Effects on data quality*. CPS Overlap Analysis Team Technical Report 2, dated April 6.

Rothgeb, J. M., and Cohany, S. R. (1992). The revised CPS questionnaire: Differences between the current and proposed questionnaires. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 649-654.

Rothgeb, J. M., and Hess, J. C. (1992, May). *The role of response distribution and item non-response analyses in evaluating alternative question wordings during the redesign of the Current Population Survey*. Paper presented at the annual meeting of the American Association for Public Opinion Research, St. Petersburg, FL.

Rothgeb, J. M., Polivka, A. E., and Creighton, K. P. (1992). *Measurement changes in the redesigned CPS questionnaire*. Paper presented at the Joint Committee Meeting of the Census Advisory Committee of the American Statistical Association and the Census Advisory Committee on Population Statistics.

Rothgeb, J. M., Polivka, A. E., Creighton, K. P., and Cohany, S. R. (1991). Development of the proposed revised Current Population Survey questionnaire. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 56-65.

Schwartz, N. (1987, June). *Cognitive aspects of labor surveys in a multi-national context*. Paper prepared for the OECD Working Party on Employment and Unemployment Statistics.

Shoemaker, H. H. (1993). Results from the Current Population Survey CATI Phase-In project. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 710-714.

Shoemaker, H. H., Bushery, J. M., and Cahoon, L. S. (1989). Evaluation of the use of CATI in the Current Population Survey. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 361-366.

Stein, R. L., and Levine, D. B. (1965). Research in labor force concepts. *American Statistical Association Proceedings of the Social Statistics Section*, 218-226.

Tanur, J. M., and Lee, J. (1992). Youth unemployment measured by the CPS. *American Statistical Association Proceedings of the Section of the Section on Survey Research Methods*, 667-671.

Tanur, J. M., Lee, J., and Shin, H. (1991). Intensity and conceptualizations of job search among youths. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 391-396.

Tanur, J. M., Lee, J., and Shin, H. (1992). Reports of youths' job searches in the CPS. *Bureau of the Census Proceedings of the Annual Research Conference*, 317-352.

Tanur, J. M., and Shin, H. (1990). Reporting job search among youths: Preliminary evidence from reinterviews. *Bureau of the Census Proceedings of the Annual Research Conference*, 746-776.

Tanur, J. M., and Shin, H. (1990, August). *Youths, reporting job search, and the Current Population Survey*. Paper presented at the annual meetings of the American Statistical Association, Anaheim, CA. (Not published in conference proceedings).

Thompson, J. (1994). *Mode effects analysis of major labor force estimates*. CPS Overlap Analysis Team Technical Report 3, dated April 14.

Thompson, J., and Fisher, R. (1994, in press). Two-sample McNemar Tests for complex survey data. *American Statistical Association Proceedings of the Section on Survey Research Methods*.

Thompson, J., and Parmer, R. (1994, in press). A comparison of nonparametric methods with parametric methods for the CPS CATI/CAPI mode effects analysis. *American Statistical Association Proceedings of the Section on Survey Research Methods*.

Tiller, R. (1994, in press). Predicting the National unemployment rate that the old CPS would have produced. *American Statistical Association Proceedings of the Section on Survey Research Methods*.

Tiller, R., and Welch, M. (1994). *Predicting the national unemployment rate that the "old' CPS would have produced*. CPS Bridge Team Technical Report 2, dated March 15. (See also addenda dated March 4 and April 1)

U.S. Bureau of Labor Statistics. (1986). *Report of the BLS-Census Bureau Questionnaire Design Task Force*. Staff report dated November. Washington, DC: U.S. Department of Labor, Bureau of Labor Statistics.

U.S. Bureau of Labor Statistics. (1987). *Second Report of the BLS-Census Bureau Questionnaire Design Task Force*. Staff report dated February. Washington, DC: U.S. Department of Labor, Bureau of Labor Statistics.

U.S. Bureau of Labor Statistics. (1988, June). *Response errors on labor force questions based on consultations with Current Population Survey interviewers in the United States*. Paper prepared for the OECD Working Party on Employment and Unemployment Statistics.

U.S. Bureau of Labor Statistics and National Center for Education Statistics. (1979). *Feasibility study for a project on improvement of occupational information*. Report of the National Occupational Information Coordinating Committee, U.S. Department of Commerce Contract No. AO-A01-788-00-1349. Washington, DC: U.S. Bureau of Labor Statistics and National Center for Education Statistics.

U.S. Bureau of the Census. (1988). *Final report on planning for a post-1990 redesign of the Current Population Survey*. Unpublished report.

Waksberg, J., and Pearl, R. B. (1965). New methodological research on labor force measurements. *American Statistical Association Proceedings of the Social Statistics Section*, 227-237.

Walsh T., et al. (1988). *Evaluation of CATI data quality and costs in the Current Population Survey*. Unpublished research report dated October 27. Washington DC: Bureau of the Census.

Westat, Inc., and AIR, Inc. (1989). *Research on hours of work questions in the Current Population Survey*. Final Report, Bureau of Labor Statistics Contract No. J-9-J-8-0083. Rockville, MD: Westat, Inc.

Westat, Inc., and AIR, Inc. (1989). *Research on industry and occupation questions in the Current Population Survey*. Final Report, Bureau of Labor Statistics Contract No. J-9-J-8-0083. Rockville, MD: Westat, Inc.

Westat, Inc., and AIR, Inc. (1992). *Summary of activities to evaluate the dependent interviewing procedure on the Current Population Survey*. Final Report, Bureau of Labor Statistics Contract No. J-9-J-8-0083. Rockville, MD: Westat, Inc.

AN OVERVIEW OF THE 1992 CENSUS OF AGRICULTURE CATI SYSTEM

Jeanette K. Mon
Bureau of the Census

I. Introduction

Computer assisted telephone interviewing, or CATI, is an enumeration methodology in which responses received by telephone are interactively entered, edited, and coded into data files. The CATI system adopted by the Bureau of the Census provides call scheduling management, contains manager search functions, and produces various monitoring/progress reports.

Uses of the CATI system can vary from collecting present indicators for research polling to accepting detailed enumeration for surveys and/or censuses. Responses can be used to produce skip patterns and are subjected to consistency and magnitude checks. In a typical situation, the interviewer reads the question displayed on the computer screen to the respondent and records the response by keying the appropriate entry; then, the computer performs checks (i.e., validity, comparative), stores the response, and proceeds to the next question. This process continues until all questions have been asked.

For the 1992 Census of Agriculture, the CATI system was used to perform follow-up action for specific nonrespondents. This paper is intended to provide an overview of the developmental and processing phases of this system as well as the handling of output resulting from the 1992 Census of Agriculture CATI System.

II. Background

In 1973, Census Bureau executives became interested in CATI after seeing a demonstration by a private research firm. After several years of research and consultation with other such firms, universities, and computer vendors, the Bureau tested CATI at the University of California's CATI site during the 1978 Current Population Survey. In the early 1980's, the Census Bureau established a CATI project. Hardware and software were acquired in order to construct the Bureau's own CATI system. During the project's first year, staff reviewed the design and capabilities of CATI systems at academic institutions and private firms. Working with the Berkeley and Michigan Survey Research centers, the staff prepared basic requirements for the Census CATI system. From 1982 through 1984, Census conducted CATI research and development surveys from a Telephone Bridge Facility set up in Suitland, Maryland headquarters. The first major testing of this system was conducted for followup of nonrespondents in the 1982 National Survey of Natural and Social Scientists and Engineers, and the 1982 Census of Agriculture.

The first Census Bureau telephone facility, the Hagerstown Telephone Center, (HTC), opened in January 1985. An additional telephone facility opened in Tucson, Arizona in early 1992.

The census of agriculture is required by law under Title 13, United States Code, section 142(a) and 191, which states that an agriculture census be taken in 1979, 1983, and in every fifth year after 1983, covering the prior year. As previously mentioned, Agriculture Division's initial use of the CATI system in the 1982 Census of Agriculture was part of a test to review the system as a viable method of data collection. Agriculture Division (AGR) selected approximately 10,000 delinquent large farm cases for enumeration using this system. The follow-up process for the remaining cases used a clerical unit of operators who called respondents and manually recorded data on an agriculture report form.

For the 1992 Census of Agriculture, a CATI system was developed primarily to address the "large farms" which had not responded to the mailed questionnaire as in 1982. The reasons for using the CATI system for followup were management efficiency, cost effectiveness, and availability of the operation (staff/hardware) from the decennial census. There were also other advantages such as eliminating the data keying step, promoting a paperless census, and using CATI's management capabilities for monitoring and scheduling cases.

III. Overview of the Agriculture CATI System

A. Agriculture Division's Use of CATI

In addition to large farm followup, the AGR CATI system was used for the Nonresponse Survey and low response county projects. The Nonresponse Survey involved contacting a sample of nonrespondents from the main agriculture census. Data collected for this survey was used to determine the proportion of nonresponse cases that are farms and to weight census totals to account for the nonresponse. In the low response county project, a sample of nonrespondents in counties that had not achieved a 75% response rate were selected for contact. The goal of this project was to ensure that published 1992 Census of Agriculture data were based on responses from at least 75% of each county's mailout cases. In the Nonresponse Survey a different (shorter) version of the interviewing instrument was used while the low response county project used the large farm follow-up instrument. Since the system used for large farm followup provided the groundwork for these two projects, this paper will deal only with AGR's experience with the large farm CATI system.

B. System Development

The development of the AGR CATI system was conducted by a Committee responsible for: writing system specifications; developing essential system components; administering tests; coordinating facility schedules; and providing training.

CATI Committee

The 1992 Census of Agriculture CATI Committee consisted of 15 knowledgeable individuals from five divisions within the Census Bureau that are familiar with the various parts of the system. Even though much effort went into the brainstorming, learning, and decisionmaking involved with the development phase, this group remained functional throughout the implementation process and, also assisted in the evaluation of the system. From the onset of the planning phase (August 15, 1991) to the close of the AGR CATI operations (September 30, 1993) several committee personnel changes occurred; however, in most instances, the strategic persons remained involved in this task. The Committee consisted of representatives from:

- AGR to specify the needs for data collection;
- Economic Programming Division (EPD) to facilitate the input and output data;
- Field Division (FLD) to implement the operation with Hagerstown and Tucson staffs;
- Demographic Surveys Division (DSD) to program the QISC interviewing instrument; and,
- Systems Support Division (SSD) to provide support of the CATI system.

For the first six months, the Committee held biweekly meetings to plan the development phase. During this period, the Committee prepared a flowchart and activity schedule; identified and assigned the required specifications; and, scheduled the project for testing and production at the CATI facilities. Attachments A and B are the processing flowchart and the Committee's activity schedule, respectively. After the planning phase, the Committee divided into workgroups to facilitate the development of their respective CATI tasks and reconvened monthly to assure the ongoing progress of each assignment.

During the remainder of 1992, AGR discussed aspects of the interviewing instrument such as sequence and wording of questions, availability and content of referral screens, and consistency checks between responses. This exchange of ideas

resulted in additional refinements in the instrument.

CATI Specifications

As required of all CATI projects, basic system specifications are necessary from the sponsor to provide the framework for system development. Such standard specifications were identified by FLD and SSD committee members and delegated to appropriate members for development.

Many of these specifications required AGR personnel to go through a "learning" period prior to development. In some instances, the concept and format for these documents as recognized by CATI system personnel were complex and involved considerable time for one to become knowledgeable enough to prepare the required documents.

Agriculture Division wrote several other CATI specifications and procedures to explain issues such as training, problem solving, and handling other details not addressed by the basic CATI system specifications. Attachment C is a list of the CATI specifications.

System Tests

Three tests were conducted to refine the system for production. These tests (September 28 & October 28, 1992 and January 14, 1993) were generally conducted in the same manner. Experienced interviewers at the Hagerstown CATI facility telephoned AGR personnel for enumeration. The "mock respondents" were comprised of AGR staff from several different areas of responsibility. Some were given scripts of varying situations (i.e., nonagriculture, refusal, complete) while others presented their own scenario. The tests checked for the following items:

- questions needing rewording and/or additional clarification;
- appropriate routing of interviewing screens from question to question;
- correct output coding for case types and responses (keycoded & nonkeycoded);
- appropriate transfer and installation of files;
- appropriate input file content;
- improvement ideas from interviewers and/or "mock" respondents; and,
- other aspects within the process requiring attention.

The tests also provided sample output files and status tables for review. The output files were passed on to EPD for subsequent reformat testing. The status tables were reviewed for format changes and/or programming errors.

Training

In addition to the usual CATI system training provided to interviewers, AGR supplemented the training with a 2-day intensive instruction course. This course included technical subject-matter learning, pronunciation of practical agricultural terms, a short session in unit conversion using a calculator, practice interviews, and other exercises needed for enumeration. Attachment D is the training schedule.

Training materials for the AGR CATI project were written in FLD's Training Branch based on information submitted by AGR. The materials included:

- Self Study Guide,
- Workbook for Training,
- Paired Practice Interviews Booklet,
- Final Review Exercises,
- Guide for Training CATI Interviewers, and
- Evaluation of Self Study and Classroom Training.

During the training sessions, a reference binder of general information was given to each participant. In view of the voluminous amount of technical detail involved with the AGR subject matter, interviewers were instructed to review this material and use it for assistance as needed. The contents of this binder included a report form guide, alphabetic crop listing, unit conversion chart, and glossary of terms.

Input File Preparation

Each state was processed as a separate file/survey. EPD was responsible for the creation of forty-nine state files to install at the CATI facilities. (Hawaii cases were called in Jeffersonville because of the unique nature of their products.) These files consisted of cases not received in the 1992 Census of Agriculture universe which were preidentified as a "large farm."

Each state file was then processed in directory assistance (DA) subunits at the CATI facilities to obtain missing telephone numbers and correct inconsistent area code/telephone number combinations. At the beginning of the CATI process, the cases which did not yield a "good" telephone number from the DA subunit were left in the state input file for appropriate output coding and to facilitate accounting of all cases in the output file. However, this procedure was modified during the first wave by assigning these cases the appropriate coding upon file installation and omitting them from the calling que.

In addition to correcting the state files for missing telephone numbers or inconsistent area code/telephone number combinations, the files were updated to reflect mail returns received after the creation of each preliminary state file. EPD produced a file of satisfied Census File Numbers or CFNs (i.e., mail-ins, other resolved). They updated this "alert" file daily for use in amending the respective state CATI file by removing satisfied cases from the calling que. This amendment process was conducted one day prior to interviewing and continued daily until state closeout.

C. CATI Processing

Large Farm CATI Schedule for 1993

In order for the entire large farm CATI follow-up process to run efficiently, AGR coordinated with EPD, FLD and SSD to develop a schedule that notified each division of their timely interaction within the process. Attachment E is the schedule showing state workloads and respective dates for each step of processing. The schedule was broken down into 5 waves with approximately 10 states in each wave. States were listed in priority order according to other AGR processing dates. FLD, together with AGR, divided the states between the two telephone centers (Tucson and Hagerstown).

AGR decided to have one state file installed at each site to test the system before installing all of wave 1--Delaware was installed at Hagerstown and Oregon was installed at Tucson. Interviewing began at Hagerstown and Tucson on February 22 and March 8, respectively. As interviewing progressed, FLD installed the remaining state files in both waves 1 and 2 to allow a backlog of available states to call. As the CATI sites ran low on available cases, CATI site managers notified FLD and AGR to approve installation of other states/waves. In May, for instance, Tucson requested more cases in the Pacific time zone to accommodate interviewers who worked late. Consequently, California and Alaska became wave 3A. Wave 4 was also divided into two waves--4 and 4A.

Approximately five weeks were allotted for the interviewing process. As the CATI interviewing progressed, it became necessary to extend some states' closeout dates so that the interviewing process for other states started later than originally scheduled. When this was done, the dates for EPD and Secondary Source Unit (SSU) processing were changed accordingly.

CATI interviewing stopped one day prior to CATI closeout to allow for instrument/output file manipulation. The output files were sent to EPD for reformatting which required three days before sending the files to SSU in Jeffersonville. The

SSU needed about 13 days to resolve these cases, i.e., determine whether or not they were in the scope of the agriculture census. All in-scope cases were then merged with general census processing while out-of-scope cases were coded as such and no further processing was needed.

This schedule was instrumental in keeping all aspects of the processing on track and meeting the goal to complete the entire CATI process by October 1. The CATI sites closed out all states by September 12; and, EPD and SSU processing were completed by September 22.

Field Division Support

During the CATI operations, feedback on problems related to the interviewing process, subject matter, or operational/system efficiency, was relayed from FLD to AGR by the Field Division liaison and resolved, in most cases, via the electronic mail system or telephone. For example, in the beginning of the operation, interviewers questioned whether crops grown in years prior to 1992 but sold in 1992 should be included as 1992 sales. The FLD liaison referred this concern to AGR. AGR personnel informed the FLD liaison that these sales should be included in 1992 sales totals. This was later reiterated in a "briefing note" or bulletin and distributed to supervisors at both CATI sites.

As with any problem posed to AGR, after finding a solution, AGR periodically prepared a briefing note to document/clarify changes to the interviewing or operating process. These changes were discussed at the pre-shift meetings at both CATI sites to keep the interviewers up-to-date on the CATI process. Attachment F is an example of a briefing note.

These notes were essential for transferring information between AGR and FLD at headquarters as well as supervisors and interviewers at the CATI sites. They were also helpful in accounting and documenting each problem's resolution.

Subject Matter Support

To keep members of AGR abreast of CATI status, a "CATI Newsletter" was developed and sent out about every six weeks from March 1992 to January 1993. The newsletter was written by AGR committee members and was distributed to Agriculture Division's Chief, Assistant Division Chiefs, and Branch Chiefs. These newsletters contained information such as status of specifications development; current CATI issues; and, schedules for CATI testing, training, and production. In response to these newsletters, division personnel were able to comment on the CATI process at hand.

Also, AGR analysts made several site visits to the telephone facilities throughout the AGR CATI operation--from initial testing to closeout--to observe the process, monitor interviews, answer technical questions, collect improvement ideas from facility personnel, as well as identify and resolve any instrument and/or processing problems. For example, prior to live interviewing, staff visited the facilities to monitor three tests and to conduct initial training on the Research Operation. Site visits were also scheduled at the start of interviewing at both the Hagerstown and Tucson sites. After this initial "getting started" stage, AGR scheduled site visits about once a month--usually at the start of a new wave of states or at closeout.

Monitoring

Monitoring was an important part of the CATI process. While site visits provided a method for AGR to monitor the flow of facility processing and handling of technical information, CATI management was responsible for monitoring the quality of the interviewing process. Also, AGR received system-generated reports that provided up-to-date workload status for monitoring CATI progress.

To monitor the quality of the interview, the CATI system contains a built-in network which allows supervisors or analysts to listen to an interview on the telephone while simultaneously viewing the computer screen to see how the information is recorded by the interviewer. The facilities maintain specific standards for such monitoring. During initial monitoring (the first three months interviewers are on the job), about 10% of an interviewer's interviewing time each month are monitored. After that, (systematic monitoring) supervisors monitor at least 2.5% of each interviewer's active interviewing time as well as any "special needs" monitoring. In half-hour monitoring sessions, supervisors were able to unobtrusively observe interviewers to identify their strengths and weaknesses. The supervisor would complete a monitoring report each time an interviewer was monitored and provide feedback to the interviewer.

FLD staff sent periodic monitoring reports to AGR via the electronic mail system. At each CATI site and for each wave of the survey, these reports showed: average number of interviewers; number of login hours; and number of monitoring sessions. In the early stages of interviewing, these reports indicated areas where monitoring was inadequate and notice was given to the facility to increase monitoring. At the end of the survey, a final report was received from FLD which included the overall monitoring rate. Attachment G is a copy of this final report.

Every day during processing, AGR received two system-generated reports reflecting the previous day's work for each active state--a Sample Status Report and an Interviewer Performance Report.

Attachment H is an example of a Sample Status Report. Each day AGR extracted and recorded pertinent totals from active state reports such as numbers of cases classified as in-scope, out-of-scope, and remaining active. Cumulative numbers were also extracted and recorded for the following case types: resolved; mail receipts; duplicates; secondary source; and, refused.

Attachment I is an example of an Interviewer Performance Report. This report showed each individual interviewer's call attempts categorized by outcome code and summarized all call attempts by interviewers excluding/including supervisors and managers. Each day AGR extracted from active state reports and recorded data for: number of interviewers used; login hours; and minutes of in-scope calls. At closeout, AGR received a cumulative interviewer's report for the entire period of the survey (i.e., Monday, March 22, 1993 - Sunday, May 2, 1993).

The extracted data from these reports was used in developing several spreadsheets, such as:

- CATI PROGRESS REPORT - A separate spreadsheet for each state (survey) showing the daily progress. For each state report, data is given for each active interviewing day plus state totals.

- LARGE FARM CATI WORKLOAD - A spreadsheet of all states at both telephone facilities by wave/state showing workload totals. The totals are shown for each telephone facility as well as for the U.S.

- CATI INTERVIEWER REPORT - A spreadsheet for each month showing active states (surveys) and giving the daily counts of interviewers working, login hours, and completed interviews for in-scope cases. These monthly reports helped to verify telephone company monthly charges.

Research Operation

When a respondent indicated that they received multiple forms under different names or CFNs for the same operation, the interview was handled as a "possible duplicate". The Research Operation was set up at both CATI facilities to review and/or verify the possible duplicate situations. This involved EPD support in acquiring access to various

agriculture databases for both CATI sites. In view of the security risk, CATI personnel were granted "read only" capability when accessing the AGR Research Menu network routines.

By accessing the "name & address" and "check-in" databases, trained CATI researchers were able to search for duplicate entries in the mail list and verify receipt of the duplicate report forms as needed. The training materials for this operation were developed by AGR analysts and required AGR personnel to conduct agriculture-specific training at both facilities due to FLD's unfamiliarity with this subject.

The benefits of this operation provided notable strides toward customer service as well as improved processing. In contrast to the spontaneous handling of possible duplicate situations in 1992, in 1987 these cases were referred to clerical reviewers after the conversation was ended; and, in cases where subsequent research indicated that an interview was still needed, the respondent was recontacted. The utility of the 1992 operation was invaluable since it resulted in the identification of about 12,000-18,000 duplicates on the AGR mail list and prevented callbacks to an estimated equal number of possible duplicate cases which were unverified by the researcher.

"Claims filed" cases, or situations in which the respondent claimed that they had already mailed their questionnaire, were not routed to the Research Operation since the alert file was updated daily and any mail returns were deleted from the calling que. In these cases, the interviewer prompted the respondent to complete the interview knowing that the form had not been received.

Other Production Processing

Several other details were handled during the CATI interviewing process. For example, specific procedures were established to process "send form" and "Title 13" requests. All other situations which were not predesignated in training were handled as "supervisor referrals."

"Send form" cases, in which the respondent would not agree to be interviewed by telephone but requested a form be sent for completion and mailing, were coded as such and systematically set for a callback 10 days later. The CATI supervisors would check the system daily to pick up all cases with send-form coding and refer them to AGR for form mailing. If the report form was received within 10 days, the CFN was automatically coded as resolved and deleted from the CATI calling que via the daily alert system. Otherwise, CATI interviewers would recontact the respondent to complete the

interview by telephone.

Some respondents requested a copy of the Census Law (Title 13 of the United States Code) that requires them to respond. Again, the interviewer was directed to relay this information to the supervisor who consequently notified AGR of these cases on a daily basis.

D. CATI Output/Results

The results of the CATI attempts were transmitted to Economic Programming Division (EPD) in the form of four files:

- Answer file - interviewer coding including respondent data and/or interviewer remarks for resolved cases;
- F7 file - interviewer remarks made during the interviewing process for resolved cases;
- History file - "snapshot" of installed cases showing each time accessed; and,
- Case Master file - system management information for each installed case.

Every interviewing day, an Answer and F7 file for each active state were transmitted to EPD for their subsequent reformatting. At each state's closeout, a cumulative version of all four files was transmitted to EPD. EPD was responsible for assuring the receipt of these files and subsequent processing for merging these files into the 1992 Census of Agriculture operations.

Daily Processing

The daily state answer files were the source of in-scope and out-of-scope records.

For in-scope records, besides the answers collected which referred to the farm unit, some data items were created based on interviewer responses to reflect:

- CATI processing codes,
- flag indicators of "zero" or "none" responses for specific items,
- summing of valid duplicate items,
- geographic changes, and,
- section indicators.

In addition, these in-scope records were used to update various AGR databases. Whenever there were verbatim responses made during the interview (i.e., "other" crops/livestock not

listed, state or county changes, etc.), they were retrieved and loaded onto the "notes" database for analyst review. A check-in status code indicating "CATI satisfied" was assigned to each record in the "check-in" database. Corrections, if any, in the name, address, and/or telephone number fields were carried to the "name & address" database.

Out-of-scope records identified in the daily answer files were updated in the check-in database by assigning the appropriate status code.

Confirmed refusal cases were also included in the daily answer files to make these cases available for the next processing step (Secondary Source Unit), and, to alleviate storage space in the respective state's active case file. However, in all states, the confirmed refusal cases were processed by EPD from the cumulative state answer file.

The daily F7 files, which consisted of auxiliary notes made by the interviewer during the interview, were reformatted and loaded onto the notes database to provide analysts with supplemental information for in-scope cases.

Closeout Processing

The closeout answer file was cumulative and provided a single source of confirmed refusals and other unresolved cases requiring SSU processing. Cases routed to SSU were identified by specific final code. Attachment J is a list of these case types and the respective final codes.

Similar to the reformatting of in-scope records, interviewer comments, if any, and processing data items were created for each SSU case and forwarded to the Jeffersonville facility for assistance in determining the resolution.

Other Output Processing

The history file consisted of one record for each time a case was accessed. This file was transmitted to EPD at the closeout of each state survey. A separate history record was created for each case to show all the calls that were made as well as other actions where calls were not made (such as, when one "quits out" of a case before dialing). This file was routinely used by facility management to track the progression of a particular case. From this file, AGR manipulated information using SAS software to produce tables and graphs for management analysis. See Attachments K through N for examples of AGR charts produced from various history files. These state history files are in storage for potential studies at a later date.

A case master file was also transmitted to EPD at the closeout of each state survey. This file consisted of one record per case which contained essential information for that case for case-management purposes. Information in the case master file was always kept current--former values of variables were not retained. CATI facility personnel usually accessed this file as a convenient source for specific case details. For the purpose of analyzing the system, data from this file could be used for reviewing number of call attempts for specific outcomes, callback time preferences, and so on. Since AGR's plans for CATI evaluations are not complete at this time, these state case master files are being stored for possible later use.

IV. Summary of Agriculture CATI System Results

A. Final Case Accounting

Of the 152,815 cases installed for CATI resolution, 57,708 cases (37.8%) were enumerated as in the scope of the agriculture census while 15,148 (9.9%) were out-of-scope. Through research, 4,924 cases (3.2%) were determined to be duplicates of satisfied cases. After CATI file installation, 29,312 cases (19.2%) were omitted from the interviewing process as they were received in the mail. A total of 45,723 unresolved cases (29.9%), including 3,188 confirmed refusals, were routed to SSU for additional processing.

The results of the 1992 large farm CATI followup with comparative statistics from the 1987 nonCATI operation are shown in Attachment O, Table 1, Results of the 1992 and 1987 Delinquent Large Farm Followup.

When comparing the results of the 1987 clerical and 1992 CATI operations, one notices the higher in-scope rate (+5.2%) and slightly lower out-of-scope rate (-1.3%) in 1992. In both operations, a large number of cases were unresolved requiring SSU processing. These SSU cases (29.9% in 1992 and 31.7% in 1987) include respondent-contacted situations (i.e., insufficient partial interviews, refusals, callbacks) as well as noncontact cases (i.e., busy, no answer, never tried). Since the level of mail receipts is independent of operation type (nonCATI or CATI), these numbers only reflect the timely creation of the input call file prior to the telephoning process.

Another observation which is evident upon review is the difference in processing claims filed and possible duplicate situations. As mentioned previously, in 1992 these cases received spontaneous handling versus 1987's procedure to discontinue the telephone call, perform research, and, if

necessary, recontact the respondent. Claims filed cases resulted in prompting for an interview in 1992 (since the calling que was updated daily by the alert file). Possible duplicate cases were verified by the Research Operation and appropriate action was taken.

The different environments provided by the 1992 and 1987 systems make it difficult to conclude whether operation type is responsible for the resulting differences. Dissimilar management philosophies, physical locations/office set up, and experience level of interviewers are some factors which inhibit any attempt to denote collection methodology as the single reason for varying results.

B. Cost Analysis

The cost of the 1992 delinquent large farm followup amounted to almost $2.1 million or $13.56 per case. This amount included staffing for supervision/interviewing, system development (excluding AGR analysts), communications, equipment, and general administrative support.

Attachment P, Table 2, shows data for comparing the cost of the 1992 and 1987 delinquent large farm follow-up operations. When the 1987 operation cost is adjusted to show 1993 dollars, the 1992 CATI system cost savings is $2.17 or 13.8% less per case than the 1987 nonCATI operation. The overall cost of the 1992 CATI followup shows an approximate increase in expenses of $146,000 or 7.6% more than the 1987 nonCATI followup.

It would be misleading to use these statistics to make any conclusions concerning cost savings between the two operations. The expenses shown in Table 2 are actual charges made to the respective projects. Costs for general staff time (AGR and other), questionnaire design, use of previously procured hardware, and communication expenses covered by blanket costs are some of the factors which need to be addressed in making a system comparison. In addition, there are operational differences (i.e., quality control, supervisor/interviewer ratio) and post-interview processing (i.e., data entry needs, file reformatting) which should be addressed to gauge the benefits of each operation's yield versus cost.

V. Evaluation of the Agriculture CATI System

Reactions to the CATI system are necessary for the proper assessment of the operation. They provide valuable input towards building future CATI systems. To develop an improved efficient data collection system, AGR requested feedback from various persons

involved with the 1992 CATI operation.

A. Feedback from CATI Committee/Site Staff

After interviewing concluded, the CATI Committee was asked to submit any ideas or comments on the entire CATI developmental and operational process. In general, the suggestions consisted of need for: 1) more lead time for system development; 2) additional in-depth testing; 3) periodic retraining for interviewers during the process; 4) tracking of file transmissions; and, 5) prompt downloading of closed out state files.

In conjunction with AGR, FLD held five debriefing sessions at both CATI sites. These open forum meetings were conducted by the FLD liaison with AGR personnel present to answer questions and monitor delivery of the presentation.

To gather feedback for all three AGR CATI projects (i.e., large farm, nonresponse, low response counties), general questions as well as separate questions pertaining to the specific project were asked. The time set for these sessions was such that the majority of the states/surveys were closed out or approaching close-out. Large farm interviewing officially ended on September 12. Nonresponse and low response county projects closed out on August 20 and September 30 respectively.

The three debriefing sessions at the Tucson CATI facility were held on August 31 (10:00 AM and 2:00 PM) and September 1 (10:00 AM). Nine people attended the first morning session and eight people attended each of the other two sessions. In Hagerstown, two sessions were held on September 9 with nine attendees at 10:00 AM and seven attendees at 2:00 PM.

All personnel involved with the CATI operation (interviewers, supervisors, directory assistance callers, researchers, on-site analysts) were given an opportunity to complete the handout of questions (Attachment Q). However, a sample of these persons were invited to attend the debriefing sessions for extended discussions on the AGR CATI operation.

Overall, the sessions yielded many constructive suggestions for improving the system. The majority of problems cited were repeated in each of the sessions. A detailed summary of the responses and comments were consolidated into the 1992 Census of Agriculture Manual for reference. Such documentation will be useful to AGR in prospective CATI systems.

B. Evaluation Documentation

As traditionally conducted for the majority of census processing systems, AGR staff is currently documenting the CATI process in an evaluation paper. This paper will include: 1) system development and processing details; 2) CATI merge processing; and, 3) 1992 CATI versus 1987 nonCATI data review. Aspects of CATI recognized by CATI Committee members, CATI site staff, and AGR subject matter analysts as requiring "enhancements" or additional address are also noted in this synopsis for improvement of future AGR CATI systems.

Other CATI-associated studies in AGR are underway as well. A 1997 Census of Agriculture planning team has been formed to establish criteria for a CATI system for the next census based on review of CATI versus nonCATI data in the 1992 census. For example, this team has solicited feedback from AGR subject matter analysts to isolate problematic data items in CATI cases. Subsequently, the team will create and evaluate tallies of these suspect items for CATI and nonCATI records to verify their source.

C. CASIC Presentation

After the AGR CATI operations concluded in September 1993, AGR presented a brief overview to the Computer Assisted Survey Information Collection (CASIC) group. This group at the Census Bureau is dedicated to the automation of data collection through computerized technologies. In view of the fact that CATI is undergoing "redesign" in the Bureau to produce a centralized computer assisted data entry system across three sites (present facilities: Hagerstown and Tucson; new third site: Jeffersonville, IN.), and that the 1992 Census of Agriculture was the largest single project to use the CATI system, CASIC requested feedback from AGR to identify system strengths and weaknesses. Many of the comments submitted to CASIC are also included in this paper.

VI. Suggestions for Future Agriculture CATI Systems

Based on the success of the 1992 AGR CATI System, especially the processing phase, it is my opinion that CATI can be an efficient tool for follow-up data collection. Considering the 1992 incurred cost, beneficial automated features, paperless reporting, and installation of the Research Operation, this system is a viable method of enumeration. However, AGR's large scale encounter with this processing technology has lent itself to many improvement ideas.

As a result of this experience with CATI, the following are suggestions for enhancing prospective AGR CATI systems:

A. Developmental Recommendations

- Learn the CATI system to become aware of its potential. Sponsors should be knowledgeable of the capabilities to maximize its use and deal with system shortcomings properly for other processing coordination.

- Allow adequate time for testing. If possible, mock respondents not associated with the survey should be used in the tests.

B. Processing Recommendations

- Search for continual improvement ideas during processing. For example, review of case management files could possibly show significant "trends" toward specific callback times. In this case, interviewer scheduling should be adjusted accordingly.

- Acknowledge retraining needs. Even though AGR training was perceived to be complete, several interviewers indicated in the debriefing sessions that they had additional questions and/or situations requiring reverification after initial interviewing.

- Acquire on-site support staff for communications problems. Operations can be severely hampered awaiting personnel from headquarters to address problems.

C. Systems Recommendations

- Allow for customized output answer files. Files with standardized CATI output pose inefficient storage/handling situations when only a portion of the data is needed.

- Allow for customized progress reporting. System generated reports are primarily used by facility management and are difficult for sponsors to use.

- Develop interviewer paths for various situations. Allow for regional crops/livestock, basic information collection for reluctant respondents, and easy-skip paths for respondents indicating involvement with few commodities.

- Permit sponsor "read only" access to CATI management system. Since full access can be a security problem "read only" capabilities would allow the sponsor to make spontaneous status checks, track specific case

responses, etc. without FLD intervention.

- Automate system file transmissions. Routing of input and output files was manually monitored requiring much attention for operations consisting of numerous files/surveys.

- Explore hardware/software for CATI enhancements such as automated dialing to prevent misdialing, instrument software to allow for word processing capabilities, etc.

VII. Conclusion

Due to the use of the CATI system, improvements in this census' followup of delinquent large farms are apparent in the production phase regarding more interactive features and in eliminating subsequent processes.

Recognizing the long-term goals of the Bureau to use the latest technologies for more efficient and productive systems, CATI has offered several improved features over the 1987 Census of Agriculture clerical follow-up unit. Most notable of these are the systematic call scheduler, management research capabilities, and automated status report generation. In addition, other attributes yielded by CATI usage are savings from eliminating a separate data keying phase, omitting the need for paper questionnaires and their handling/storage, and providing better customer service through spontaneous research of possible duplicate cases requiring fewer recontacts.

Comparing the CATI to nonCATI follow-up operations is difficult. As mentioned in Section IV, part B, Cost Analysis, incurred costs are an unfair indicator of cost savings between 1992 CATI and 1987 nonCATI since savings and production efficiencies are not considered. Also, to conclude that any data disparity is a result of the collection methodology would be a misinterpretation in view of operational differences. Because of these factors which make comparison of CATI versus nonCATI systems very complex, AGR will be cautious in drawing conclusions--other than the realized progress achieved in process automation and better service to the respondent.

Aside from the deficiencies in comparing CATI and nonCATI operations for conclusive statements, AGR plans to review CATI-originated data with nonCATI data in the 1992 Census of Agriculture by examining the level of edited and/or imputed statistics. Rather than conclude that the operation source is the cause of any differences, the 1997 Census of Agriculture CASIC planning team will use these comparisons to detect dissimilarities in the data capture (i.e., wording and/or sequence of questions).

In summary, the 1992 Census of Agriculture CATI System was successful in terms of completing the follow-up phase in an efficient and timely manner. The automated features including the customized Research Operation provided many enhancements over the 1987 clerical operation. AGR hopes to build upon their experience with this initial system and use CATI in the 1997 Census.

REFERENCES

Ferrari, Pamela W. "An Evaluation of Computer-Assisted Telephone Interviewing Used During the 1982 Census of Agriculture," Research Paper, Agriculture Division, January 1986.

Mon, Jeanette K. and King, Elizabeth A. "Evaluation of the 1992 Census of Agriculture CATI System," in 1992 Census of Agriculture Manual, Project Manual MC, Chapter 27, Subchapter A, Document 2, Draft version, April 1994.

LIST OF ATTACHMENTS

A. Agriculture Division Large Farm CATI Flowchart

B. CATI Committee Schedule

C. 1992 Agriculture Census CATI Specifications

D. Schedule of Training

E. Large Farm CATI Schedule for 1993

F. Large Farm Followup Briefing Note (sample)

G. Monitoring Report as of 9/25/93

H. Sample Status Report (sample)

I. Interviewer Performance Report (sample)

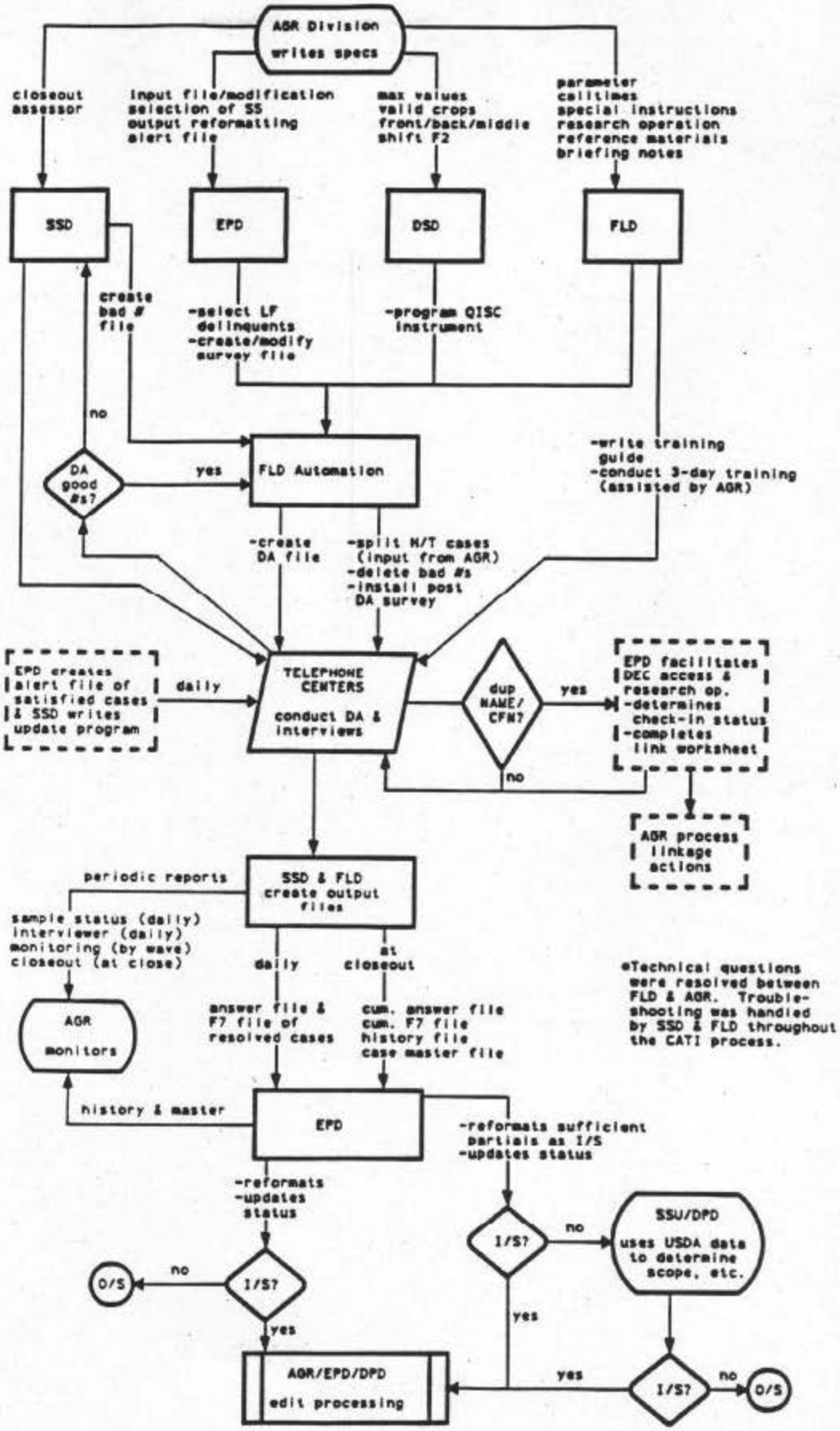
J. List of SSU Designated Case Type by Final Code

K-N. Agriculture charts/graphs from history record data

O. Table 1, Results of the 1992 & 1987 Delinquent Large Farm Followup

P. Table 2, Cost Comparison of the 1992 and 1987 Large Farm Followup

Q. Telephone Center Debriefing Handout


AGR Division
writes specs
closeout assessor
input file/modification selection of SS output reformatting alert file
max values valid crops front/back/middle shift F2
parameter calltimes special instructions research operation reference materials briefing notes
SSD
EPD
DSD
FLD
create bad # file
-select LF delinquents
-create/modify survey file
-program QISC instrument
no
DA good #s?
yes
FLD Automation
-write training guide
-conduct 3-day training (assisted by AGR)
-create DA file
-split H/T cases (input from AGR)
-delete bad #s
-install post DA survey
EPD creates alert file of satisfied cases & SSD writes update program
daily
TELEPHONE CENTERS
conduct DA & interviews
dup NAME/ CFN?
yes
no
EPD facilitates DEC access & research op.
-determines check-in status
-completes link worksheet
AGR process linkage actions
periodic reports
SSD & FLD create output files
sample status (daily)
interviewer (daily)
monitoring (by wave)
closeout (at close)
daily
at closeout
AGR monitors
answer file & F7 file of resolved cases
cum. answer file
cum. F7 file
history file
case master file
•Technical questions were resolved between FLD & AGR. Trouble-shooting was handled by SSD & FLD throughout the CATI process.
history & master
EPD
-reformats sufficient partials as I/S
-updates status
-reformats
-updates status
I/S?
no
SSU/DPD
uses USDA data to determine scope, etc.
O/S
no
I/S?
yes
yes
AGR/EPD/DPD
edit processing
yes
I/S?
no
O/S

Revised 3/2/92

CATI COMMITTEE SCHEDULE

Activity	Start Date	Complete Date	Responsible Person/Div
1. QISC Specs	9/91	3/92	Battle/King Mon
Modifications	11/91	8/92	DSD/AGR
2. Testing			
a. QISC	3/92	8/92	Battle/King/ Mon/DSD
b. Transmission to/from EPD	6/92	1/93	FLD/SSD/AGR/ EPD
c. Complete system (Pretest)	8/92	8/92	AGR/FLD/SSD/ DSD
d. State test	2/93	2/93	AGR/FLD/SSD/ EPD/DSD
3. Other Specifications	2/92	12/92	AGR/FLD/SSD/ EPD/DSD
4. Training Package/ Personnel	6/92	2/93	AGR/FLD
5. Implementation of System	3/93	8/93	FLD/AGR/SSD/ EPD
6. Evaluation	9/93	5/94	Committee

Attachment C

1992 AGRICULTURE CENSUS CATI SPECIFICATIONS

Manual Number	Title (explanation)
92EAG-A-MC-10-E-04	CATI Specs for Large Delinquent Telephone Followup (QISC programming specs for middle of interview--sections 1-31 of report form--designating outcome codes)
92EAG-A-MC-10-E-06	CATI Input File Specs (components and layout of the input file)
92EAG-A-MC-10-E-08	Valid Crop Listing for CATI QISC Specs (listing valid crops for each state)
92EAG-A-MC-10-E-10	CATI Assessor Specs (assignment of agendum and final codes which designate the next action)
92EAG-A-MC-10-E-12	CATI Front/Back for Delinquent Large Farm Followup (QISC programming specs for introduction and closing portions of the interview designating outcome codes)
92EAG-A-MC-10-E-13	CATI Test File Specs (input file specs for testing)
92EAG-A-MC-10-E-14	CATI Closeout Specs (assignment of final, outcome and agendum codes which designate whether sufficient partial or secondary source processing)
92EAG-A-MC-10-E-15	CATI Parameter Specs (a consolidation of several specs for a file used to pass individual survey info to the CATI system and specifies other file management requirements)
92EAG-A-MC-10-E-19	CATI Responses to Commonly Asked Questions (shift F2 option available to interviewers during interview)
92EAG-A-MC-10-E-22	Research Operation for Large Farm CATI (instructions for accessing AGR database to verify "claims filed" or "duplicate" situations)
92EAG-A-MC-10-E-23	Special Instructions for CATI Supervisors (supervisor duties regarding the research operation and certain agendums, i.e., refusals, language barrier, etc.)

92EAG-A-MC-10-E-24	CATI Output Reformatting Specs (info concerning the raw CATI answer file and needed reformatting/data manipulation)
92EAG-A-MC-10-E-30	CATI Calltimes Specs (setting times and days for making calls)
92EAG-A-MC-10-E-31	Large Farm CATI Schedule for 1993 (dates set for each step of processing)
92EAG-A-MC-10-E-38	Selection of Secondary Source Cases (specs for creating SSU file)
92EAG-A-MC-10-E-40	Change in CATI Interviewing Process for Acres (K787)
92EAG-A-MC-10-E-41	CATI Workbook for Training (interviewer training)
92EAG-A-MC-10-E-42	CATI Self Study for Interviewers (part of interviewer training)
92EAG-A-MC-10-E-43	CATI Final Review Exercise (part of interviewer training)
92EAG-A-MC-10-E-44	Large Farm CATI Briefing Notes (notes to supervisors/interviewers instructing how to handle particular situations)
92EAG-A-MC-10-E-45	Large Farm CATI Newsletters (information to keep AGR staff abreast of what was happening regarding AGR CATI)

Schedule of Training

Day 1

Chapter	Topic	Estimated Length	Total Elapsed Time
Chapter A	Introduction to the 1992 Census of Agriculture	1/2 hr.	1/2 hr
Chapter B.	Using the Agriculture Census Reference Materials	1 hr.	1 1/2 hrs
	Break	1/4 hr.	1 3/4 hrs
Chapter C	Farm Operations	1 1/2 hrs.	3 1/4 hrs
	Break	1/4 hr.	3 1/2 hrs
Chapter D	Walk-through Training Interview	3/4 hrs.	4 1/4 hrs
Chapter E	Concepts and Procedures	3/4 hrs.	5 hrs
Chapter F	Two Paired Practice Interviews	1 1/4 hrs.	6 1/4 hrs

Day 2

Chapter G	Another Walk-through Interview	1 1/4 hrs.	1 1/4 hrs
	Break	1/4 hr.	1 1/2 hrs

Day 2 (continued)

Chapter	Topic	Estimated Length	Total Elapsed Time
Chapter H	Additional Concepts and Procedures	3/4 hr.	2 1/4 hrs.
	Break	1/4 hr.	2 1/2 hrs.
Chapter I	Two Paired Practice Interviews	1 1/4 hrs.	3 3/4 hrs.
	Break	1/4 hr.	4 hrs.
Chapter J	Final Review Exercise	3/4 hr.	4 3/4 hrs.

Attachment E

Revised 10/5/93

LARGE FARM CATI SCHEDULE FOR 1993

WAVE	STATE	ESTIMATED MAXIMUM WORKLOAD	ACTUAL WORKLOAD	ACTUAL DATES FOR: CREATE FILE	DIRY ASST.	BEGIN INTVIEW	END INTVIEW	CATI CLOSEOUT	EPD PROCESS	SECOND. SOURCE	DATA KEY'G CLOSEOUT
1-H	DE (51)	726	1,147	2/4	2/8	2/22	3/27	3/28	5/4	5/7	5/19
1-H	IN (32)	4,001	3,982	2/25	2/26	3/2	4/17	4/18	5/5	5/7	5/18
1-T	OR (92)	3,482	3,986	2/25	3/1	3/8	4/10	4/11	5/21	5/24	6/3
1-T	WI (35)	3,152	3,028	3/4	3/8	3/16	4/17	4/18	5/14	5/14	5/28
1-T	WA (91)	3,528	3,528	3/4	3/8	3/23	5/1	5/2	5/27	5/27	6/10
1-T	IA (42)	5,966	5,796	3/4	3/8	3/23	5/1	5/2	5/18	5/19	6/1
1-T	MO (43)	4,789	5,251	3/4	3/8	3/30	5/13	5/14	5/25	5/27	6/8
1-H	OH (31)	3,562	3,475	3/4	3/8	3/18	5/1	5/2	5/28	6/7	6/15
1-H	MD (52)	1,376	1,377	3/4	3/8	3/18	5/1	5/2	6/9	6/10	6/24
1-H	WV (55)	825	857	3/4	3/8	3/18	4/17	4/18	6/10	6/11	6/25
2-H	VA (54)	3,018	3,007	3/4	3/8	3/22	5/1	5/2	6/11	6/15	6/28
2-T	KS (47)	3,394	3,282	3/4	3/8	4/2	5/8	5/9	6/18	6/21	7/7
2-T	IL (33)	4,158	3,830	3/4	3/8	4/5	5/8	5/9	6/25	6/28	7/14
2-H	MI (34)	3,069	3,282	3/4	3/8	3/22	5/1	5/2	7/5	7/8	7/26
2-T	WY (83)	2,352	2,808	3/4	3/8	4/7	5/8	5/9	7/18	7/21	8/5
2-H	ME (11)	718	830	3/4	3/8	3/29	5/1	5/2	7/11	7/14	8/2
2-H	NH (12)	234	243	3/4	3/8	3/29	4/17	4/18	7/11	7/14	8/2
2-H	VT (13)	1,030	1,217	3/4	3/8	3/29	5/1	5/2	7/12	7/15	8/2
2-H	CT (16)	427	534	3/4	3/8	3/29	5/1	5/2	7/12	7/15	8/2
2-H	MA (14)	597	685	3/4	3/8	3/29	5/1	5/2	7/17	7/20	8/2
2-H	RI (15)	84	113	3/4	3/8	3/29	4/17	4/18	7/17	7/20	8/2
3-T	CO (84)	2,895	2,249	4/8	4/12	5/3	6/10	6/11	7/20	7/23	8/13
3-T	ID (82)	2,790	2,406	4/8	4/12	4/22	6/1	6/2	7/24	7/27	8/18
3-H	NJ (22)	1,800	1,805	4/8	4/12	4/22	5/22	5/23	7/24	7/27	8/18
3-H	PA (23)	6,084	4,940	4/8	4/12	4/22	6/5	6/6	7/31	8/3	8/20
3-H	NY (21)	4,550	4,053	4/8	4/12	4/22	6/5	6/6	7/28	7/29	8/20
3-H	KY (61)	3,672	3,472	4/8	4/12	4/22	6/5	6/6	7/27	7/30	8/24
3-H	TN (62)	3,414	3,407	4/8	4/12	4/22	6/5	6/6	7/29	8/3	8/30
3-T	UT (87)	2,404	2,391	4/8	4/12	5/1	6/1	6/2	7/30	8/5	9/5
3-T	NE (46)	5,861	5,064	4/8	4/12	4/19	6/14	6/15	8/2	8/5	8/2
3-T	MN (41)	8,247	7,515	4/8	4/12	4/21	6/24	6/25	8/3	8/12	9/20
3-T	MT (81)	4,480	3,950	4/8	4/12	5/10	6/24	6/25	8/4	8/13	9/24
3A-T	CA (93)	8,803	7,627	5/5	5/10	6/4	7/17	7/18	8/5	8/10	9/17
3A-T	AK (94)	825	789	5/5	5/10	6/4	7/17	7/18	9/7	9/10	12/10
4-H	TX (74)	9,311	8,022	6/3	6/4	6/15	8/14	8/15	8/17	8/25	10/5
4-H	NC (56)	4,018	3,105	6/3	6/4	6/15	7/17	7/18	8/13	8/17	10/22
4-H	FL (59)	6,735	6,396	6/3	6/4	6/15	7/17	7/18	8/25	9/2	11/24
4-T	NV (88)	800	525	6/3	6/15	6/21	7/17	7/18	8/4	8/12	10/5
4A-H	ND (44)	4,008	3,419	6/3	6/11	6/15	7/31	8/1	8/16	8/19	11/5
4A-H	SD (45)	3,787	3,480	6/3	6/11	6/15	7/31	8/1	8/18	8/31	11/5
4A-T	AR (71)	6,374	4,870	6/3	6/15	6/22	7/17	7/18	8/19	8/1	11/5
4A-T	AZ (86)	1,432	1,141	6/3	6/10	6/22	7/17	7/18	8/23	8/1	11/16
4A-T	NM (85)	3,040	2,578	6/3	6/10	6/22	7/17	7/18	8/24	9/7	12/1
5-T	AL (63)	3,805	2,717	7/12	7/14	7/20	8/11	8/12	8/14	8/20	12/3
5-H	MS (64)	2,987	2,241	7/12	7/26	8/2	8/11	8/12	8/13	8/15	12/2
5-T	OK (73)	5,770	4,822	7/12	7/14	7/20	8/11	8/12	8/16	8/22	12/3
5-T	LA (72)	3,213	2,727	7/12	7/14	7/19	8/11	8/12	8/13	8/15	12/8
5-H	SC (57)	1,934	1,717	7/12	7/26	8/2	8/11	8/12	8/14	8/17	12/21
5-H	GA (58)	4,841	4,291	7/12	7/26	8/2	8/11	8/12	8/18	8/21	12/5
	TOTAL	166,778	152,815								

H = Hagerstown

T = Tucson

*Each state was updated for mail receipts the night before interviewing began

1992 Census of Agriculture CATI System
Large Farm Followup Briefing Note #5 - March 16, 1993

Thank you for your continued efforts in our data collection phase! Please note the following points:

- If the respondent indicates he/she has grapes, please read "grape vines" instead of "grape trees" as shown on the screen. (This will be corrected ASAP.)

- If respondent indicates he has peppermint, select "mint for oil" in section 2 or 7. They are the same.

- Only call the Research Operator when there is an actual duplicate situation (duplicate forms received for one operation with DIFFERENT names and/or CFNs). Remember to read the second sentence on the >multiforms< screen and the >claimsfile< screen before accepting a response.

- When "READ LISTING" appears on the screen, read the entire list.

  When "READ, IF NECESSARY" appears, you may use information previously given in the interview to decide if you need to read the entire list. For instance, in S9 the letter "P" appears next to the crops produced, so you want to probe for those particular items and then ask "Any others?" (it may not be necessary to read the entire list). If in doubt, read the entire list.

- The Shift F1 option brings up the >info-ref< screen which gives you the name and full address of the respondent. This option is useful for the interviewer in filling out the top portion of the worksheet when research is unavailable.

  After getting into the middle of the interview (beginning with S1), the Shift F1 >info-ref< screen will also give you the option "C" to change the respondent.

- Any CFNs reported as duplicates on the >check< screen will be displayed on the >research< screen. If the CFN is displayed, MOVE THE CURSOR by pressing enter to get to where you indicate whether the dup was found. DO NOT type in anything else on the line where the cursor first appears, unless there was no CFN displayed and you need to enter a CFN found by the Research Operator.

- On the >hello-2< screen, if you enter "11" (deceased & sold farm) you go to the >intro-b< screen where you should enter "8" (sold farm). The next question asks if operated during any part of 1992--if yes, the interview continues; if no, the interview ends.

Below find the requested Ag monitoring data as of September 25:

Hagerstown:

	Ave. # Intvrs.	Login Hours	Mon. sessions	Mon. rate
Wave 1	88	5757	311	5.4
Wave 2	73	4524	236	5.2
Wave 3	146	9550	269	2.8
Wave 4 & 5	73	14424*	375	2.6
Wave 5	198	3538	204	5.8

* these login hours include an underestimated amount of 14.7 for state 74

Tucson:

	Ave. # Intvrs.	Login Hours	Mon. sessions	Mon. rate
Wave 1	107	9283	144	1.55
Wave 2	122	4176	71	1.7
Wave 3	153	12281	440	3.6
Wave 4	141	9094	335	3.7
Wave 5	93	7200	243	3.4

TTC has included AG Model Drop in their counts.

AG CENSUS USA

ACRLF57 CREATED: 09/06/93 01:52:05

SAMPLE STATUS REPORT

FOR: SUNDAY SEPTEMBER 5, 1993

REMAINING ACTIVE CASES BY TIME ZONE

		TOTAL	EASTERN	CENTRAL	ROCKY	PACIFIC	YUKON	HAWAII
NO TELEPHONE NUMBER	(1)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
NEW CASE	(0,2)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
UNREACHED CASE	(3)	1 0.02	1 0.02	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
DIRECTORY ASSISTANCE NEEDED	(4)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
CONTACT, RESP NOT REACHED	(6)	4 0.07	3 0.06	1 0.14	0 0.00	0 0.00	0 0.00	0 0.00
CONTACT, RESP REACHED	(7)	11 0.20	10 0.21	1 0.14	0 0.00	0 0.00	0 0.00	0 0.00
AWAITING MAILBACK	(8)	9 0.16	8 0.17	1 0.14	0 0.00	0 0.00	0 0.00	0 0.00
REFUSAL/NO PROGRESS	(9)	5 0.09	5 0.10	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
PARTIAL INTERVIEW	(14)	25 0.45	21 0.44	4 0.57	0 0.00	0 0.00	0 0.00	0 0.00
HOSTILE BREAK/OFF REFUSAL	(15)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
OTHER	(5,10-13,16-20)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
TOTAL/MARK	(0-20)	55 1.00	48 1.00	7 1.00	0 0.00	0 0.00	0 0.00	0 0.00

TODAY'S RESOLVED CASES BY TIME ZONE

		TOTAL	EASTERN	CENTRAL	ROCKY	PACIFIC	YUKON	HAWAII
COMPLETE (ALL ITEMS)	(1)	5 0.83	4 0.80	1 1.00	0 0.00	0 0.00	0 0.00	0 0.00
PARTIAL TYPE 1	(2)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
PARTIAL TYPE 2	(3)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
PARTIAL TYPE 3	(4)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
PARTIAL TYPE 4	(5)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
PARTIAL TYPE 5	(6)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
MAILED FORM/SAME CFM	(7)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
MAILED FORM/DUP CFM	(8)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
OTHER DELETION	(9)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
DECEASED/NO REFERRAL	(10)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
LANDLORD ONLY	(12)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
NONAGRICULTURE	(13)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
NONAGRICULTURE - 10 AC	(14)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
SOLD FARM	(15)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
SOLE OPERATOR MOVED/NO REFERRAL	(17)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
SOLE OPERATOR ILL/NO REFERRAL	(18)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
OTHER NONINTERVIEW	(19)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
UNCONVERTIBLE LANGUAGE BARRIER	(21)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
UNPUBLISHED NUMBER	(22)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
NO LISTING (LEARNED FROM DA)	(23)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
NO INITIAL NUMBER SUPPLIED	(24)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
CONFIRMED REFUSAL	(25)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
SEARCH CUTOFF	(26)	1 0.17	1 0.20	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
CLAIMS FILED/UNCONFIRMED	(35)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
MAIL RETURN UPDATE	(104)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
CLOSEOUT CUTOFF	(115-127)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
OTHER	(1,16,20,27-34,36-103,105-114, 128-200)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
TOTAL/FINAL	(1-200)	6 1.00	5 1.00	1 1.00	0 0.00	0 0.00	0 0.00	0 0.00

CUMULATIVE RESOLVED CASES BY TIME ZONE

	TOTAL	EASTERN	CENTRAL	ROCKY	PACIFIC	YUKON	HAWAII

COMPLETE (ALL ITEMS)	(1)	818 0.49	781 0.54	37 0.18	0 0.00	0 0.00	0 0.00	0 0.00
PARTIAL TYPE 1	(2)	9 0.01	9 0.01	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
PARTIAL TYPE 2	(3)	2 0.00	2 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
PARTIAL TYPE 3	(4)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
PARTIAL TYPE 4	(5)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
PARTIAL TYPE 5	(6)	103 0.06	97 0.07	6 0.03	0 0.00	0 0.00	0 0.00	0 0.00
MAILED FORM/SAME CFN	(7)	65 0.04	61 0.04	4 0.02	0 0.00	0 0.00	0 0.00	0 0.00
MAILED FORM/DUP CFN	(8)	79 0.05	78 0.05	1 0.00	0 0.00	0 0.00	0 0.00	0 0.00
OTHER DELETION	(9)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
DECEASED/NO REFERRAL	(10)	3 0.00	3 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
LANDLORD ONLY	(12)	30 0.02	26 0.02	4 0.02	0 0.00	0 0.00	0 0.00	0 0.00
NONAGRICULTURE	(13)	73 0.04	71 0.05	2 0.01	0 0.00	0 0.00	0 0.00	0 0.00
NONAGRICULTURE = 10 AC	(14)	75 0.05	75 0.05	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
SOLD FARM	(15)	17 0.01	16 0.01	1 0.00	0 0.00	0 0.00	0 0.00	0 0.00
SOLE OPERATOR MOVED/NO REFERRAL	(17)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
SOLE OPERATOR ILL/NO REFERRAL	(18)	7 0.00	7 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
OTHER NONINTERVIEW	(19)	11 0.01	11 0.01	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
UNCONVERTIBLE LANGUAGE BARRIER	(21)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
UNPUBLISHED NUMBER	(22)	18 0.01	17 0.01	1 0.00	0 0.00	0 0.00	0 0.00	0 0.00
NO LISTING (LEARNED FROM BA)	(23)	261 0.16	110 0.08	151 0.72	0 0.00	0 0.00	0 0.00	0 0.00
NO INITIAL NUMBER SUPPLIED	(24)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
CONFIRMED REFUSAL	(25)	17 0.01	16 0.01	1 0.00	0 0.00	0 0.00	0 0.00	0 0.00
SEARCH CUTOFF	(26)	23 0.01	22 0.02	1 0.00	0 0.00	0 0.00	0 0.00	0 0.00
CLAIMS FILED/UNCONFIRMED	(35)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
MAIL RETURN UPDATE	(104)	51 0.03	50 0.03	1 0.00	0 0.00	0 0.00	0 0.00	0 0.00
CLOSEOUT CUTOFF	(115-127)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
OTHER(11,16,20,27-34,36-103,105-114,	128-200)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
TOTAL/FINAL	(1-200)	1662 1.00	1452 1.00	210 1.00	0 0.00	0 0.00	0 0.00	0 0.00

Attachment H (cont'd)

ACRLF57 CREATED: 09/06/93 01:52:05

SAMPLE STATUS REPORT

FOR: SUNDAY SEPTEMBER 5, 1993

CUMULATIVE RESOLVED CASES SPECIFIC BY FINAL

		TOTAL	EASTERN	CENTRAL	ROCKY	PACIFIC	YUKON	HAWAII
COMPLETION SERIES	(1-6)	932 0.56	889 0.61	43 0.20	0 0.00	0 0.00	0 0.00	0 0.00
FORM REC/OTH DEL	(7-9)	144 0.09	139 0.10	5 0.02	0 0.00	0 0.00	0 0.00	0 0.00
CONTACT MADE/NO DATA	(10,12-15,17-19)	216 0.13	209 0.14	7 0.03	0 0.00	0 0.00	0 0.00	0 0.00
UNREACHABLE	(21-24)	279 0.17	127 0.09	152 0.72	0 0.00	0 0.00	0 0.00	0 0.00
CONFIRMED REFUSAL	(25)	17 0.01	16 0.01	1 0.00	0 0.00	0 0.00	0 0.00	0 0.00
SEARCH CUTOFF	(26)	23 0.01	22 0.02	1 0.00	0 0.00	0 0.00	0 0.00	0 0.00
UNCONFIRMED CLAIMS FILED	(35)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
MAIL RETURN UPDATE	(104)	51 0.03	50 0.03	1 0.00	0 0.00	0 0.00	0 0.00	0 0.00
CB SCHEDULED/SUFF PARTIAL	(115)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
PREREFUSAL/HOB BREAKOFF	(116)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
NEEDS RESEARCH WORK	(117)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
CB SCHEDULED/PARTIAL INT	(118)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
TEMP UNAVAILABLE	(119)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
UNCOMPLETED CALL	(120)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
LANGUAGE BARRIER	(121)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
UNCONFIRMED CLAIMS FILED (HOLD)	(122)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
WILL FILE/SEND FORM	(123)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
NEV CONT/CONF NUMBER	(124)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
NEV CONT/UNCONF NUMBER	(125)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
ANSWER SERV/MACH. LEFT MESSAGE	(126)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
NEVER TRIED	(127)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
OTHER	(11,16,20,27-34,36-103,105-114, 128-200)	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
TOTAL/FINAL	(1-200)	1662 1.00	1452 1.00	210 1.00	0 0.00	0 0.00	0 0.00	0 0.00

Attachment H (cont'd)

INTERVIEWER PERFORMANCE REPORT

September 6, 1993 01:52:42

Sunday, September 5, 1993

Survey: AGRLF57

TOTAL ALL INTERVIEWERS
(Excluding supervisors/managers)

OUTCOME OF CALL ATTEMPTS		COUNT OF CASES	PERCENT OF TOTAL CASES	TOTAL TIME (MINUTES)	PERCENT OF TOTAL TIME	MEAN TIME (MINUTES)
A. Complete interviews	(01)	5	0.08	128	0.54	25.60
B. Partial interviews	(02-06)	0	0.00	0	0.00	0.00
C. SUBTOTAL, COMPLETES AND PARTIALS	(01-06)	5	0.08	128	0.54	25.60
D. Other call progress	(40)	0	0.00	0	0.00	0.00
E. Out of scope	(07-16)	0	0.00	0	0.00	0.00
F. HH refusals and hostile breakoffs	(30-34)	1	0.02	8	0.03	8.00
G. Other noncooperation	(35-39)	2	0.03	2	0.01	1.00
H. Other noninterviews	(17-29)	0	0.00	0	0.00	0.00
I. No contact	(80-89)	24	0.41	33	0.14	1.38
J. Other outcomes	(41-79)	27	0.46	67	0.28	2.48
K. TOTALS	(01-89)	59	1.00	238	1.00	4.03

L. Total login time (minutes)..... 369

M. Total login time (hours)..... 6.15

N. Call progress/login hours ((C+D)/M)..... 0.81

O. Ratio of household refusals and hostile breakoffs to call progress (F/(C+D))..... 0.20

P. Call progress rate ((C+D+E+F+G+H+J)/K)..... 0.59

Q. First refusal rate..... 0.13

R. Total refusal/hostile breakoff conversion attempts with contact... 0

RESULTS OF CONVERSION ATTEMPT:

S. Complete or Partial interview.. 0

T. Out-of-Scope..... 0

U. Second refusal..... 0

V. Other noninterview..... 0

W. Refusal/Hostile breakoff conversion rate ((S+T+V)/R)..... 0.00

Attachment I

INTERVIEWER PERFORMANCE REPORT

September 6, 1993 01:52:42

TOTAL ALL STAFF
(Including supervisors)

Sunday, September 5, 1993

Survey: AGRLF57

OUTCOME OF CALL ATTEMPTS		COUNT OF CASES	PERCENT OF TOTAL CASES	TOTAL TIME (MINUTES)	PERCENT OF TOTAL TIME	MEAN TIME (MINUTES)
A. Complete interviews	(01)	5	0.08	128	0.54	25.60
B. Partial interviews	(02-06)	0	0.00	0	0.00	0.00
C. SUBTOTAL, COMPLETES AND PARTIALS	(01-06)	5	0.08	128	0.54	25.60
D. Other call progress	(40)	0	0.00	0	0.00	0.00
E. Out of scope	(07-16)	0	0.00	0	0.00	0.00
F. HH refusals and hostile breakoffs	(30-34)	1	0.02	8	0.03	8.00
G. Other noncooperation	(35-39)	2	0.03	2	0.01	1.00
H. Other noninterviews	(17-29)	0	0.00	0	0.00	0.00
I. No contact	(80-89)	24	0.41	33	0.14	1.40
J. Other outcomes	(41-79)	27	0.46	67	0.28	2.48
K. TOTALS	(01-89)	59	1.00	238	1.00	4.04

L. Total login time (minutes)..... 369

M. Total login time (hours)..... 6.15

N. Call progress/login hours ((C+D)/M)..... 0.81

O. Ratio of household refusals and hostile breakoffs to call progress (F/(C+D))..... 0.20

P. Call progress rate ((C+D+E+F+G+H+J)/K)..... 0.59

Q. First refusal rate..... 0.13

R. Total refusal/hostile breakoff conversion attempts with contact... 0

RESULTS OF CONVERSION ATTEMPT:

S. Complete or Partial interview.. 0

T. Out-of-Scope..... 0

U. Second refusal..... 0

V. Other noninterview..... 0

W. Refusal/Hostile breakoff conversion rate ((S+T+V)/R)..... 0.00

Attachment I (cont'd)

INTERVIEWER/FACILITY SUMMARY PAGE

September 6, 1993 01:52:42

Sunday, September 5, 1993

Survey: AGRLF57

INTERVIEWER CODE	LOGIN TIME (MINUTES)	LOGIN TIME (HOURS)	COMPLETED INTERVIEWS	COMPLETES PER LOGIN HOURS	CALL PROGRESS PER LOGIN HOUR	RATIO OF HH REF TO CALL PROGRESS	REFUSAL CONVERSION RATE
FACILITY	369	6.15	5	0.81	0.81	0.20	0.00
ALL INTERVIEWERS	369	6.15	5	0.81	0.81	0.20	0.00
IM38	369	6.15	5	0.81	0.81	0.20	0.00

NUMBER OF INTERVIEWERS LISTED IN SUMMARY : 1

NUMBER OF SUPERVISORS/MANAGERS LISTED IN SUMMARY : 0

TOTAL NUMBER OF ALL USERS LISTED IN SUMMARY : 1

Attachment I (cont'd)

Attachment J

LIST OF SSU DESIGNATED CASE TYPE BY FINAL CODE

	TYPE OF CASE	FINAL CODE
1.	Insufficient partial	06
2.	Unconvertible language barrier	21
3.	Unpublished number	22
4.	No listing of telephone number	23
5.	No initial number supplied	24
6.	Confirmed refusal	25
7.	Search cutoff	26
8.	Callback scheduled, sufficient partial	115
9.	Prerefusal/hostile breakoff	116
10.	Needs research work	117
11.	Callback scheduled, insufficient partial	118
12.	Temporarily unavailable	119
13.	Uncompleted call, no contact on callback	120
14.	Language barrier	121
15.	Unconfirmed claims filed	122
16.	Will file, send form	123
17.	Never contacted, confirmed number	124
18.	Never contacted, unconfirmed number	125
19.	Answering service/machine, left message	126
20.	Never tried	127

1992 Census of Agriculture
Length of Call for Completed CATI Cases

OHIO (31)

MINUTES	1-15	16-30	31-45	46-60	>60	TOTAL
NUMBER OF CASES	118	719	348	115	7	1,307
	9.0%	55.0%	26.6%	8.8%	0.5%	

NORTH CAROLINA (56)

MINUTES	1-15	16-30	31-45	46-60	>60	TOTAL
NUMBER OF CASES	148	803	447	73	1	1,472
	10.1%	54.6%	30.4%	5.0%	0.1%	

NEW YORK (21)

MINUTES	1-15	16-30	31-45	46-60	>60	TOTAL
NUMBER OF CASES	135	1,048	451	100	9	1,743
	7.7%	60.1%	25.9%	5.7%	0.5%	

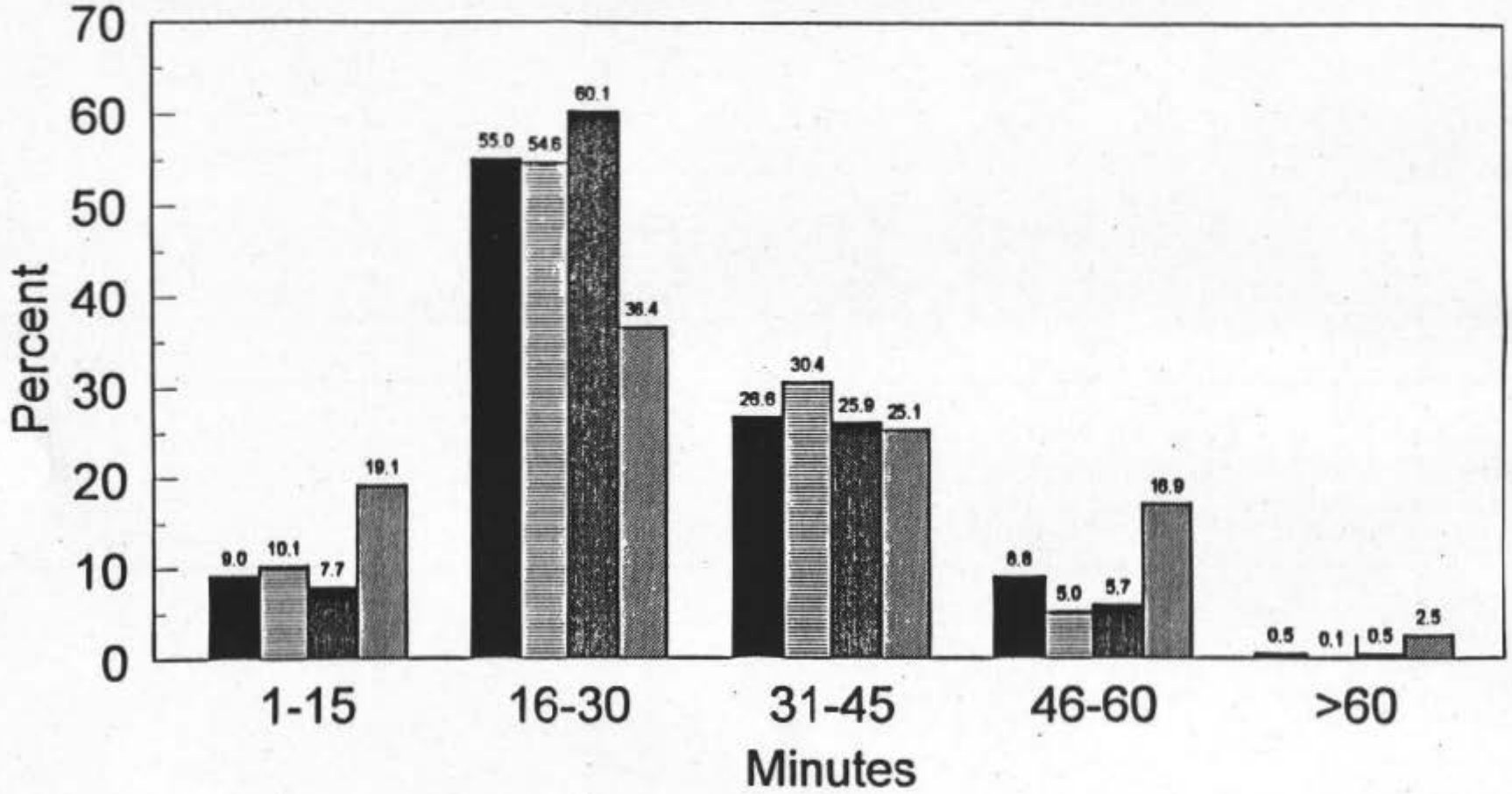
CALIFORNIA (93)

MINUTES	1-15	16-30	31-45	46-60	>60	TOTAL
NUMBER OF CASES	204	389	268	180	27	1,068
	19.1%	36.4%	25.1%	16.9%	2.5%	

AG CENSUS USA

1992 Census Of Agriculture CATI System

% of Completed Cases by Length of Call



■ Ohio ▒ North Carolina ▓ New York ▒ California

Source: US Bureau of the Census
Agriculture Division

(Rev. 6/94)

1992 Census of Agriculture
Respondent Time for Completed CATI Cases

OHIO (31)

6:00 am to 6:59 am	7:00 am to 7:59 am	8:00 am to 8:59 am	9:00 am to 9:59 am	10:00 am to 10:59 am	11:00 am to 11:59 am	12:00 to 12:59 pm	1:00 pm to 1:59 pm	2:00 pm to 2:59 pm	3:00 pm to 3:59 pm	4:00 pm to 4:59 pm	5:00 pm to 5:59 pm	6:00 pm to 6:59 pm	7:00 pm to 7:59 pm	8:00 pm to 8:59 pm	9:00 pm to 5:59 am
2	106	103	86	81	64	98	95	86	90	95	98	97	92	97	17
0.2%	8.1%	7.9%	6.6%	6.2%	4.9%	7.5%	7.3%	6.6%	6.9%	7.3%	7.5%	7.4%	7.0%	7.4%	1.3%

TOTAL = 1,307

NORTH CAROLINA (56)

6:00 am to 6:59 am	7:00 am to 7:59 am	8:00 am to 8:59 am	9:00 am to 9:59 am	10:00 am to 10:59 am	11:00 am to 11:59 am	12:00 to 12:59 pm	1:00 pm to 1:59 pm	2:00 pm to 2:59 pm	3:00 pm to 3:59 pm	4:00 pm to 4:59 pm	5:00 pm to 5:59 pm	6:00 pm to 6:59 pm	7:00 pm to 7:59 pm	8:00 pm to 8:59 pm	9:00 pm to 5:59 am
0	180	115	71	63	60	113	94	78	66	84	81	113	109	123	72
0	12.7%	8.1%	5.0%	4.4%	4.2%	7.9%	6.6%	5.5%	4.6%	5.9%	5.7%	7.9%	7.7%	8.6%	5.1%

TOTAL = 1,422

NEW YORK (21)

6:00 am to 6:59 am	7:00 am to 7:59 am	8:00 am to 8:59 am	9:00 am to 9:59 am	10:00 am to 10:59 am	11:00 am to 11:59 am	12:00 to 12:59 pm	1:00 pm to 1:59 pm	2:00 pm to 2:59 pm	3:00 pm to 3:59 pm	4:00 pm to 4:59 pm	5:00 pm to 5:59 pm	6:00 pm to 6:59 pm	7:00 pm to 7:59 pm	8:00 pm to 8:59 pm	9:00 pm to 5:59 am
0	175	139	160	164	175	135	132	95	103	139	92	94	117	108	14
0.0%	9.5%	7.5%	8.7%	8.9%	9.5%	7.3%	7.2%	5.2%	5.6%	7.5%	5.0%	5.1%	6.4%	5.9%	0.8%

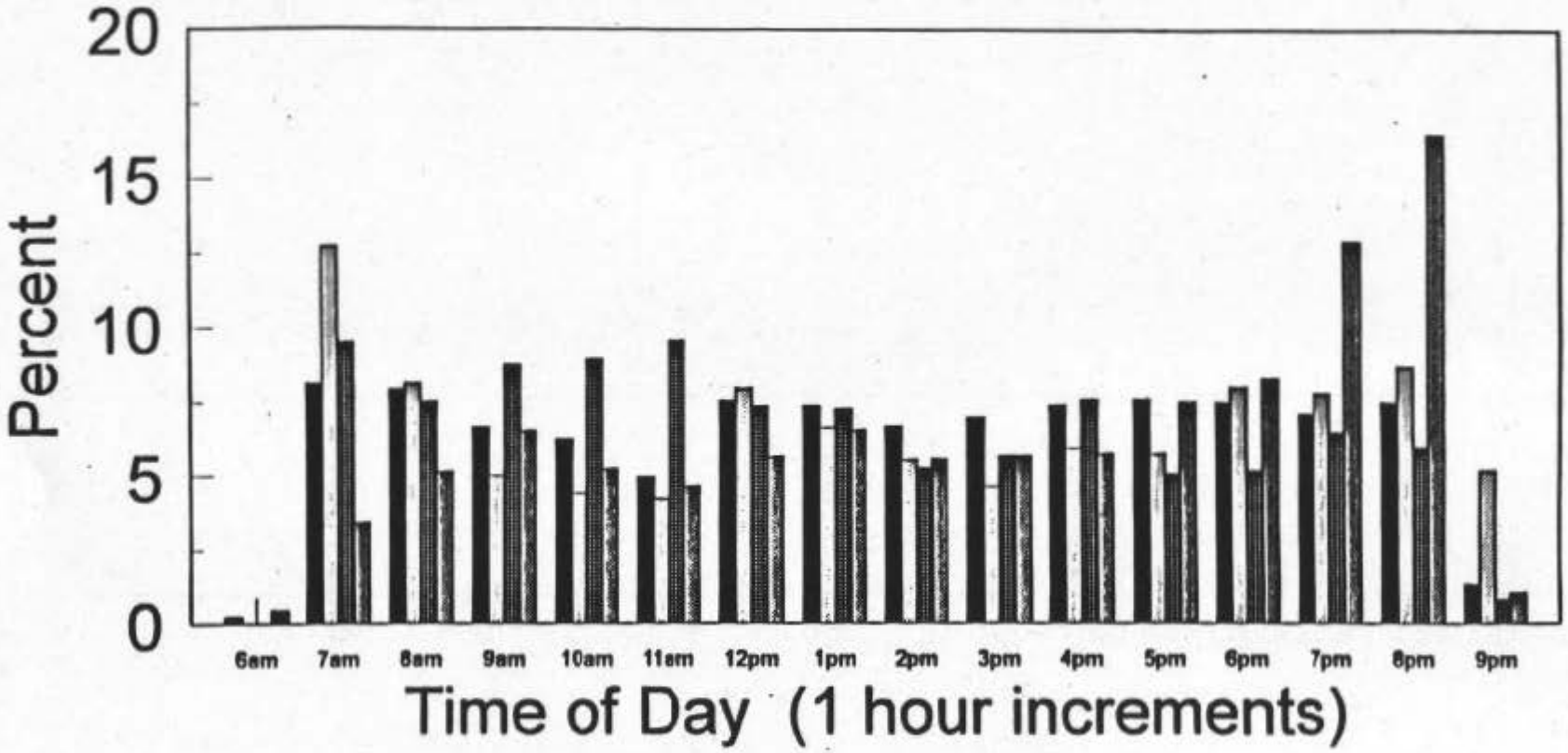
Total = 1,842

CALIFORNIA (93)

6:00 am to 6:59 am	7:00 am to 7:59 am	8:00 am to 8:59 am	9:00 am to 9:59 am	10:00 am to 10:59 am	11:00 am to 11:59 am	12:00 to 12:59 pm	1:00 pm to 1:59 pm	2:00 pm to 2:59 pm	3:00 pm to 3:59 pm	4:00 pm to 4:59 pm	5:00 pm to 5:59 pm	6:00 pm to 6:59 pm	7:00 pm to 7:59 pm	8:00 pm to 8:59 pm	9:00 pm to 5:59 am
8	72	109	140	112	99	121	140	118	119	123	159	175	274	352	21
0.4%	3.4%	5.1%	6.5%	5.2%	4.6%	5.6%	6.5%	5.5%	5.6%	5.7%	7.4%	8.2%	12.8%	16.4%	1.0%

1992 Census of Agriculture CATI System

% of Completed Cases by Respondent Time of Day



Ohio
North Carolina
New York
California

Source: U.S. Bureau of the Census
Agriculture Division

Attachment N
(Rev. 6/94)

Table 1

Results of the 1992 and 1987 Delinquent Large Farm Followup

CATEGORY	1992	1987
In-scope cases	57,708 (37.8%)	39,893 (32.6%)
Out-of-scope cases	15,148 (9.9%)	13,677 (11.2%)
Claims filed cases	NA[1]	20,030 (16.4%)
Duplicates	4,924 (3.2%)	NA[2]
Mail receipts	29,312 (19.2%)	10,000 (8.2%)
Secondary source	45,723 (29.9%)	38,800 (31.7%)
Total Workload	152,815	122,400

[1]Number of claims filed cases was not recorded in 1992. In these situations, if the respondent indicated mailing took place within the last week, a callback was systematically set for 7 days later. If mailing took place prior to the last week, the respondent was prompted to proceed with the interview. Any resulting refusals were sent to SSU.

[2]Number of duplicates was not recorded in 1987.

Table 2

Cost Comparison of the 1992 and 1987 Large Farm Followup

CATEGORY	1992	1987
Facility supervisory staff	$889,266	
Interviewing staff	$264,797	$626,000[1]
Benefits, other applied costs	$315,327	$323,000
Communications	$500,048	$122,000
Equipment, micro-computers	$49,982	$20,000
Facility administrative support, training, travel	$2,761	$158,000
Data Keying		$106,500
Other	$50,619	$210,795
TOTAL COST	$2,072,800	$1,566,295 $1,926,543[2]

FINAL COST PER CASE:	
1992	1987
$13.56	$12.79 $15.73[2]

[1]Includes supervisory staff

[2]inflated by 23% to reflect 1993 dollars

1992 CENSUS OF AGRICULTURE CATI SYSTEM

TELEPHONE CENTER DEBRIEFING HANDOUT

Name:_____

Position: (check all that apply)

- ☐ **Interviewer**
- ☐ **Research Operator**
- ☐ **Supervisor**
- ☐ **Directory Assistance**

Date started working on Ag CATI:

_____

AGENDA

I. Introduction

II. Debriefing on Training

III. Debriefing on the Interviewing Instrument

IV. Debriefing on Special Processing

V. Other

I. INTRODUCTION

II. DEBRIEFING ON TRAINING

Large Farm:

1. Did the training prepare you for conducting interviews?

2. Were AG concepts and definitions sufficiently covered during training?

3. Were the reference materials you received sufficient? Were they necessary? Did you use them?

4. Were the practice interviews helpful? Were there enough?

5. Do you have any ideas for improving training in the future?

6. Is there anything you would like to see covered in the training that was not included?

7. Were there topics in training that needed more or less time? If so, what were they?

Nonresponse:

1. Did you think the self-study was adequate for nonresponse training? If not, what other types of training would you suggest?

2. Did you feel that the nonresponse self-study prepared you for conducting nonresponse survey interviews adequately? Did you need more or less time?

3. Did the 4 practice interviews provide enough practice?

Low Response:

1. Was the purpose of the Low Response County follow-up surveys made clear to you?

2. Was the introductory briefing of this survey sufficient for your interviewing? If not, what other types of training/information would be helpful?

III. DEBRIEFING ON THE INTERVIEWING INSTRUMENT

Front of Instrument:

1. In the >review< screen, information about the operation (1987 acreage, value of sales, type of organization, and so on) was available for the interviewer.

   Did you find this information useful?

2. Did you find that the "front" screens accommodated for most situations?

3. Were the screens for "claims filed" or "multiforms" situations sufficient?

Large Farm/Low Response Middle of Instrument:

1. Are there any general screen changes that you would recommend for the next census' instrument?

2. Included in the instrument were "menu" screens in which selection was made from a listing.

   Were these screens difficult to collect responses? Any problems with these screens?

3. There were screens for verification of responses.

   Were there any problems with using these screens?

4. There were screens with indication of prior responses.

   Were these hints helpful?

5. For some "menu" screens, the "other" option was available for selection 3 times.

   Was this sufficient? Did you need more than 3 "other" selections at times?

6. Many of the basic screens presented a question followed by an "If Necessary" statement which offered more detail. These were to be used in case the respondent needed more explanation of the basic question.

   Were these "If Necessary" statements helpful? Did you need to read these often, sometimes, or rarely?

7. Were there any differences between the Low Response and Large Farm surveys? (Differences such as, amount of respondent cooperation, kinds of respondent reactions, wording in the instrument that did not apply, unique problems, etc.)

8. There were screens in which more than one response was requested.

   Are these types of screens preferable to asking for a single item per screen?

9. There were screens available for "help or further explanation."

   Did you use these screens often, sometimes, or rarely?

10. The "F7" key was available for additional note taking. Did you use this often, sometimes, or rarely?

11. Did you find the "Shift F2" Q & A function helpful? Did you use this often, sometimes, or rarely? Were other Q & A needed? If so, give some examples.

12. Did you find it necessary to use the "F1" key to back up often? Did this function work well when needed?

Nonresponse Middle of Instrument:

1. Were the skip patterns logical? If not, what would you suggest?

2. Was the instrument adequate in terms of collecting all pertinent informatin to areas covered?

3. Were there any specific types of agricultural production for which the instrument did not adequately provide questions and/or answers?

4. Were there any questions which were not clear to the respondents? What questions were they, and how could the question(s) be reworded?

5. Was the instrument sufficient in questioning "small" and/or borderline farms? If not, explain.

Back of Instrument:

1. Were the callback screens sufficient for making a callback?

2. Was the >inotes< screen sufficient?

3. Was there any confusion as to whether a callback was a "soft" or "hard" appointment? If so, explain.

IV. DEBRIEFING ON SPECIAL PROCESSING

1. Did you find it difficult to connect with a research operator? Was it often, seldom, or rarely that a research line was busy or unavailable?

2. Were the roles of the respondent, researcher, and interviewer clear whenever research was being conducted?

3. Was the procedure clear as to what was to be done if a fax number instead of a telephone number was identified?

4. Did you find it helpful to have a calculator for the data collection process?

5. Were the periodic "Briefing Notes" informative?

V. OTHER

QUESTIONS FOR SUPERVISORS

- Were you given adequate training for your job as supervisor? If no, what type of training would have been helpful?

- Do you have any other comments in regards to your supervisory function?

QUESTIONS FOR RESEARCH OPERATORS

- Were you given adequate training for your job as research operator? If no, what type of training would have been helpful?

- What was the most serious problem encountered?

- Do you have any suggestions for improving the computerized research operation or the research process?

- Do you have any other comments in regards to your research operator function?

QUESTIONS FOR DIRECTORY ASSISTANCE CALLERS

- Do you have any suggestions for improving your job in calling DA for telephone numbers? If yes, what are they?

- Do have any comments in regards to your job in calling DA for telephone numbers? If yes, what are they?

TABLE OF CONTENTS

		Page
I.	Introduction	1
II.	Background	1
III.	Overview of the Agriculture CATI System	2
	A. Agriculture Division's Use of CATI	2
	B. System Development	3
	C. CATI Processing	6
	D. CATI Output/Results	11
IV.	Summary of Agriculture CATI System Results	13
	A. Final Case Accounting	13
	B. Cost Analysis	14
V.	Evaluation of the Agriculture CATI System	14
	A. Feedback from CATI Committee/Site Staff	15
	B. Evaluation Documentation	16
	C. CASIC Presentation	16
VI.	Suggestions for Future Agriculture CATI Systems	16
	A. Developmental Recommendations	17
	B. Processing Recommendations	17
	C. Systems Recommendations	17
VII.	Conclusion	18
	References	19
	List of Attachments	20

DISCUSSION: WHAT CAN CAI LEARN FROM HCI?

Mick P. Couper
Joint Program in Survey Methodology

1. Introduction

This discussion will not focus on the papers presented in this session. These are two good papers that demonstrate the feasibility of computer-assisted interviewing (CAI) for large-scale data collection, both CAPI (computer-assisted personal interviewing) and CATI (computer-assisted telephone interviewing). Instead, this discussion will take to heart the "New Directions" part of the title of this seminar, and attempt to take a glimpse at the future of CAI or, more broadly, CASIC. In doing so, I will focus on certain aspects of computer-assisted interviewing, and particularly the question of the user interface.

Sixty years ago, writing about the industrial revolution, Lewis Mumford (1934: 6) wrote: "So far we have embraced the machine without fully understanding it...". I believe the same may be said of CAI at its present stage of development. We know that CAI "works", as these two papers ably demonstrate, and we are enthusiastically advocating the application of computer technology to virtually all areas of survey data collection. The CASIC movement is well-established in government, academic and private survey research organizations. But how much do we understand this new method of data collection, and its impact on the data collected, on the people who provide it, and on the people who collect and process it?

To quote further from Mumford, "In order to reconquer the machine and subdue it to human purposes, one must first understand it and assimilate it" (1934: 6). This is a view of the machine as a tool in the hands of the user. Rather than making the human conform to the machine, attention should be turned to the needs of the user. There are many areas of research in CASIC that remain unexplored (see Couper, Groves and Kosary (1989), Groves and Nicholls (1986) and Baker (1992) for some examples), and much work that needs to be done to optimize the use of such systems. I will focus on only one of these areas, namely the question of usability.

Whereas feasibility addresses the question "Can it be done?", usability focuses rather on "How best can it be done?". "Best" in this case should be defined in part from the perspective of the users of the system. It is in this area that I believe we have a great deal to learn from the field of human computer interaction research or HCI. Marchionini and Sibert (1991) define HCI research as being "concerned with the design of interfaces that allow easy and efficient use of computer systems." Hix and Hartson (1993) offer a less formal definition of usability: "If your computer were

a person, how long 'til you punch it in the nose?".

The Dippo et al. paper in this session talks at length about the "new methodology" of using cognitive psychological theories and methods in questionnaire development. A similar opportunity presents itself with the application of HCI research methods and findings to CAI. There is much we can learn about usability, both in terms of findings from existing research in other domains, and in terms of methods for usability testing and evaluation of user interfaces.

By usability is meant simply that the focus of our attention turns from the system to the user. This means person-centered design rather than system-centered design. At present, the capabilities and limitations of the hardware and software we use are driving the design of CAI systems. We are making the user adapt to idiosyncracies of the system rather than the other way around. The notion of "user-friendliness" or the subjective reaction of the user to the system, is only one component of usability. Shneiderman (1992: 18) defines usability in terms of the following five measurable components:

(a) time to learn
(b) speed of performance
(c) rate of errors by users
(d) retention over time
(e) subjective satisfaction

In the spirit of continuous quality improvement, if we focus on measurable aspects of usability, we will be able to demonstrate concrete improvements in the design, development and implementation of computer-assisted survey instruments. In doing so, we should achieve measurable gains in data quality, defined by Kalton (keynote address, this conference) to include not only accuracy, but also timeliness, cost effectiveness, relevance and accessibility.

2. Types of Users of CAI Systems

In promoting a user-centered view in CAI, we need to define who the users are. I have four sets of people in mind: (a) the programmer or instrument designer, (b) the interviewer, (c) the manager or supervisor, and (d) the end-user or analyst. To this list could be added a fifth set of users, the respondents[1]. Each of these users faces a different set of usability and other issues. Some of these concerns, and areas for further research and development, are as follows:

---

[1]I thank Judith Lessler for reminding me of this group of users.

2.1. Programmer/author/instrument designer

Part of the ambivalence about what to call the people who create the CAI instrument reflects uncertainty regarding the combination of people and skills needs for this work. Should non-programmers be able to create and test a CAI instrument? What is the role of questionnaire designers versus computer programmers in instrument design? What do we mean by "programmers"? Developing an understanding of the process of instrument design will facilitate identification of the optimal skill combinations required to enhance the product of the CAI design process. Some exploratory work has been done in CAI on the development of new software tools or application of existing tools to facilitate the instrument development process (see for example Baker, 1988; Balestrino, Fortunato and Montagna, 1992; Dibbs and Hale, 1993; Pierzchala, 1993). Another area that needs further development is in the tools for testing and debugging CAI instruments (see Connett, Mockovak and Uglow, 1994).

In part, the mix of skills required for these tasks may be dictated by the design of the CAI systems used. Some CAI systems may use more natural language interfaces, while others use cryptic code in authoring specifications. Should we be expecting this group of users to adapt to the complexities of the systems being used? A human factors perspective would argue not. The first generation CAI software was relatively unsophisticated in its design interface, but we should expect more of future CAI systems.

2.2. Interviewer

In some senses this is the most critical group of users for a number of reasons. The large number of interviewers and the diversity of their computer skills and knowledge relative to other groups of users, the minimal training they receive on the computer hardware and software they will use with little close supervision, and the potential they have to impact the data collection process (in terms of both costs and errors), all make it imperative to design systems to maximize interviewer efficiency and minimize errors. We will return to this group later.

2.3. Manager and/or supervisor

This set of users requires detailed information on the process of data collection, including cost and production data. While CAI can provide vast amounts of timely information that were previously not available in paper-and-pencil data collection, we have yet to devise methods to manage the information flow in ways that would facilitate the work of these users. This group has the potential for information overload unless such tools are provided.

2.4. End-user, analyst

The production of analytic data sets in a form that analysts can readily use, and in a way that they can understand how the data were collected is an important area of usability. This includes both the data itself and the metadata (codebook, variable labels, information on skips and edits, etc.). Is the analyst provided with a hard copy version of the questionnaire to review, or are all users expected to load the interviewing software on their system to look at a particular set of questions they may wish to analyze? Some of the information that may be needed by this group of users includes: (a) where did this question appear (what questions came before or after)? (b) which respondents were asked this question (skip patterns)? (c) what edit, range and consistency checks were built into this question? and (d) how was this variable created (recode, combination of multiple questions, etc.)?

These needs speak to the integration of the survey data collection process with the production of useful data sets. This view acknowledges that CAI systems are more than just a set of interviewing tools, and are (or should be) a fully integrated system of data collection, management and data preparation (see Kreighton, Matchett, and Landman, 1994).

2.5. Respondent

In interviewer-administered surveys respondents may have little direct contact with the computer, other than through the interviewer. However, in a variety of self-administered surveys using CAI (such as CASI, CSAQ, TDE, VRE, etc.; see Weeks (1992) for a review), respondents may interface with the system directly. These respondents may have had little or no training on the use of the system, may have limited prior experience with computers, and may not be highly motivated to participate in the survey. Thus, in addition to concerns about interacting with the computer, they may be uncertain about the nature of the interview task itself. For these reasons, the design of the interface is especially critical for this group of users.

All of these sets of users (and there are others) vary in their information needs, the tools they need to access or use this information, their computer skill levels and/or knowledge of the particular system being used, and so on. Much of our energy seems to have gone into the task of getting working CAI instruments up and running. We have expected the various groups of users to adapt to the idiosyncracies and shortcomings of the systems we currently have at our disposal. A more user-oriented approach would be to systematically determine what the needs of each set of users are and understand the nature of their work, then design systems that specifically meet those needs or that facilitate the completion of their tasks.

One of the critical lessons from HCI or human factors research is the importance of involving users in the design and evaluation process (see Galitz, 1993; Gould and Lewis, 1983; Norman, 1983). This cannot be emphasized enough. Users are a valuable resource that we have tended to ignore or pay only lip service to in CAI. We often proceed from the assumption that we know best, and design systems with little regard for those who will attempt to use them. As Powell (1990: 31) notes caustically, "Dumbo the elephant used his ears to fly. Use yours to listen to the users."

In the remainder of this discussion, I will focus my remarks on the second group of users, the interviewers. This is not because the other users are less important, but rather that some of the problems they face may be relatively intractable in the short run, whereas measurable improvements can be made to the instruments used by interviewers will relatively little investment.

3. Design Principles for CAI

Thus far, I have talked in the abstract about the need to pay attention to the human-computer interface, and of the importance of designing for usability. Note that usability is more than simply screen design, it is the entire system as experienced by the users. As Jagodzinski (cited in Davis and Bostrom, 1992) notes, to most users (and this would certainly include interviewers), the interface <u>is</u> the system. Usability considerations cannot be separated from other aspects of system design and development (see Gould and Lewis, 1983; Gould, 1988).

Because usability or "user friendliness" can be a quite nebulous concept, let me offer a set of design guidelines for CAI systems. These are adapted from a variety of sources in the HCI literature, including Hix and Hartson (1993), Galitz (1993), Mayhew (1992), Norman (1983), Powell (1990), Ravden and Johnson (1989), and Shneiderman (1992). Some of these are empirically-based principles, others are more prescriptive. Nonetheless, these are a set of desirable qualities of computer systems generally agreed upon in the field of HCI that may be applicable to CAI. This list may serve as a starting point to focus our attention on some of the issues that need to be addressed in terms of enhancing the usability of the systems we use for CASIC. A well-designed CAI system should exhibit the following qualities:

4.1. Functionality

The system should meet the needs and requirements of users when carrying out the tasks (Ravden and Johnson, 1989). Note that this is functionality from a usability perspective. It is not what the designer thinks the users should do, but rather what the user needs to do in order to complete the task correctly and efficiently. Furthermore, it is not enough that the <u>system</u> can do

X; the critical question is whether the user can do X with the system.

Functionality is a necessary but not sufficient condition for any CAI system. Shneiderman (1992: 10) notes that if the functionality of a system is inadequate, it does not matter how well the human interface is designed. The remaining guidelines or principles are essentially ways in which system functionality can be presented to the user to facilitate successful completion of the task.

4.2. Consistency

This refers to the look and feel of the system. At its simplest level, consistency refers to the placement of items on screens, including the use of fonts, upper or lower case, color, highlighting, etc., to distinguish between questions, interviewer instructions, response options, and so on. However, consistency should also include input modes, mapping of function keys and movement and navigation around the instrument.

Consistency can be viewed at a number of different levels:

(a) Consistency within a particular survey instrument. There are probably few who would disagree with this in principle, but I have seen a number of production CAI instruments where this is not achieved in practice.

(b) Consistency between the instrument and other interviewer tools (case management, transmission software, e-mail, etc.). We give interviewers a variety of tools to use, often without taking much effort to integrate them in a consistent fashion. Do the function keys assigned to operations in case management, for example, have consistent effects when used in the survey instrument? Many of the case management systems used by survey organizations (see Nicholls and Kindel, 1993) are written in-house, usually with little consideration of the CAI interface with which they will be used.

(c) Consistency across different surveys instruments within a particular organization. This is an area where organizational standards or guidelines in the authoring of CAI instruments would be beneficial (see Hunter, 1993). It appears that many programmers or authors have a particular style, which may be internally consistent within the instrument (or module) they develop, but differs from other survey instruments interviewers have used. An extension of this is interviewers actually using different CAI systems for different surveys.

(d) Consistency across organizations. Although I am not advocating that this be done, we ought to acknowledge that interviewers may work for multiple organizations using different hardware and software systems. There are no universally accepted design guidelines for CAI systems, and this may impact on the transferability of knowledge.

The first of these levels may be the most easy to implement, but

the other levels are no less important to consider in CAI.

Another component of consistency is predictability. System actions should be expected within the context of actions that are performed by the user (Galitz, 1993). In other words, if the interviewer does X, the system should always do Y. Thus, there is not only consistency in terms of what the user sees and does, but also consistency in terms of what the system does in response to user inputs.

Although consistency is probably the most universally endorsed principle, there are those who caution against its rigid application without consideration of other design principles. Grudin (1989) shows examples of how blindly following the maxim of consistency to the exclusion of other interface considerations can lead to poor usability design decisions (see also Reason, 1990).

4.3. Informative feedback

For every user action there should be some system feedback (Shneiderman, 1992: 73). This may take the form of immediate execution, change in state or value, correction message, confirmation message or in-progress message (see Ravden and Johnson, 1989: 56). System feedback is especially critical when system time is slow. Such feedback should be clear, concise and intelligible to the user.

4.4. Transparency

The system should permit the user's attention to be focussed entirely on the task being performed, without concern for the mechanics of the system (see Galitz, 1993). The computer is ideally suited for automating routine functions, and these should not detract from those activities requiring human attention. In CAI these may include time stamps, range and consistency checks, read-write operations, and other system functions. The interviewer's focus should be on the interviewing task, rather than on the operation of the CAI system.

However, there may be times when it is necessary for the user to see what the system is doing. One example in CAI may concern skip patterns. Usually these would be transparent to the interviewer, but there may be times when s/he needs to make judgements about an appropriate response to a root question. Without knowing the logic of the skip and the outcome of a particular choice, the interviewer cannot make an informed judgement as to the appropriate response.

4.5. Explicitness

Whereas the actions that the system performs without human intervention should be transparent to the user, the steps that the

user needs to take should be obvious. Norman (1988) uses the notion of affordances in evaluating the design of everyday objects. Essentially, affordances are properties of objects that suggest what sort of operations and manipulations can be done. For example, the design of a door handle suggests the operation to be performed, by affording pulling, pushing, turning or twisting actions. Affordances can be similarly applied to the human-computer interface: the computer screen should make the required user actions explicit or self-evident. Norman (1983) also cautions against the "tyranny of the blank screen" in DOS, where the "c:\" prompt provides the user with no clue as to what operation need to be performed. Many CAI systems assume that the user knows what to do in a particular situation, sometimes without providing any hints as to the expected action or guidance on where to find help to complete the task. Well-designed systems should make both the semantics (what can be done) and the syntax (how to do it) of the system explicit (Mayhew, 1992).

4.6. Comprehensibility

Systems should be understandable to users. Jargon, idiosyncratic language and abbreviations should be avoided. Norman (1988: 179) suggests ways to violate this guideline: "Be arbitrary. Use nonobvious command names or actions. Use arbitrary mappings between the intended action and what must actually be done." Where possible, natural language and real-world analogies should be used (Hix and Hartson, 1993). Ravden and Johnson (1989: 32) note that "The way the system looks and works should be compatible with user conventions and expectations." For example, using the Page Up, Page Down, and arrow keys for movement may make more sense than using function keys. The layout of dates, telephone numbers, etc. in the CAI system should match users' expectations or common conventions for the presentation of such information.

4.7. Tolerance

The system should be tolerant of human capacity to make errors. Galitz (1993: 26) writes: "The fear of making a mistake and not being able to recover from it is a primary contributor to a fear of dealing with computers." System design should recognize that errors will be made, and should include appropriate error prevention, detection and correction facilities (see Reason, 1990). Efforts should be made to prevent serious errors while facilitating easy recovery from more common errors. The more potentially disastrous an action, the more difficult it should be to perform. Thus, backing up to change a previous answer in CAI should be easier to do than suspending an interview in midstream. Actions should be easily reversible.

4.8. Efficiency

The system should be designed to minimize effort and maximize

efficiency on the part of the user. System response time is only one aspect of efficiency. As Mayhew (1992: 508) notes, overall task time is a function of both system and user response time. In other words, task time = system response time + system display rate + user scan/read time + user think time + user response time + time making errors + time recovering from errors. By improving the speed of the system (without attending to interface issues to reduce user time or errors), only the system side of efficiency (response time and display rate) will be addressed, without affecting other components of overall task time. All aspects of good interface design should facilitate the overall efficiency of operation. For example, user response time can be optimized by avoiding complex sequences of actions for common operations. User scan/read time can be reduced through effective screen design. Galitz (1993: 48) notes that system responsiveness should match the speed and flow of human thought processes, and offers some specific guidelines for various types of operations (see also Shneiderman, 1992: 284-297).

4.9. Supportiveness

This is closely related to the principles of explicitness and comprehensibility. Tolerance of errors and facilities for easy recovery from errors is another characteristic of a supportive system. The limited cognitive capacities of users should be recognized and accommodated. This can be done by reducing the amount of memorization of commands, codes, syntax and rules required by users (Brown, 1988: 97). Reliance on recognition rather than recall will help reduce cognitive burden for the user. Norman (1991: 6) writes, "It is typically the case that for systems with 40 plus commands, only about 7 commands show any frequency of use". Complex sets of commands and those that are rarely used are less likely to be remembered. Supportive systems provide online help and make it readily accessible to the user. If one needs to consult a manual to find out how to get online help, something is gravely wrong with the system.

4.10. Optimal Complexity

The early dictums on design (on both screen and paper) called for keeping things simple and maximizing the use of blank space. This view has given way to a recognition from a growing body of research (see Tullis, 1983; Coll and Wingertsman, 1990; Staggers, 1993) that users' preference for complexity exhibits an inverted U shape. Users both prefer and perform better with a moderate amount of complexity, rather than too simple or too complex. Galitz (1993: 35) notes that complexity should be commensurate with the capabilities of the system users. Complexity refers not only to the amount or density of information of the screen, but to all aspects of screen design. Hix and Hartson (1993: 49) recommend organizing the screen to manage complexity.

5. Lessons to be Learned from HCI

These general design guidelines for CAI systems should be seen as a set of goals for improving CAI design, rather than principles set in stone. There may be other characteristics of effective systems not mentioned here. There is also a recognition that compromises among these qualities or guidelines may be necessary to achieve optimal usability of CAI systems. Nonetheless, I have found examples of violations of each of the guidelines in various production instruments used by a number of different organizations. The CAI systems we use are clearly not perfect, and there is much room for improvement. These guidelines may help us on the road to quality improvement in instrument design to facilitate the work of our interviewers.

Many of these guidelines are not new to CAI. A number of these principles have already been articulated with regard to CAI. Nicholls and House (1987), for example, note that one of the general objectives of CAI systems is that they should meet interviewer needs. They explicate further: " displays must be quickly comprehensible [Principle 6], interviewers should have access to all needed information [Principle 9], opportunities for interviewer error should be minimized [Principle 7], and interviewer movement through the questionnaire, either forward or backward, should be expedited [Principle 8]" (Nicholls and House, 1987: 96). Despite these and other efforts to articulate design guidelines for CAI systems, it appears that little progress has been made.

If the only contribution made by human-computer interaction research to CAI was in the development of a set of general design principles (such as those outlined above), we would not have gained much. There are two additional keys to the applicability of HCI research to CAI. The first is a theoretically-grounded understanding of the interaction between human and computer and how the interface impacts the user and his/her task. Human factors research traces much of its theoretical roots to cognitive psychology (see for example Carroll, 1991), and it is this body of literature that will be most helpful to CAI design. The second critical lesson to be learned from HCI is the application of research methods to measure and understand the usability aspects of CAI systems (and user interfaces in general). The utility of HCI research lies not only in <u>what</u> was found, but also <u>how</u> it was found. A variety of methods are used in HCI research that can be readily adapted for use in CAI. These include usability testing in laboratory settings, experimental studies, observation and so on (see Shneiderman, 1992). Both theory and measurement are important to the partnership between HCI and CAI.

With regard to theory it is important to note that not all the findings of HCI research are equally applicable to computer-assisted interviewing. The nature of the interviewing task may be

very different from that studied in HCI (use of programming languages, word processors, spreadsheets, etc.). The partnership with HCI does not mean uncritically applying all findings in that field to CAI. Rather, by focusing on similarities and differences between CAI and other tasks involving human-computer interaction, we can distinguish between what is useful and what requires further exploration.

As far as measurement is concerned, many of the techniques already used in survey research can be applied to study interaction aspects of the task. These include:

(a) Cognitive laboratory investigations of interviewer-computer interaction using observation, protocol analysis, think alouds, etc.
(b) Laboratory-based experiments testing alternative designs or focusing on particular issues and actions interviewers face in CAI.
(c) Scripted mock interviews which may include tests of particular types of actions.
(d) Observation of production interviewing using computers.
(e) Experiments embedded in production data collection.
(f) Measurement of interviewer production and process (e.g. keystroke files, time stamps, monitoring, behavior coding).

Many of these methods parallel those used to study human-computer interaction, and can be productively applied to CAI.

6. Conclusion

So where do we go from here? In this discussion I have tried not to be too prescriptive in terms of ways to design user interfaces for CAI. Rather, I am advocating more of a design philosophy that (a) explicitly takes the users into account; (b) involves measurement of progress toward usability goals (e.g reducing learning, minimizing errors, maximizing user satisfaction, etc.); and (c) attempts to extract empirically-derived principles and guidelines that have general applicability beyond the particular system or interface on which they were tested or developed. These tasks can be greatly facilitated by learning from the field of HCI research.

In terms of action steps, I believe the field of computer-assisted interviewing can make great strides by doing the following:

(a) Apply what is already known about human-computer interaction and usability to CAI.
(b) Adapt HCI research methods to understand and explore usability issues in CAI. Conduct both qualitative and quantitative research on the interface between interviewer and computer in CAI.
(c) Explicitly incorporate usability testing as an integral part of the instrument development process.

(d) Identify gaps in our knowledge in the human-computer interface in CAI, and undertake research to close these gaps.
(e) Think ahead to new technologies and what we require of them, rather than being constrained by the limitations of existing CAI systems.

In this discussion, I have tried to turn our attention to the future rather than the immediate past. I see these issues both as challenges and as great opportunities for survey research. Let us not just embrace the machine, let us understand it and thereby unleash its full potential.

In doing so, we can learn a great deal from HCI research. In the same way that the field of questionnaire design has reaped great benefits from the partnership with cognitive psychology, so too can CAI benefit from interaction and collaboration with human-computer interaction or human factors researchers. Indeed, the benefit may well be mutual.

REFERENCES

Baker, R.P. (1992), "New technology in survey research: Computer-assisted personal interviewing (CAPI)." Social Science Computer Review, 10 (2): 145-157.

Balestrino, R., Fortunato, E. and Montagna, S. (1992), "A survey questionnaire design methodology." Pp. 489-501 in Westlake, A., Banks, R., Payne, C. and Orchard, T., Survey and Statistical Computing. New York: North-Holland.

Bayerl, J.P., Millen, D.R. and Lewis, S.H. (1988), "Consistent layout of function keys and screen labels speeds user responses." Proceedings of the Human Factors Society 32nd Annual Meeting, pp. 344-346.

Brown, C.M. (1988), Human-Computer Interface Design Guidelines. Norwood, NJ: Ablex.

Carroll, J.M. (ed.) (1991), Designing Interaction: Psychology at the Human-Computer Interface. New York: Cambridge University Press.

Coll, R. and Wingertsman, J.C. (1992), "The effect of screen complexity on user preference and performance." International Journal of Human-Computer Interaction, 2 (3): 255-265.

Connett, W.E., Mockovak, W.P. and Uglow, D.A. (1994), "Future CAI needs from a user perspective." Paper presented at the Census Bureau's 1994 Annual Research Conference and CASIC Technologies Interchange.

Couper, M.P., Groves, R.M. and Kosary, C. (1991), "Methodological issues in CAPI." Proceedings of the American Statistical Association, Survey Research Methods Section.

Davis, S. and Bostrom, R. (1992), "An experimental investigation of the roles of the computer interface and individual characteristics in the learning of computer systems." International Journal of Human-Computer Interaction, 4 (2): 143-172.

Dibbs, R. and Hale, A. (1993), "Questionnaire design in a paperless society." Proceedings of the Bureau of the Census 1993 Annual Research Conference. Washington: Bureau of the Census.

Falzon, P. (ed.) (1990), Cognitive Ergonomics: Understanding, Learning and Designing Human-Computer Interaction. San Diego: Academic Press.

Galitz, W.O. (1993), User-Interface Screen Design. Boston: QED.

Gould, J.D. (1988), "Designing for usability: The next iteration is to reduce organizational barriers." Proceedings of the Human Factors Society 32nd Annual Meeting, pp. 1-9.

Gould, J.D. and Lewis, C. (1983), "Designing for usability -- key principles and what designers think." Proceedings of CHI '83: Human Factors in Computing Systems. New York: ACM, pp. 50-53.

Grabinger, R.S. (1989), "Screen layout design: Research into the overall appearance of the screen." Computers in Human Behavior, 5: 175-183.

Groves, R.M. and Nicholls, W.L. (1986), "The status of computer-assisted telephone interviewing: Part II -- Data quality issues." Journal of Official Statistics, 2 (2): 117-134.

Grudin, J. (1989), "The case against user interface consistency." Communications of the ACM, 32 (10): 1164-1173.

Hix, D. and Hartson, H.R. (1993), Developing User Interfaces: Ensuring Usability Through Product and Process. New York: Wiley.

Hunter, P. (1993), "Introducing CAI standards within OPCS." Paper presented at the Second International Blaise Users Conference.

Kreighton, K., Matchett, S. and Landman, C. (1994), "Building integrated systems of CASIC technologies at the U.SA. Bureau of the Census." Paper presented at the Census Bureau's 1994 Annual Research Conference and CASIC Technologies Interchange.

Marchionini, G. and Sibert, J. (1991), "An agenda for human-computer interaction: Science and engineering serving human needs; Report of an invitational workshop sponsored by the National

Science Foundation." SIGCHI Bulletin, 23 (4): 17-32.

Mayhew, D.J. (1992), Principles and Guidelines in Software User Interface Design. Englewood Cliffs, NJ: Prentice-Hall.

Mumford, L. (1934), Technics and Civilization. New York: Harcourt, Brace and Company.

Nicholls II, W.L. and House, C.C. (1987), "Designing questionnaires for computer-assisted interviewing: A focus on program correctness." Proceedings of the Bureau of the Census Annual Research Conference. Washington, DC: Bureau of the Census.

Nicholls II, W.L. and Kindel, K.K. (1993), "Case management and communications for computer assisted personal interviewing." Journal of Official Statistics, 9 (3): 623-640.

Nielsen, J. and Levy, J. (1994), "Measuring usability: Preference vs. performance." Communications of the ACM, 37 (4): 66-75.

Norman, D.A. (1983), "Design principles for human-computer interfaces." Proceedings of CHI '83: Human Factors in Computing Systems. New York: ACM, pp. 1-10.

Norman, D.A. (1988), The Design of Everyday Things. New York: Doubleday.

Norman, D.A. (1993), Things that Make Us Smart: Defending Human Attributes in the Age of the Machine. Reading, MA: Addison-Wesley.

Norman, K.L. (1991), The Psychology of Menu Selection: Designing Cognitive Control of the Human/Computer Interface. Norwood, NJ: Ablex.

Pierzchala, M. (1993), "Computer generation of mega-version instruments for data collection and interactive editing of survey data." Proceedings of the Bureau of the Census Annual Research Conference. Washington, DC: Bureau of the Census.

Powell, J.E. (1990), Designing User Interfaces. San Marcos, CA: Microtrends Books.

Ravden, S. and Johnson, G. (1989), Evaluating Usability of Human-Computer Interfaces. Chichester: Ellis Horwood.

Reason, J. (1990), Human Error. Cambridge: Cambridge University Press.

Shneiderman, B. (1992), Designing the User Interface: Strategies for Effective Human-Computer Interaction. (2nd. ed.) Reading, MA: Addison-Wesley.

Staggers, N. (1993), "Impact of screen density on clinical nurses' computer task performance and subjective screen satisfaction." International Journal of Man-Machine Studies, 39 (5): 775-792.

Tullis, T.S. (1983), "The formatting of alphanumeric displays: A review and analysis." Human Factors, 25 (6): 657-682.

Weeks, M.F. (1992), "Computer-assisted survey information collection: a review of CASIC methods and their implications for survey operations." Journal of Official Statistics, 8 (4): 445-465.

DISCUSSION OF TWO PAPERS ABOUT CASIC

Sandra Sperry
Westat, Inc.

This is such an interesting time in survey research. After many years in which there was very little change in the approaches used to collect data and prepare it for analysis, the past ten years have produced enormous change. First there was the growth of centralized telephone interviewing using computer-assisted data collection methods. Then in the late 80's, government agencies began to experiment with using laptop computers to collect data for in-person interviews. Now it seems to be completely accepted that an agency will, at the very least, consider computer-assisted survey information collection (CASIC) techniques for all its surveys.

The impact of these two papers about CASIC comes in part from the fact that the Census Bureau and the Bureau of Labor Statistics do things on such a grand scale. How often do survey researchers get to casually mention that their pretest of a CATI system included 10,000 cases? This is what Jeanette Mon tells us about the CATI test done in 1982 to prepare for the 1992 Census of Agriculture. And Cathy Dippo tells us that the Parallel Survey, which was carried out for 18 months to assess the overall effects of redesigning the Current Population Survey (CPS), was "relatively small," that is, only 12,000 households per month.

The scale of these two projects, along with their emergence at a time when so many organizations are considering how best to implement CASIC, makes these two papers especially valuable. The papers are, among other things, historical records that document aspects of the development and application of CASIC at a time when CASIC is moving from being unusual and experimental to being the data collection mode of choice.

Both papers tell us a lot about process, although each paper focuses on different aspects of the process of applying CASIC methods. Cathy Dippo's paper tells us about the methodological processes used to redesign the CPS questionnaire for computer-assisted data collection. Jeanette Mon's paper focuses more on the processes used to manage all aspects of developing and using CASIC methods. Both include much helpful information for the researcher embarking on a project that will use CASIC. I would like to mention a few of the points that I found most relevant.

I will start with Cathy Dippo's paper. She notes that preliminary work on designing a new CPS questionnaire included laboratory work on how interviewers and respondents understand labor force terminology. This work led to developing questionnaires that were also tested with laboratory techniques. Then special respondent debriefing techniques were used with 2,300 CPS telephone respondents. The results of this debriefing verified the laboratory findings. These results are reassuring about the value of using laboratory techniques for questionnaire design. Most researchers do not have the resources to carry out a respondent debriefing of this scope and, therefore, are not able to evaluate the results of their laboratory work with this precision.

I was also very interested in how the CPS staff used CASIC techniques to help them with questionnaire design. The use of behavior coding while monitoring CATI interviews is an excellent way to maximize the benefits of a CATI or CAPI pretest. Programming follow-up probes to debrief respondents about specific answers that they gave during the main interview seems like another inspired idea for getting the most out of a pretest.

The design features of a CASIC questionnaire that have been incorporated into the new CPS instruments, that is, complex skips, wording that is specially tailored to the respondent's situation, built-in consistency checks, and dependent interviewing in longitudinal data collections, are aspects of CASIC that I think have been incorporated into a number of CATI and CAPI surveys. It was interesting to read about some of the ways in which the CPS instrument has changed to make use of CASIC, but the concepts presented in this part of the paper were not as new to me as some of the approaches used for instrument development. To learn even more about the instrument development process, I think many researchers would be interested in reading some of the technical reports written by the CPS Overlap Analysis Team. The technical reports present more detail about work that was done on this project, a project that included much more developmental research than most projects can afford.

I would like to turn now to Jeanette Mon's paper. As I mentioned earlier, this paper addresses more of the issues encountered in managing a large CASIC design and data collection. I was particularly struck by the number of different Census divisions (five) that needed to cooperate and communicate with one another in order to carry out the 1992 Census of Agriculture's CATI follow-up. Because different divisions had different responsibilities with regard to the same data, it was very important for them to meet regularly and to document carefully everything that they did. It is often easy for researchers working on smaller studies or in organizations with a less structured approach to dividing up responsibilities to assume that they do not need the level of formal communication and documentation that was required for the Census of Agriculture follow-up. I would hypothesize, however, that all survey research projects would be better off if they included more preparation of the formal specifications that Ms. Mon mentions in her paper.

Ms. Mon also mentions that to develop the basic system specifications that are required by the CASIC programming staff, the subject matter experts needed to go through a learning period. She notes that in some instances, the requisite specs were so complicated that it took a considerable amount of time to learn how to prepare them. I think this is a problem that many of us have encountered in developing CASIC instrumentation. The process requires much collaborative work among staff involved in all different aspects of a CASIC project, and we must occasionally spend time just learning how to communicate with one another.

The description of how different states were scheduled to be called in each of two telephone centers presents another lesson in managing a large scale CATI operation. It was very important that the scheduling of states be kept flexible so that when one or the other of the phone centers began to run low on work at particular times (as happened on two different occasions), new states could be installed to provide the level of work needed to keep interviewers working efficiently.

The paper about the Census of Agriculture's CATI follow-up included one more section that I found particularly enticing. The CATI software created a file with a complete history of all actions made with regard to a case. Staff used these files "to produce tables and graphs for management analysis." Some of the tables and graphs are included as attachments to the paper. I found these tables and graphs very intriguing and would like to know more about how the management staff used them during the field operations phase of the project.

In conclusion, I will not suggest the usual call for more research. Rather, I will end with a call for more papers from the research on which these two papers are based. There is a great wealth of material here to be explored, and I think there is much for all of us to learn from it.

Session 8

LONGITUDINAL SURVEYS

PANEL DESIGN AND ESTIMATION STRATEGIES IN THE NATIONAL MEDICAL EXPENDITURE SURVEY

Steven B. Cohen

Agency for Health Care Policy and Research
2101 E. Jefferson Street
Rockville, Maryland 20852

Presented at the COPAFS Seminar on New Directions in Statistical Methodology
May 25-26, Bethesda Holiday Inn

The views expressed in this paper are those of the author and no official endorsement by the Department of Health and Human Services or the Agency for Health Care Policy and Research is intended or should be inferred. The author wishes to acknowledge the contributions made by the following individuals who were co-authors of papers whose findings have been synthesized in this presentation: D.E.B, Potter, Barbara L. Carlson, and Ayah Johnson.

PANEL DESIGN AND ESTIMATION STRATEGIES IN THE NATIONAL MEDICAL EXPENDITURE SURVEY

Steven B. Cohen, Agency for Health Care Policy and Research

1. Introduction

The National Medical Expenditure Survey (NMES-2) was established to provide an assessment of the health care utilization, expenditures, sources of payment and health insurance coverage of the U.S. civilian noninstitutional population and the population using nursing and personal care homes and facilities for the mentally retarded for calendar year 1987. The core of the data collection effort for the non-institutionalized population was a series of interviews with a household sample that collected detailed information on health status, use of health care services, expenditures and sources of payments, insurance coverage, employment, income and assets, and demographic characteristics for calendar year 1987. The institutional population component was designed to obtain similar types of information for the institutionalized population residing in nursing and personal care homes (NH), and in facilities for the mentally retarded (MR) for calendar year 1987.

The NMES household survey was designed as a panel survey with an initial screening interview conducted in the fall of 1986 for a sample of approximately 35,600 addresses, to obtain information required for oversampling of specific policy relevant population subgroups. A subsample of about 15,000 households was selected for the detailed interviews. To reduce the deleterious impact of long recall periods on measurement error, data collection specifications required four separate interviews conducted with selected households at three to four month

intervals over a fifteen month period to obtain the required heath care data for calendar year 1987. The adoption of a panel design was also motivated by the analytical goal of measuring changes in health insurance coverage over the course of calendar year 1987.

The NMES Institutional Population Component was designed to yield unbiased national and regional estimates at the facility level and for the overall institutional user population. More specifically, the primary objective of the survey was to estimate the use of and expenses for health care services for all persons residing in institutions at any time during calendar year 1987. To obtain a nationally representative sample of the 1987 institutional user population, the survey included a sample of residents residing in selected facilities as of January 1, 1987, in addition to a representative sample of admissions to the selected facilities over the course of 1987. The union of these samples served to represent the 1987 institutional user population. The operational implications of a selection of both residents as of 1/1/87 and admissions over the course of 1987 required the adoption of a panel design for the survey. Interviewers made four distinct visits to each cooperating facility at approximately four month intervals to facilitate sample selection and data collection in the institutions.

This paper provides a summary of the sample design and estimation strategies adopted in the National Medical Expenditure Survey associated with the longitudinal features of the survey. For the Household Survey, particular attention is given to the consequences of adopting an address sample design in meeting specified survey design goals. The NMES Household survey also benefitted from a design strategy that included the selection of a sample of

households in the first round of data collection for the main study that were non-respondents to the screener interview. The impact of this design strategy on the overall survey response rate is also summarized. The paper includes a discussion of the estimation strategy adopted to adjusted for part-year nonresponse. With respect to the Institutional Population Component, both the sample design and the estimation strategy used to correct for the representation of individuals with multiple opportunities for sample selection over the course of 1987 are described in detail.

2. Address Sample Design in the NMES Household Survey

The Household Component of the National Medical Expenditure Survey (NMES) is a panel survey with 1987 as the reference period, collecting measures of health status, use of health care services, expenditures and sources of payment, insurance coverage, employment, income and assets, as well as demographic information for the U. S. civilian noninstitutional population. To meet analytical objectives, the survey included an oversample of the following policy relevant population subgroups: blacks, Hispanics, the poor and near poor, the elderly and persons with functional limitations. A separate screening interview was conducted in the fall of 1986 to facilitate the identification of these population subgroups. The screened households were selected for the main NMES household survey on the basis of the characteristics of the persons they included at the time of the screening interview. However, for the purposes of cost-efficiency and to maximize the response rate, the actual NMES Round One sample was characterized by an address sample design. An address sample design requires the interviewer to go back to the originally sampled address, whether the same household resides there or not. Consequently, the Round One sample consisted of all households living at the sampled addresses

at the time of the Round One interview, whether or not the screened households were still present.

The sample design considerations inherent in the selection of an address sample as an alternative to a household sample design are clearly not limited to the NMES, but generally applicable to all national household surveys that consider non-concurrent screening interviews to identify particular population subgroups targeted for oversampling. When the basic sampling units are defined to be the sample addresses themselves, rather than the households residing at the addresses at the time of the screening interview, a reduction in survey costs should be realized for the subsequent interview since movers would not have to be traced and interviewed at their new location. However, when the address sample design is used, each sample listing (or a representative subsample) should be returned to in the round of data collection following the screening interview, to interview all eligible respondents residing at the address (Cohen and Johnson, 1992). Addresses that were vacant at the time of the screening interview would have to be checked to determine their occupancy status. Dwelling units whose occupants refused to participate in the screener interview would need to be recontacted in addition to those units in which no eligible respondents were at home or available for being interviewed for the screening interview (Cox and Cohen, 1985; Cox et al., 1979).

The adoption of a household sample design would define households residing within the sampled addresses at the time of the screening interview as the basic sampling units. Individuals or families who moved out of the sample dwelling units would have to be traced and followed for interview. This design imposes greater control over the sample to insure that sample size targets are satisfied for the oversampled population subgroups. Furthermore, this design would

generally yield more precise survey estimates for the oversampled population subgroups than an address sample design, which is more vulnerable to the adverse effects of greater sampling weight variation. However, the advantages of the household sample design are less likely to be realized as the time period between the screening and follow-up interviews increases. Additional interviewers would have to be hired to conduct the Round One interview for the selected screener respondents that moved outside the PSUs that comprised the NMES screener sample. Furthermore, the field period would have to be extended to accommodate the time needed to locate the movers and conduct their interview.

The choice of design requires an evaluation of the potential loss in the precision of surveys estimates for the oversampled population subgroups, further contrasted with the potential savings in survey costs and higher response rate achieved by the address sample design. In this study, the consequences of adopting an address sample design for the National Medical Expenditure Survey are evaluated in terms of the realization of specified survey design goals.

The adopted NMES household survey sample design is a stratified area probability design with the following stages of selection: (1) selection of 165 primary sampling units (PSU's) which are counties, parts of counties or groups of contiguous counties; (2) selection of 2,317 segments within PSU's; (3) selection and screening of dwelling units within segments; and (4) selection of dwelling units based on demographic characteristics from the set of screened dwelling units (Cohen, DiGaetano and Waksberg,1991). The survey was sponsored by the Agency for Health Care Policy and Research and the NMES sample represents a union of the national sample frames of Westat, Inc., the prime contractor, and NORC, a subcontractor.

The NMES screener interview was conducted in the fall of 1986. The final NMES screener sample consisted of 35,634 addresses, of which 3,091 were identified as vacant, another 1,085 identified as not a dwelling unit and 250 addresses were determined to be ineligible for the survey. Of the 31,208 remaining dwelling units sampled, 28,458 responded to the screener interview, achieving a 91.2 percent response rate.

The Round One household sample of dwelling units was then selected from the screened households, using sampling rates that achieved the desired sample for specified population subgroups (DiGaetano,1987). The 15,130 dwelling units that constituted the targeted Round One sample, based on the demographic and health status profiles of their members at the time of the NMES screener interview, initially consisted of 16,615 responding reporting units. Reporting units were defined as individuals related by blood, marriage or adoption within a dwelling unit. These reporting units contained 39,885 individuals that represented the civilian non-institutionalized population as of Fall, 1986.

Figure 1 illustrates the transitions in sample composition that would occur between the screener and Round One interviews, as a consequence of the address sample design. The expected transitions include the movement of targeted households out of sampled addresses, changes in household composition, and the inclusion of replacement households that have moved into sampled addresses.

FIGURE 1 ABOUT HERE

It was also expected that some of the addresses contacted which were vacant in the screener field period would become occupied at the time of the Round One interview. Excluding

them from the sample would understate the number of recent moves in the sample. Consequently, a sample of 1,464 vacant addresses was selected to supplement the occupied addresses sampled from all screened households.

o Field Results

Field results indicated that the majority of the 15,130 sampled addresses consisted of the same reporting units that completed the screener interview. Transitions occurred as a consequence of the following situations:

1. movement by screener identified families out of sampled dwelling units,
2. creation of new reporting units, consisting of individuals related by blood, marriage or adoption, that have moved into sampled addresses since the time of the screener interview, and
3. changes in the Round One composition of reporting units that were initially identified at the time of the screener interview.

At the end of the Round One field period, it was determined that 722 (4.5 percent) of the 15,130 sampled addresses responding to the NMES screener interview were vacant. Another 253 sampled addresses were determined to be ineligible for the interview as a consequence of the death or institutionalization of targeted respondents, or due to a change in the original household composition to all military or student members. Furthermore, 847 reporting units that responded to the screener moved out of the sampled addresses, and were replaced by new reporting units at the time of the Round One interview. Overall, more than ten percent of the sampled addresses with screener respondents experienced at least one household move during the five to six month period that passed between interviews..

Within the 15,130 sampled addresses that constituted the targeted Round One sample, there were 15,590 reporting units identified as eligible or NMES interviews during the Round One field period out of 17,412 reporting units linked to the addresses during the screener or Round One interview. Of these, 14,060 responded (90.2 percent) to the Round One interview. The remaining 1,530 eligible RUs were classified as nonrespondents (9.8 percent) due to a refusal to complete the interview, unavailability during the field period, illness, or other nonresponse. Relative to the 1,464 sample addresses vacant at the time of the screener, 1,016 (69.4 percent) remained vacant at the time of the Round One interview. Of the 479 reporting units new to the sample, 408 (85.2 percent) responded to the first household interview. The joint screener-round one response rate for these targeted Round One addresses, including the screener vacant sample, was 82.1 percent. This was derived by multiplying the screener response rate (.912) with the combined Round One response rate for targeted sample and the vacant sample ((14,060 + 408)/(15,590 + 479) = .90).

At the person level, it was noted that 32,205 (80.7 percent) of the targeted 39,885 screener respondents also responded to the Round One interview. Of the 7,680 screener only-respondents, 3,150 individuals (41 percent) were in targeted reporting units that refused to complete the Round One interview. Another 224 individuals (2.9 percent) resided in reporting units that were determined to be ineligible at the time of the Round One interview. Consequently, 43.9 percent of the screener only- respondents would not have participated in the Round One interview, independent of the address sample design.

It was also determined that 1,673 of the screener only- respondents (21.8 percent) departed from the sample as a consequence of the movement of their reporting units (847 RUs)

away from the sampled address, which gained a new replacement household available for the Round One interview. Another 1,529 individuals (19.9 percent) left the sample due to the movement of their reporting units out of the sampled address, which was vacant at the time of the Round One interview. The remaining 1,104 screener-only respondents that departed from their sampled addresses (14.4 percent), were associated with reporting units for which at least one respondent completed both the screener and Round One interview at the sampled address. Consequently, 56.1 percent of the screener only- respondents (4,306) were not sampled in Round One as a function of the address sample design. The final set of respondents to the first round of the NMES household survey consisted of 36,259 individuals. Of the 4,054 individuals that were Round One only-respondents, 841 (20.7 percent) were associated with initially sampled reporting units that did not respond to the NMES screener interview, but were refielded to improved the overall NMES response rate. Independent of an address sample design, a decision to refield these cases could have been incorporated in the NMES sample design. The remaining 3,213 Round One only-respondents (79.3 percent) were added to the NMES sample as a function of the address sample design.

o A COMPARISON OF THE DEMOGRAPHIC CHARACTERISTICS OF NMES MOVERS

Of the 7,680 screener only-respondents, 43.9 percent would not have completed a Round One interview, independent of the address sample design for NMES. Under an alternative sample design that tracked all screener respondents that moved, efforts would have been made to locate or determine the status of 4,306 individuals linked to more than 1,500 reporting units.

The demographic characteristics of these targeted sample movers are presented in Table 1, and contrasted with the characteristics of the 3,213 replacement individuals that are new to the sample as a function of the address sample design (excluding the 841 new Round One respondents associated with RUs that were screener nonrespondents).

TABLE 1 ABOUT HERE

According to Current Population Reports, (U.S. Bureau of the Census, 1987) 18 percent of the population experienced a move during the course of the year. The NMES national estimate of population transition (10 percent) during the five to six month period that transpired between the screener and Round One interviews, compares well with census figures.

Overall, the 3,213 individuals that were new to the sample in Round One represented a 25 percent shortfall in targeted sample size, relative to the movers. The population subgroup that consistently experienced the greatest proportionate differential from sample targets were the elderly. In addition, a greater shortfall in targeted sample size characterized the black and Hispanic population subgroups, relative to the domain that represented whites and other races. It was recognized, however, that even if the targeted sample of movers were to be traced, the expected Round One yield would need to reflect nonresponse and loss in sample due to death, institutionalization, or inability to locate.

Under the assumption of a ten percent loss in sample due to nonresponse (which was the experience for non-movers), and a conservative assumption of an additional five percent loss due to death, institutionalization, or inability to locate, the expected Round One yield for the targeted NMES movers was 85 percent. Relative to the expected number of movers completing the Round One interview, the effective overall shortfall in sample size was only 12 percent. When

focused on specific population subgroups, a more pronounced shortfall is sample size was noted, particularly for the elderly. The address sample design was more effective for subpopulation subgroups defined by race or ethnicity, where black or Hispanic households were likely to be replaced by other black or Hispanic households.

The additional measures of poverty status and functional impairment considered in the sampling scheme were not examined, as a consequence of transitions in classification over time that characterize these measures. Screening for these measures is further complicated by the considerable degree of movement into and out of poverty in any two years (Moeller and Mathiowetz, 1990) and the potential movement of elderly individuals from a state of good health to that of disability over time.

Although addresses that experienced a sample movement out were characterized by a shortfall in expected sample size by virtue of the address sample design, a more important consideration was the effect on sample yields for the overall Round One sample. A comparison of the targeted overall Round One sample (36,511) and the resultant Round One sample (35,418) revealed a three percent shortfall. Relative to the targeted Round One sample, a greater shortfall in sample size characterized the black and Hispanic population subgroups (four percent), when contrasted with the white and other races population subgroup (2 percent). Clearly, the overall effect of the address sample design on the Round One sample yields was minimal.

o EFFECT OF ADDRESS SAMPLE DESIGN ON PRECISION

The precision of the NMES survey estimates for the population subgroups of analytical

interest was expected to decrease only marginally, as a function of the small loss in sample size attributable to the address sample design. Even for the population subgroup characterized by the largest relative sample loss (4.3 percent for Hispanics), the expected increment in the standard errors of their associated survey estimates was only two percent. The major cause for concern, however, was the greater variation in sampling weights for the population subgroups of analytical interest (Kish. 1965). Table 2 presents a summary of the variation in the sampling weights that characterized the targeted and actual Round One samples. More specifically, the distributions of the unadjusted sampling weights, and their coefficient of variation are compared across the alternative sample designs for Round One.

TABLE 2 ABOUT HERE

The actual Round One sample was characterized by a higher level of variation in sampling weights across all population subgroups targeted for oversampling. This was most obvious for the black and Hispanic population subgroups. The greater variation in their sampling weights was primarily a function of the movement of new individuals, with characteristics targeted for oversampling, into sampling addresses selected in NMES at a lower sampling rate based on their composition at the time of the screener interview.

The impact of the address sample design on the precision of survey estimates was directly examined by a comparison of the standard errors for a representative set of demographic and health status measures that were available from the screener interview. It was necessary to restrict this comparison to these measures, since the screener only-respondents did not have any

of the information that was collected only in the Round One interview.

The person level demographic measures under investigation included region, size of city, marital status and veteran status. In addition, the health status measures under investigation included all questions in the NMES screener data base that identified functionally impaired individuals for the purposes of oversampling. For each of the survey measures under consideration, the ratio of the standard errors derived from the address sample design and from the potential Round One sample was computed. Study findings indicate only a slight increase in the standard errors of the survey estimates that characterize the address sample design (ratios generally ranged from 1.01 to 1.04). Even for the population subgroup characterized by the greatest loss in precision (white/other race, 65+), the average increment in the standard errors of their associated survey estimates was only six percent.

The cost savings achieved by the address sample design were estimated to be $175,000, and were attributable to not having to locate the targeted individuals that had a change of address since responding to the NMES screener interview. The availability of these additional funds allowed for the implementation of more intensive survey nonresponse conversion techniques to enhance the overall NMES household survey response rate. As a consequence of the address sample design, the response rate that characterized the NMES Round One interview was not affected by a component of nonresponse due to inability to locate.

Under a NMES design that attempted to include the movers in the sample, the field period for the Round One interview would have had to be extended. This would be required in order to accommodate the time needed to locate the movers and conduct the interview. This

extension in the field period would have a potentially adverse effect on the quality of data obtained from the movers, due to the extended length of the recall period that characterized their interview. Use of an address sample design for NMES eliminated the component of response error that would be attributable to this particular extension in recall period. Alternatively, the problems inherent with tracing movers in a NMES survey that considered a household design could be noticeably reduced by a field redesign that facilitated a significant reduction in the average lag period between the screener and Round One interviews.

The NMES address sample design was cost effective and achieved an acceptable response rate relative to a design that would have tracked movers. In addition, the quality of the Round One data was expected to be improved, as a consequence of the shorter length of recall period (where respondents were required to provide health care information at the time of the Round One interview relative to 1/1/87) that characterized the respondents of the address sample design. The major potential limitation of the adopted design was with respect to the precision of survey estimates. Even on this sensitive dimension, study findings indicated only a slight loss in the precision of survey estimates derived from the address sample design. Consequently, the overall benefits of the address sample design were realized in the National Medical Expenditure Survey, with only minimal effects on the precision of survey estimates.

3. Refielding A Sample of Screener Refusals and Other Nonresponding Dwelling Units in the First Round of the NMES Household Survey.

A supplemental sample of refusals to the screener and other nonresponding dwelling units was

also included in the round 1 sample, to improve the overall NMES response rate over the four rounds of data collection. In addition to 1,600 dwelling units classified as screener interview refusals, another 916 dwelling units were categorized as "other nonresponse". The other nonresponse classification included dwelling units where no one was home after four calls, where the potential household respondents were unavailable during the screener field period, and where respondents were considered too ill to complete the interview.

Refusals and other nonresponding dwelling units were classified into the same four strata used for sampling vacant addresses. A sample of 645 addresses with dwelling units categorized as refusals to the screener was then selected from the 1,600 addresses with such eligible units and a sample of 376 addresses with dwelling units categorized as other nonresponding units were selected from 916 addresses with eligible units with this classification. Consequently, the only set of screener nonrespondents that were not considered for selection in the first round of the NMES household survey consisted of dwelling units with a language problem, dwelling units where the interviewer was unable to enter the structure, and those classified in the residual nonresponse category. These 234 dwelling units were not targeted for selection in round 1 because of the low probability of converting them to participate in the four rounds of the household survey. By recontacting the nonrespondents to the NMES screener interview, 99.26 percent of the sample of all households were given a chance to participate in the NMES round 1 interview. It should be noted, however, that any new dwelling units built during the five month lag time between the screener and the round 1 interview were not included in the round 1 sample frame. The final round 1 sample is summarized in Table 3.

TABLE 3 ABOUT HERE

Round One Field Results

As noted, the joint screener-round one response rate for the targeted round 1 address sample based on responses to the screener interview, including the screener vacant sample, was 82.1 percent. This round 1 response rate was improved by recontacting nonrespondents to the NMES screener interview. In the set of refielded addresses which were characterized by screener nonresponse, 372 (40.1 percent) of the 928 eligible reporting units responded in round 1. This sampling approach contributed an additional 3.3 percent to the joint response rate. Overall, the joint screener round one response rate for the NMES was 85.4 percent.

4. Part-Year Nonresponse in the NMES Household Survey

Panel surveys are subject to wave nonresponse, which occurs when responses are obtained for some but not all waves of the survey. Prior to selecting an adjustment strategy, it is essential that the patterns of nonresponse and potential correlates are examined. When levels of sample attrition are high, it has been suggested that wave nonresponse is related to level of respondent burden (Czajka, 1986). Generally, the process of sample attrition is non-random (Short and McArthur, 1986). Partial respondents are often distinguished from their complete respondent counterparts on a number of dimensions. These distinctions should be considered in the determination of a nonresponse adjustment strategy.

Two general strategies to correct for sample attrition have gained acceptability in the statistical community: sampling weight nonresponse adjustments and imputation (Kalton, 1986;

Czajka, 1986; Herringa and Lepkowski, 1986). Weight adjustments for sample attrition are rather straightforward to implement and avoid the increase in sampling variance expected in survey estimates as a function of imputation. The primary limiting features of the technique are the loss of large amounts of useful data provided by partial respondents, and the deleterious impact of large nonresponse adjustments to sampling weights on the precision of survey estimates. Imputation for panel attrition allows for the inclusion of partial respondents in the derivation of survey estimates, and the use of the data they have provided. Within the imputation framework, there are several general methodologies that are considered to compensate for panel nonresponse: longitudinal hot deck imputation (Herringa and Lepkowski, 1986). Due to the large number of time dependent analytical measures that are directly affected by sample attrition and the sophisticated software requirements to implement the technique, imputation compensation strategies are the more costly of the two types of strategies.

o Characteristics of the NMES Part-Year Respondents

Of 36,753 key participants in the NMES household survey, 2,294, or 6.2 percent, responded for some, but not all, of the time period in 1987 for which they were eligible. Key sample respondents to the household survey consisted of all civilian non-institutionalized individuals who responded to the Round One interview, in addition to individuals who joined responding Round One reporting units and did not have an opportunity for selection during the period of time that spanned the Round One field period (new babies, military returning to civilian status, individuals in institutions or outside the country returning to their primary

residence). To ascertain the potential level of nonresponse bias that was attributable to partial response in NMES, it was necessary to determine whether the part-year respondents differed systematically from their full year counterparts. To facilitate these comparisons, demographic profiles of these two distinct respondent groups were compared for the following measures: gender, age, race/ethnicity, marital status, years of school completed, census region, size of city, and indicators of functional status.

Since the nonresponse adjustment strategies employed to correct for part-year nonresponse would be dependent on a respondent's data profile for his period of participation in the survey, it was necessary to impose a thresshold on what constituted a minimally acceptable time representation of partial data for making annual national health care estimates. In NMES-2, the minimum part-year response requirement of data for more than one-third of a respondent's period of eligibility followed the approach taken in the 1980 National Medical Care Utilization and Expenditure Survey (sponsored by the National Center for Health Statistics). In NMES-2, 48.6 percent of the part-year respondents (1,114 individuals) who constituted 6.2 percent of the person level sample, did not satisfy this criterion. Consequently, they were treated as total nonrespondents and a standard weighting class adjustment for non-response was applied to the sampling weights of the remaining respondents to correct for their exclusion.

The comparisons of the race/ethnicity distributions for the part-year respondents with data for at least one-third of their period of eligibility in 1987 (henceforth referred to as part-year respondents) and full year respondents revealed a significantly greater representation of whites

and other races (non-black, non-Hispanic) among the individuals that provided complete data when compared to the partial respondents (Cohen, Johnson and Carlson, 1989). This pattern was also observed in the 1977 National Medical Care Expenditure Survey (NMCES), which was the predecessor of NMES (Cohen, 1982). Alternatively, the partial respondents had a higher representation of Hispanics than their complete data counterparts. There was also a significantly higher representation of the partial respondents living in the 19 largest Standard Metropolitan Statistical Areas (SMSAs) in the nation, indicating a greater likelihood of sample attrition in NMES in the large urban metropolitan areas. Furthermore, partial respondents were more likely to reside in the Northeast region of the United States than individuals who provided full year response profiles.

No significant gender differentials were noted across respondent groups. With respect to marital status, the full year respondents were more likely to be married than their partial respondent counterparts. Alternatively, the partial respondents had a greater representation of never married individuals, which mirrored the 1977 NMCES experience. Furthermore, a comparison of the age distributions that characterized the respective respondent groups revealed that partial respondents were more likely to be aged 20-29. Since this age group represents a highly mobile population subgroup, this suggests that the sample attrition that they displayed in the NMES was partially a function of migration. Furthermore, complete respondents had a higher representation of elderly individuals between the ages 70-74 than their partial respondent counterparts. With respect to years of education completed, the partial respondents had a higher representation of individuals with some high school training, as indicated by at least 9-11 years

of education.

To minimize the nonresponse bias in survey estimates attributable to partial response, an appropriate estimation strategy is needed to adjust the data for the remaining sample members who did not respond for their entire period of eligibility. In view of the programming time and cost necessary to implement an imputation strategy to correct for partial nonresponse, and the relatively small representation of partial nonrespondents in NMES, the advantages of this technique for NMES application were not obvious. The technique would require the linkage of partial respondents to complete respondents with matching demographic and health status profiles, the extraction of data from the complete respondent which corresponded to the nonresponding time period of the partial respondent, and its imputation to the partial respondent for each time dependent variable in NMES-2, representing a complex and expensive process.

Traditionally, when the level of partial response is low, it is often preferable to treat partial respondents as complete nonrespondents. Using this approach, only those sample participants providing complete data would be used in the analysis. This was the approach taken for the remaining part-year respondents in the NMES-2 household survey. Weighting classes were formed by cross-classifications of the following measures: race/ethnicity, age and gender. The person level sampling weights for the full year respondents were further post-stratified to poverty status estimates derived from the Current Population Survey.

5. An Estimation Strategy to Represent the Institutional User Population in the NMES-2

Institutional Population Component

The Institutional Population Component (IPC) of the National Medical Expenditure Survey (NMES) was established to provide an assessment of the health care utilization, costs, sources of payment and health insurance coverage of the U.S. institutionalized population residing in nursing and personal care homes (NH), and in facilities for the mentally retarded (MR). The primary objective of the survey was to estimate the use of and expenses for health care services for all persons residing in institutions at any time during calendar year 1987. To obtain a nationally representative sample of the 1987 institutional user population, the survey included a sample of residents residing in selected facilities as of January 1, 1987, in addition to a representative sample of admissions to the selected facilities over the course of 1987. The union of these samples served to represent the 1987 institutional user population.

o Sample Design

The adopted NMES institutional population survey is a stratified, two stage probability design with two phases of facility selection. Current residents (residents on January 1, 1987) and admissions (persons admitted between January 1, and December 31, 1987) were sampled within participating facilities at the second stage.

The IPC facility sample consisted of 851 eligible nursing and personal care homes and 730 eligible facilities for the mentally retarded. Facilities were considered to be respondents to the survey when they completed a Facility Questionnaire. Consequently, the IPC facility level response rate was 95.2 percent for nursing and personal care homes, and 94.7 percent for

facilities for the mentally retarded.

The design of the survey required that the institutional use and expenditure data for current residents were to be collected for their entire period(s) of institutionalization in 1987. In contrast, IPC data collection for the admissions sample began with their first admission to a sampled IPC eligible facility, independent of prior institutional stays over the course of 1987. Consequently, their 1987 institutional data collection period was constrained. For estimation purposes, individuals who responded for at least a third of their eligibility period of institutional data collection were considered respondents.

In the nursing and personal care IPC sample, 805 participating facilities (94.6 percent) allowed for the selection of a sample of their residents as of January 1, 1987. Overall, 3,392 eligible residents were selected, representing a national nursing and personal care home population of 1.5 million residents. Similarly, in the IPC sample of facilities for the mentally retarded, 685 participating facilities (93.8 percent) allowed for "current" resident sampling. Overall, 3,738 eligible residents were selected, representing a national population of 212,000 residents in facilities for the mentally retarded.

Overall, the response rate in the IPC for current residents providing data for at least one-third of their period of institutionalization in 1987 was 89.5 percent for residents in nursing and personal care homes (.946 facility level response rate x .946 resident level response rate), and 88.4 percent for residents in facilities for the mentally retarded (.938 facility level response rate x .942 resident level response rate). This data was to be obtained in the IPC through the administration of the Institutional Use and Expenditure Questionnaire (IUEQ), to be completed

by facility staff (Edwards and Edwards, 1989).

The admissions sample consisted of 2,608 eligible sampled admissions to nursing and personal care homes, and 889 eligible sampled admissions to facilities for the mentally retarded. Sampled admissions were defined to be individuals who were admitted to the sampled IPC facility during 1987 and had no prior admissions to that facility during the survey year.

In the nursing and personal care home sample, 758 participating facilities (89.1 percent) allowed for the sample selection of admissions at all rounds of data collection. Similarly, 657 facilities for the mentally retarded (90 percent) allowed for the sample selection of new admissions at all rounds of data collection. .

Overall, the response rate for new admissions providing data for at least one-third of their period of institutionalization in 1987 was 81.2 percent for those sampled in nursing and personal care homes (.891 facility level response rate x .911 admission response), and 81.3 percent for admissions sampled in facilities for the mentally retarded (.900 facility level response rate x .903 admissions response rate).

Data collected from facility respondents included facility level characteristics, physical and mental health status and functional limitations of sampled persons, and their socio-demographic characteristics and residential history in and immediately before admission to sampled facilities. Information collected on health care services use and expenses included facility services provided, charges and sources of payment, hospitalizations during the institutionalized period and associated conditions, number of physician contacts, and contacts with other medical care providers and therapists.

This data collection effort was referred to as the Survey in Institutions (SII). During each visit,

interviewers obtained or constructed lists of residents from each cooperating facility and proceeded to select the sample (Edwards and Edwards, 1989). The current resident sample was selected from a list of all residents in sample facilities as of January 1, 1987. Similarly, the admission samples were selected on three separate occasions in cooperating facilities from separate lists of all admissions that occurred during the following time periods in 1987: January 1 to April 30, May 1 to August 31, and September 1 to December 31. Sampled persons were followed throughout 1987. For those who left the facilities in which they were selected, facility use and expenditure data were collected up to the time of discharge. If a sample person entered another IPC-eligible facility, the institutional data collection procedures were continued in the new facility.

Since study objectives required data that facility staff could not be expected to provide, the IPC also included a Survey of Next of Kin. This survey consisted of a set of questionnaires administered to community respondents who knew about sampled persons and their lives outside of institutions. Data were obtained on use and expenditures linked to specific residence periods, living arrangements outside of sampled institutions, perceptions of health status and functional limitations, and arrangements for informal care.

o The NMES Institutional User Population

The IPC sample design consisted of two distinct selections of 1987 institutional users: the first selection was designed to provide a representative national sample of residents in IPC

eligible facilities as of January, 1, 1987 (current residents); and the other selection was designed to provide a nationally representative sample of 1987 admissions to IPC eligible facilities. The strict requirement of a single day of sample eligibility for the current resident sample resulted in a single opportunity of selection for each sampled current resident as of 1/1/87. Imposition of a similar restriction for the selection of admissions, requiring the selection of individuals experiencing their first institutional stay in 1987, would have simplified the sample design by allowing each sampled institutional user a single opportunity of selection. Since this information regarding an individual's prior periods of institutionalization was not available at the time of sample selection, and often unavailable from facility records, such a restriction could not be imposed. Resident history information for sampled admissions was often obtained through the IPC Survey of Next of Kin, whereby community respondents who knew about sampled persons would be the primary source for information regarding prior institutional stays.

As a consequence of the sample selection scheme that was employed, an individual who experienced more than one institutional stay over the course of 1987 had multiple chances of selection into the IPC sample. Furthermore, a subset of sampled admissions was determined to have also resided in an IPC eligible facility on 1/1/87, indicating an overlap with the independent sample of January 1 residents. In order to identify the sample of institutional users that had multiple opportunities of selection in the IPC sample, it was necessary to further classify the IPC sample of institutional users according to their institutional experience over the course of 1987.

o Classification of Current Resident Sample

With respect to the population of institutional users that resided in an IPC eligible facility on January 1, 1987 (referred to as current residents), four mutually exclusive and exhaustive classifications are specified in order to characterize their institutional experience over the course of 1987 (Figure 1). More specifically, the first group of current residents consists of institutional users who remained in the same facility over the course of 1987. Residents in this class are referred to as **static full year residents (Group 1)**. The next class of current residents consists of institutional users who remained in the same facility for only part of calendar year 1987, with no subsequent admissions to IPC eligible facilities (i.e., nursing and personal care homes and facilities for the mentally retarded that met the definition for eligibility in the NMES IPC) over the course of 1987. These institutional users are referred to as **single stay part-year residents (Group 2)**. Current residents with this classification could have returned to the community as a member of the civilian non-institutionalized population, been transferred to an out-of scope facility or institution (e.g., acute care hospital, psychiatric institution), or died while in the institutional setting.

FIGURE 2 ABOUT HERE

The remaining current residents experienced at least one subsequent admission to an IPC eligible facility over the course of 1987. They are distinguished in the following manner. The first group consists of current residents who were formally discharged from the facility they resided in as of January 1, 1987 and subsequently were readmitted to the same facility over the course of 1987. These institutional users are referred to as **current residents with re-admissions to same facility (Group 3)**, and consist of residents with one or more re-admissions restricted to the same facility over the course of 1987. Alternatively, the remaining group consists of

current residents with admissions to different facilities (Group 4). This classification also includes institutional users who were re-admitted to the same facility they resided in on 1/1/87 and who also experienced at least one admission to a different IPC eligible facility over the course of 1987.

Classification of Individuals Sampled as Admissions

In a complementary manner, four mutually exclusive and exhaustive classifications are specified in order to characterize individuals who experienced at least one admission to an IPC eligible facility over the course of 1987: these individuals are referred to as sampled admissions (Figure 1). The first class of affected institutional users consists of individuals who were not residents in IPC eligible facilities as of 1/1/87 and whose first institutional admission in 1987 was in a sampled IPC facility (Primary Sample Facility (PSF)). Institutional users in this class are referred to as **sampled admissions with initial 1987 admission to an IPC Primary Sample Facility (Group 5)**. This group includes individuals with one or more unique admissions to eligible institutions over the course of 1987 (Groups 5a or 5b, and Groups 5c or 5d, respectively) .

The next classification identifies individuals with a 1987 institutional admission to an IPC Primary Sample Facility, who were also residents in the same facility as of 1/1/87. Institutional users in this group are referred to as **residents sampled as admissions with 1987 admission(s) to the same IPC Primary Sample Facility (Group 6)**. This group of institutional users was already represented in the NMES Institutional Population Component Survey by current residents classified as **residents with re-admissions to same facility (Group 3)**, and by a subset of the

residents with admissions to different facilities (Group 4) who were also re-admitted to the Primary Sample Facility. As a consequence of the ease in identifying these current residents who were re-admitted to the same facility over the course of 1987, they were not considered eligible for IPC data collection. A related group of institutional users consists of individuals with an admission to an IPC Primary Sample Facility who were also residents in an non-sampled IPC eligible facility as of 1/1/87. Such individuals are referred to as **residents in non-sampled facilities with 1987 admission(s) to an IPC Primary Sample Facility (Group 7)**. This class of institutional users was also dually represented in the NMES IPC sample by a subset of the **current residents with admissions to different facilities (Group 4)**. These sampled admissions were not excluded from IPC data collection as a consequence of being unable to determine, at the time of sampling, whether they were institutionalized in some other IPC eligible facility on January 1, 1987.

The remaining group of institutional users with 1987 admissions consists of individuals who did not reside in IPC eligible settings as of 1/1/87, and who were admitted to non-sampled IPC eligible facilities in 1987 prior to an admission to a Primary Sample Facility. Institutional users in this class are referred to as **admissions in IPC Primary Sample Facilities with initial 1987 admission to a non-sampled IPC eligible facility (Group 8)**. This class of institutional users was also dually represented in the NMES IPC sample by a subset of the **new admissions with initial 1987 admission to an IPC Primary Sample Facility (Group 5d)**.

IPC Sample Distribution of Institutional Users

A summary of the IPC sample distribution of institutional users, further distinguished by

facility setting, is presented in Table 4. These tabulations include individuals with response profiles for utilization and expenditure data for at least a third of their period(s) of institutionalization in 1987, beginning with their **sampled stay.** For current residents, this translated to their entire period(s) of institutionalization in 1987. Alternatively, individuals with 1987 admissions were classified as respondents when response profiles were obtained for at least a third of their institutional experience in 1987, beginning with their **sampled admission.** Inclusion of these partial respondents in the derivation of national health care utilization and expenditure estimates for the institutional user population requires implementation of an imputation procedure to adjust for missing time dependent data (Cohen and Potter, 1990). Since IPC data collection for the admission sample began with their first admission to a sampled IPC facility, their 1987 institutional data collection period was constrained. When resident history information was not obtained either through the IPC Survey of Next of Kin or the IPC Survey in Institutions for periods in 1987 prior to their **sampled admission**, resident history profiles were imputed for the missing time gaps in 1987 (Potter and Cunningham, 1990). Inclusion of these sampled admissions with prior periods of institutionalization in the derivation of national health care utilization and expenditure estimates for the institutional user population also requires implementation of **additional** imputation procedures to correct for missing time dependent data (Cohen and Potter, 1990).

The sample of institutional users in nursing and personal care homes consisted of 5,585 respondents, with 3,209 (57.5 percent) sampled as current residents and 2,376 (42.5 percent) sampled as 1987 admissions (Table 4). Relative to the current resident sample, 2,586 (80.6 percent) were classified as static full year residents (Group 1), another 150 (4.7 percent) were

single stay part-year residents (Group 2), with the remaining 473 (14.7 percent) experiencing subsequent admissions to eligible facilities in 1987 (Groups 3 and 4). After excluding the 448 sampled admissions that were considered ineligible for IPC data collection (Group 6), the admission sample was dominated by 2,002 (84.3 percent) institutional users who were not institutionalized on 1/1/87 (Groups 5 and 8).

Alternatively, the sample of institutional users in facilities for the mentally retarded consisted of 4,323 respondents, with 3,520 (81.4 percent) sampled as current residents and 803 (18.6 percent) sampled as 1987 admissions (Table 4). As a consequence of the low representation of **sampled admissions** in these types of facilities in any given year, IPC sample size specifications for the admission sample in facilities for the mentally retarded did not assume separate national estimates would be made for the **sampled admissions**. Relative to the current resident sample, 3,089 (87.8 percent) were classified as static full year residents (Group 1), and another 73 (2.1 percent) were single stay part-year residents (Group 2). The remaining 358 experiencing subsequent admissions to eligible facilities in 1987 (Groups 3 and 4) with the majority (316) experiencing admissions to non-sampled facilities (Group 4). After excluding the 76 sampled admissions that were considered ineligible for IPC data collection (Group 6), the admission sample was primarily represented by 432 (53.8 percent) institutional users who were not institutionalized on 1/1/87 (Groups 5 and 8).

TABLE 4 ABOUT HERE

o Estimation Strategy for the Institutional User Population

The stratified, multi-stage probability sample design adopted for the IPC institutional user population allows for the derivation of approximately unbiased estimates of health care parameters at the national level. This is conditioned upon the application of sampling weights to the sample data that properly reflect the sample selection scheme. The sampling weight for a sample member is defined as the reciprocal of a sample unit's probability of selection (Cox and Cohen, 1985). The estimation strategy for the IPC includes additional adjustments for all levels of nonresponse experienced in the survey. Nonresponse adjustments to the sampling weights have been implemented at the facility level and the institutional user level. To further improve the precision of survey estimates that characterize the IPC sample, post-stratification adjustments on facility level and resident level characteristics have also been implemented, using information from the 1986 Inventory of Long Term Care Places (Flyer, 1992).

One estimation strategy under consideration attempted to maximize the precision in survey estimates that characterize the institutional user population by the inclusion of all responding sampled institutional users in the estimation process. However, the implementation of this estimation strategy is not without penalty. In order to derive national estimates of the health care utilization and expenditure experience for the institutional user population, an imputation strategy must be considered to correct for missing time dependent health care data associated with institutional stays in 1987 prior to the sampled admission. Greater programming resources are required to implement the imputation process that corrects for missing time dependent health care utilization and expenditure data. The approach requires a determination of the exact time period for which institutional data is missing, a linkage between the institutional user with missing time dependent data to the best matching donor with a complete

data profile (using minimum distance function techniques), and imputing the appropriate time dependent data from donor to recipient (Cohen, 1992). The inclusion of sampled units that have multiple opportunities of selection in the derivation of national estimates requires additional adjustments to the survey sampling weights that reflect corrections to sample unit selection probabilities. Furthermore, the inclusion of a multiplicity adjustment to the estimation weights of institutional users selected from the IPC admission sample adds greater variability to their sampling weight distribution, partially limiting the expected gain in precision associated with an increase in sample size.

Given the complexities associated with the implementation of this strategy, and the need to provide timely national health care expenditure estimates of the institutional user population, an alternative approach was adopted for the derivation of use and expenditure estimates and other time dependent measures. More specifically, the alternative estimation strategy restricted the admission sample to institutional users whose first institutional stay in 1987 was in a sampled facility (i.e., Group 5). Adoption of this approach obviates the need for an imputation strategy to correct for missing time dependent data associated with institutional stays in 1987 prior to an institutional user's sampled admission. Furthermore, the restriction of the IPC admission sample to a sample of first institutional stays in 1987 obviates the need for a multiplicity adjustment to estimation weights and an adjustment for dual frame representation of residents in facilities as of 1/1/87. This is a consequence of limiting the sample of institutional users to a single opportunity of selection. Institutional users determined to have experienced institutional stays prior to their sampled admission (Groups 7 and 8) would be defined as ineligible for the

purposes of estimation.

Implementation of this approach results in a sample diminution of 513 out of 5585 respondents for the sample in nursing and personal care homes (9.2 percent reduction) and a comparable loss of 403 out 4,323 respondents in the sample of facilities for the mentally retarded (9.3 percent reduction). The impact of this sample restriction on the precision of survey estimates translates to a 4.9 percent increase in the standard errors that characterize the survey estimates of all institutional users in nursing and personal care homes and a corresponding 5 percent increment for the sample in facilities for the mentally retarded. Greater increments in standard errors are to be noted for the subset of institutional users that experience admissions during 1987. Although the magnitude of this loss in precision is non-negligible, it falls within acceptable levels when contrasted with the time and resource demands inherent in the implementation of an imputation strategy to correct for missing time dependent data associated with institutional stays in 1987 prior to an institutional user's sampled admission. Furthermore, consideration of the restricted first admission sample for the purposes of estimation eliminates exposure to a component of nonresponse bias due to missing time dependent data for prior institutional stays. This component of bias is often only partially reduced through application of imputation techniques. Implementation of imputation strategies to correct for missing time dependent health care utilization and expenditure data associated with prior institutional stays will inform future methodological investigations regarding the impact on survey estimates and their precision due to the inclusion of individuals whose sampled admission was not their first institutional stay in 1987.

6. Summary

The next cycle of the National Medical Expenditure Survey (NMES-3) will be in the field from 1996 through 1997, in order to obtain health care utilization, expenditure and insurance coverage information that characterizes the health care experience of the civilian non-institutionalized population and the population in nursing home for calendar year 1996. A number of design strategies that have been adopted in NMES-2 will also be considered for the NMES-3. One design feature of the NMES-2 Household Survey that will not be adopted for NMES-2 is the address sample design. This decision was not based on any design limitations that were identified in NMES-2 with the adoption of the address sample design, but driven by additional analytical demands placed on the survey. One of the primary motivations for the choice of 1996 as the time period for data collection was the need to have baseline data to assess the impact of health care reform on the nation's health care experience. A number of fast-track states are already in the process of implement health care reform initiatives. Consequently, the households identified for sample selection through the administration of a screening interview will also be subject to an additional interview in the fall of 1995. to gauge their satisfaction with the health care delivery system, their perceptions regarding access, and their current level of health insurance coverage. Since one of the analytical objectives of the NMES-3 household survey will be to assess the longitudinal changes in insurance coverage and access to health health care system over time, including data from the fall of 1995, it will be necessary to follow the same individuals that complete the screening and baseline interviews in the fall of 1995.

The NMES-2 was successful with its strategy to re-field a supplemental sample of

refusals to the screener and other nonresponding dwelling units in the first core round of data collection for the Household Survey. This strategy resulted in a three percent improvement to the overall NMES-2 response rate over the four rounds of data collection. As response rates are monitored for both the NMES-3 household and nursing home surveys, this design strategy will again be given serious consideration.

As a consequence of the panel designs of the NMES-3 household and nursing home surveys, wave nonresponse will remain a concern. As in NMES-2, methodological investigations will be conducted to determine whether the part-year respondents differ systematically from their full year counterparts. Both the level of nonresponse encountered in NMES-3 and the results of the evaluation of the patterns of nonresponse will guide the choice of the nonresponse adjustment strategy that is to be implemented.

Based on the estimation strategy considered in NMES-2 to represent the institutional user population, the NMES-3 nursing home survey design will employ a data collection scheme that limits the likelihood of multiple opportunities of selection into the survey. The planned design will restrict the admission sample for estimation purposes to a sample of individuals experiencing their first institutional stay in a sampled facility for the targeted survey year. Results from the NMES Institutional Population Component Feasibility Study revealed that facility respondents are able to provide accurate information regarding an individual's prior period of institutionalization in a given year. It is recognized that the restriction of the sample of admissions to the individuals experiencing their first institutional stay in a sampled facility for

the targeted survey year cannot be completely implemented at the time of sample selection. However, the ability to use facility information on prior admissions for the selected sample will result in significant cost savings to the survey, based on a reduction in unnecessary data collection activities associated with cases that are not eligible for estimation purposes.

7. References

Cohen, S. B. (1982). "An Analysis of Alternative Imputation Strategies for Individuals with Partial Data in the National Medical Care Expenditure Survey", Review of Public Data Use, 10:153-165.

Cohen, S. B., P. Flyer and D.E.B. Potter (1987). "Sample Design of the Institutional Population Component of the National Medical Expenditure Survey". Presented at the Annual Meeting of the American Public Health Association, New Orleans, LA.

Cohen, S. B., R. DiGaetano and J. Waksberg (1991). Sample Design of the National Medical Expenditure Survey Household Survey, Methods 3. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research.

Cohen, S. B., D.E.B. Potter and P. Flyer (1993). Sample Design of the National Medical Expenditure Survey Institutional Population Component, Methods 6. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research.

Cohen, S.B., A.E. Johnson and B.L. Carlson (1989). "An Analysis of Part-Year Nonresponse in the Household Component of the National Medical Expenditure Survey". Journal of Economic and Social Measurement, 15:281-299.

Cohen, S.B., and D.E.B. Potter (1993). "An Estimation Strategy to Represent the Institutional User Population in the 1987 National Medical Expenditure Survey". Journal of Economic and Social Measurement, 19:159-177.

Cohen, S.B. and A.E. Johnson (1992). "Address Sample Effects the National Medical Expenditure Survey". The American Statistician 46: 100-105.

Cohen, S.B.and D.E.B. Potter (1990). "An Analysis of Part-Year Nonresponse in the Institutional Population Component of the National Medical Expenditure Survey". American Statistical Association, Proceedings of the Section on Survey Research Methods.

Cox, B.G., and S.B. Cohen (1985). Methodological Issues for Health Care Surveys. New York, Basel: Marcel Dekker, Inc.

Cox, B., R. Folsom, B. Moser and T. Virag (1979). Housing Unit Versus Household as the Sampling Unit, Report No.RTI/1725/01-01S, Research Triangle Park, North Carolina.

Czajka, J.L. (1986). "Discussion of Papers on Nonresponse in the Survey of Income and Program Participation". Proceedings of the Section on Survey Research Methods.

DiGaetano, R.G. (1987). Sampling Report for the Selection of Dwelling Units for the Household Survey of the NMES. NMES Report No. 1.00. Westat, Inc., Rockville, Maryland.

DiGaetano, R.G., and J. Waksberg (1986). Sample Design of the Household Survey (HHS) and the Survey of American Indians and Alaska Natives (SAIAN). NMES Report No. 1.46. Westat, Inc., Rockville, Maryland.

Edwards, W. S. and B. Edwards (1989). National Medical Expenditure Survey: Questionnaires and Data Collection Methods for the Institutional Population Component, Methods 1. Department of Health and Human Services, Public Health Service, National Center for Health Services and Health Care Technology Assessment, DHHS Pub. No. (PHS) 89-3440.

Flyer, P. (1992). Sample Design for the Institutional Population Component (IPC) of the National Medical Expenditure Survey. NMES Internal Report. Westat, Inc., Rockville, Maryland.

Herringa, S.G. and J.M. Lepkowski (1986). "Longitudinal Imputation for SIPP". ASA.

Proceedings of the Section on Survey Research Methods.

Kalton, G. "Handling Wave Nonresponse in Panel Surveys". Journal of Official Statistics, 2:303-314.

Kish, L. (1965). Survey Sampling. New York: New York, John Wiley and Sons.

Moeller, J.F. and N. A. Mathiowetz (1990). Problems of Screening for Poverty Status. American Statistical Association: Proceedings of the Section on Survey Research Methods.

Potter, D.E.B. and P.J. Cunningham (1990). "Constructing Year-Long Profiles of Residence Histories for a National Sample of Nursing Home Users." American Statistical Association, Proceedings of the Section on Survey Research Methods.

Sirken, M. G. (1970). "Household Surveys with Multiplicity." Journal of the American Statistical Association, 65,257-266.

Short, K. and E. McArthur (1986). "Life Event and Sample Attrition in the Survey of Income and Program Participation". ASA, Proceedings of the Section on Survey Research Methods.

U.S. Bureau of the Census (1987). Current Population Reports, Series P-25, No. 957.

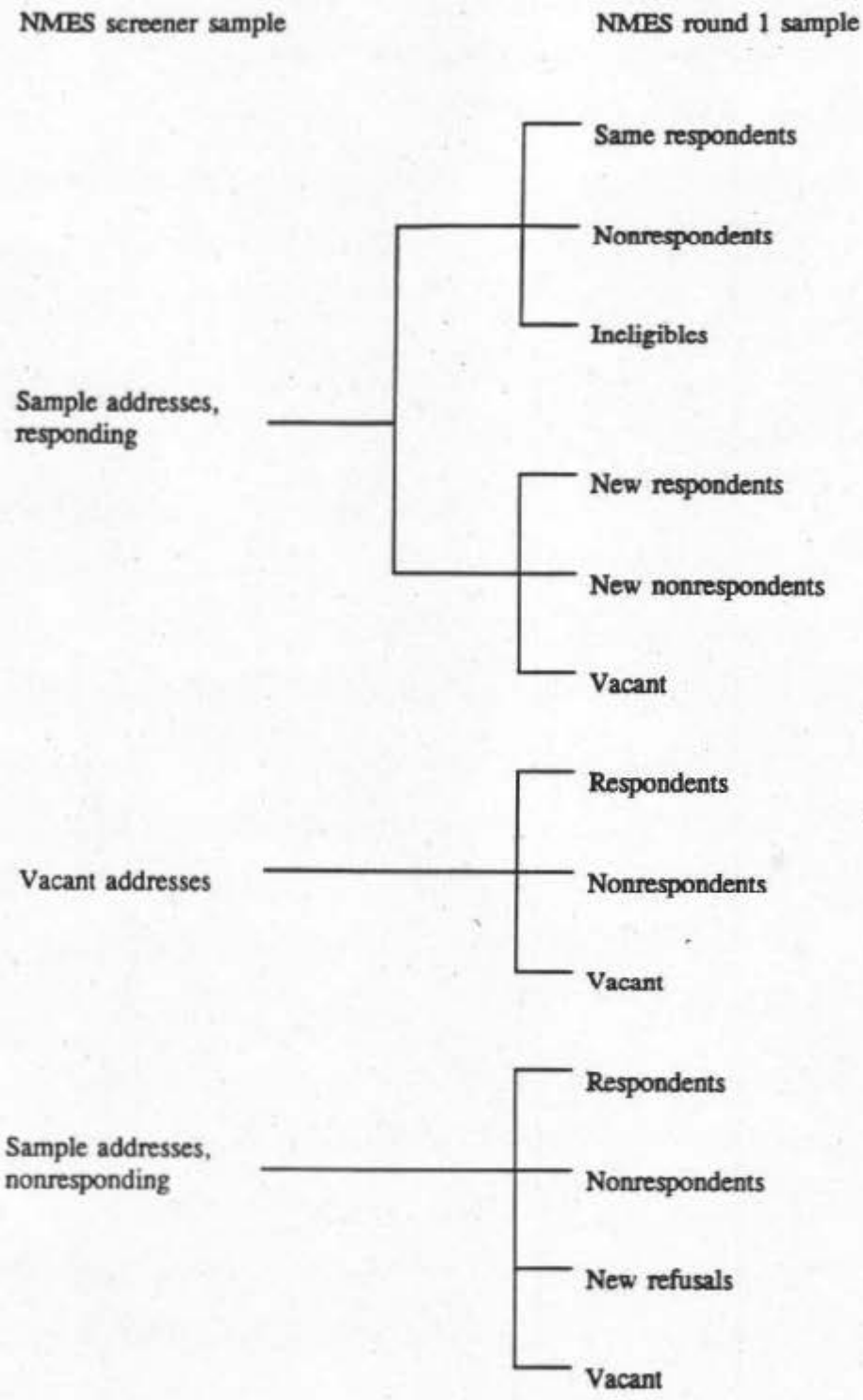


Figure 1. Results of NMES Address Sample Design: Transitions from Screener to Round 1.

Table 1. Comparison of Demographic Characteristics of NMES Movers

Race/ethnicity	Age	Screener-only respondents[a]	Expected[b] Round 1 response	Round 1-only respondents[c]	Round 1-only respondents relative to: Targeted sample (%)	Round 1-only respondents relative to: Expected sample (%)
Hispanic	<65	615	523	447	72.7	85.5
	65+	13	11	6	46.2	54.6
Subtotal		628	534	453	72.1	84.8
Black/non-Hispanic	<65	978	831	686	70.1	82.6
	65+	33	28	20	60.6	71.4
Subtotal	<65	1,011	859	706	69.8	82.2
White/other	65+	2,400	2,040	1,926	80.3	94.4
		267	227	128	47.9	56.4
Subtotal		2,667	2,267	2,054	77.0	90.6
Total		4,306	3,660	3,213	74.6	87.8
National estimates (in thousands)		(24,060)		(21,500)		

[a] Screener-only respondents are those who moved away from sampled addresses prior to the administration of the NMES Round 1 interview.

[b] Under assumption of 85% response rate.

[c] Round 1-only respondents are individuals new to the NMES sample as a function of the address sample design.

Source: Agency for Health Care Policy and Research: National Medical Expenditure Survey, United States, 1987.

Table 2. Variation in Sampling Weights and Mean Standard Error Ratios

		Distribution of weights						
		Potential Round 1 sample			Actual Round 1 sample			
		Coefficient of variation	Percentiles		Coefficient of variation	Percentiles		
Race/ethnicity	Age		1st	99th		1st	95th	99th
Hispanic	<65	.08	2562	3420	.19	2562	3420	7935
	65+	.08	2562	3925	.19	2562	3420	7935
Black/non-Hispanic	<65	.07	2562	2659	.21	2562	2959	5572
	65+	.05	2562	2949	.12	2562	2949	7935
White/other	<65	.31	2659	8235	.25	2863	8235	8235
	65+	.21	2863	8235	.36	2659	5193	8235
Total		.46	2562	8235	.46	2562	8215	8235

* Mean of standard error ratios for address sample design relative to potential Round 1 sample from household sample design.
Source: Agency for Health Care Policy and Research. National Medical Expenditure Survey, United States, 1987.

Table 3. NMES Household Survey field results at the RU level, round 1

Screener disposition	Number of addresses targeted for round 1	Round 1 responding	Round 1 refusal	Round 1 other non-response	Movers	Other ineligible	Vacant, not a dwelling unit	Total	Overall response rate
Responding	15,130	14,060	1,234	296	847[a]	253[a]	722[a]	17,412	
Response rate	0.912	0.902	0.079	0.019					0.821[b]
Vacant	1,464	408	47	24	0[a]	15[a]	1,016[a]	1,510	
Response rate	--	0.852	0.098	0.050					
Nonresponding	1,021	372	470	86	2[a]	11[a]	111[a]	1,052	
Response rate	0.081	0.401	0.506	0.093					0.033
Nonresponding not fielded	234								
Response rate	0.007	–	–	–	–	–	–	–	0.000
Total		14,840	1,751	406	849	279	1,849	19,974	0.854

[a]Ineligible for round 1.
[b]Combined round 1 response rate for responding and vacant addresses based on screener interview.
Source: Agency for Health Care Policy and Research. National Medical Expenditure Survey.

Figure 2.
Graphic Representation of Institutional User Population in the 1987 National Medical Expenditure Survey

User Type	*Institutional Experience Over the Course of 1987*			
Users Samples as Current Residents				
	January 1			December 31
Group 1: Static full year residents	Sample Facility			
Group 2: Single stay part-year residents	Sampled Facility			
Group 3: Current residents with re-admission(s) to the same sampled facility[1]	Sampled Facility		Sampled Facility	▲
Group 4: Current residents with admission(s) to different facilities	Sampled Facility		Non-Sampled Facility	▲
Users Sampled as Admissions				
Group 5a: Admissions with initial 1987 admission to sampled facility, remained in facility for remainder of year		Sampled Facility		
Group 5b: Admissions with initial 1987 admission to sampled facility, single stay part-year residence		Sampled Facility		
Group 5c: Admissions with initial 1987 admission to sampled facility and readmission(s) to the same facility[2]		Sampled Facility	Sampled Facility	▲
Group 5d: Admissions with initial 1987 admission to sampled facility and admission(s) to non-sampled facilities		Sampled Facility	Non-Sampled Facility	▲
Group 6: Residents sampled as admission(s) with 1987 admission(s) to the same sampled facility[3]	Sampled Facility	Sampled Facility		▲
Group 7: Residents in non-sampled facilities with 1987 admission(s) to sampled facilities	Non-Sampled Facility	Sampled Facility		▲
Group 8: Admissions in sampled facilities with initial 1987 admission to a non-sampled facility		Non-Sampled Facility	Sampled Facility	▲

[1]Sampled persons meeting the criterion for User Type 3 and User Type 4 were classified as User Type 4.
[2]Sampled persons meeting the criterion for User Type 5c and User Type 5d were classified as User Type 5d.
[3]Sampled persons meeting criterion for User Type 6 were classified as ineligible.
▲ Potential additional admissions.

Table 4
Distribution of the Institutional User Population, by User and Facility Type, Unweighted (National Medical Expenditure Survey – Institutional Population Component: United States, 1987)

User type	Nursing and personal care homes		Facilities for the mentally retarded	
	Number	Percent	Number	Percent
Users Sampled as Current Residents				
Group 1: Static full year residents	2,586	80.6	3,089	87.8
Group 2: Single stay part-year residents	150	4.7	73	2.1
Group 3: Current residents with re-admission(s) to the same sampled facility	269	8.4	42	1.2
Group 4: Current residents with admission(s) to different facilities	204	6.3	316	8.9
Total current residents	3,209	100.0	3,520	100.0
Users Sampled as Admissions				
Group 5a: Admissions with initial 1987 admission to sampled facility, remained in facility for remainder of year	1,106	46.6	283	35.2
Group 5b: Admissions with initial 1987 admission to sampled facility, single stay part-year residence	432	18.2	68	8.5
Group 5c: Admissions with initial 1987 admission to sampled facility and readmission(s) to the same facility	159	6.7	21	2.6
Group 5d: Admissions with initial 1987 admission to sampled facility and admission(s) to non-sampled facilities	166	6.9	28	3.5
Group 6: Residents sampled as admission(s) with 1987 admission(s) to the same sampled facility	448[a]	0.0[a]	76[a]	0.0[a]
Group 7: Residents in non-sampled facilities with 1987 admission(s) to sampled facilities	374	15.7	371	46.2
Group 8: Admissions in sampled facilities with initial 1987 admission to a non-sampled facility	139	5.9	32	4.0
Total sampled as admissions	2,376[a]	100.0[a]	803[a]	100.0[a]
Overall Total	5,585[a]		4,323[a]	

[a]Users classified into Group 6 were ineligible for data collection and were thus excluded from the totals and percent distributions.
Source: Agency for Health Care Policy and Research.

Council of Professional Associations on Federal Statistics

SEMINAR ON NEW DIRECTIONS IN STATISTICAL METHODOLOGY
Session on Longitudinal Surveys

Methodological Issues Encountered in Following a Cohort of Eighth Graders

Steven J. Ingels, NORC; Jeffrey A. Owings, NCES

Bethesda, Maryland
May 26, 1994

This paper is based on the National Education Longitudinal Study of 1988 (NELS:88). NELS:88 is a longitudinal micro-database sponsored by the National Center for Education Statistics, U.S. Department of Education. The ideas and opinions expressed in this paper are those of the authors and do not necessarily represent positions or conclusions endorsed by NCES.

Introduction. Longitudinal studies typically employ a probability sample of a unit (for example, individuals, institutions [e.g., schools], groups [e.g., families] or inanimate objects [e.g., dwelling units]) that is drawn at one point in time, then repeatedly observed, so that change in units can be measured over time. Longitudinal designs provide a powerful vehicle for reliably[1] measuring individual-level change and development as well as for describing the dynamics of change and the processes that are associated with it. At the same time, longitudinal studies have both inherent and potential limitations (see, for example, Pearson 1989 for a discussion of advantages and disadvantages of longitudinal surveys).

Some of the most important of these limitations can in large measures be overcome if the study is properly executed or if potential limitations are explicitly addressed in the design. This is the case, in particular, for three specific threats to longitudinal sample representativeness.[2]

(1) *Undercoverage*. Potentially biasing undercoverage[3] may arise from any of several sources. It may arise from deliberate or inadvertent exclusion of part of the baseline "target" sample, or may arise if baseline nonrespondents are not pursued in subsequent waves. Undercoverage problems may also arise if the eligibility of ineligible baseline students whose eligibility status is subject to change is not reassessed in succeeding rounds.

(2) *Need for Freshening*. The unit or cohort being studied may become less statistically representative of the target population (or less policy-relevant) over time. For example, a sample of individuals in a given geographical area may become less representative of that area as sample members disperse and other individuals move into the area. A sample of eighth graders two years later is not fully representative of the nation's tenth graders at the second point in time.

(3) *Attrition and Nonresponse*. Sample attrition poses substantial risks for a longitudinal study's representativeness. This danger can be overcome if high response rates are maintained across all rounds, and may be partially compensated for in weighting.

In this paper, the National Education Longitudinal Study of 1988 (NELS:88) serves as an example of how these three specific problems of representativeness in a longitudinal study may be approached and overcome. We more briefly comment on some additional potential sources of survey error.

1. Description of NELS:88

As a point of entry into our topic, we briefly describe NELS:88 by summarizing its goals, surveys administered, response rates, and analysis potential.

1.1 *Goals of NELS:88*

Beginning in 1988 with a cohort of 26,432 eighth graders attending 1,052 public and private schools across the nation, NELS:88 was designed to provide longitudinal data about critical transitions experienced by students as they leave eighth grade school settings, progress through high school (or drop out), enter and leave postsecondary institutions, and enter the work force. The 1988 eighth grade cohort has been followed at two-year intervals (specifically, first follow-up – 1990; second follow-up – 1992) with a third follow-up currently (spring 1994) underway.

Major features of NELS:88 include:

- the integration of student, dropout, parent, teacher, school administrator and school records (transcript) surveys;
- the initial concentration on an eighth grade student cohort with follow-ups at two year intervals;

- the inclusion of supplementary components to support analyses of geographically or demographically distinct subgroups (for example, selected state supplements; oversamples of Asians and Hispanics, and of students in private schools); and

- the design linkages to previous longitudinal studies (High School and Beyond [HS&B], the National Longitudinal Study of the High School Class of 1972 [NLS-1972]) and other current studies (for example, the National Assessment of Educational Progress [NAEP] testing program and high school transcript data collections).

The longitudinal design of NELS:88 permits the examination of change in young people's lives and the role of schools, teachers, community, and family in promoting growth and positive life outcomes. In particular, data from NELS:88 can be used to investigate issues in the context of the family, community, school, and classroom including:

- Students' academic growth over time;

- The transition from eighth grade to high school and the transition from high school to the labor market or postsecondary education;

- The process of dropping out of school, as it occurs from the end of eighth grade on;

- The role of schools in helping the disadvantaged;

- The school experiences and academic performance of language minority students;

- The influence of ability grouping, program type, and coursetaking patterns on future educational persistence and achievement; and

- The features of effective schools.

1.2 Surveys administered

NELS:88 components, by wave, are summarized in Figure 1.

Figure 1: Base Year Through Fourth Follow-Up – NELS:88 Components

BASE YEAR	FIRST FOLLOW-UP	SECOND FOLLOW-UP	THIRD FOLLOW-UP
spring term 1988 GRADE 8 **Students:** Questionnaire, Tests* **Parents:** Questionnaire **Principals:** Questionnaire **Two Teachers per student:** (taken from English, social studies, mathematics, science)	spring term 1990 MODAL GRADE = SOPHOMORE **Dropouts, Students:** Questionnaire, Tests **Principals:** Questionnaire **Two Teachers per student:** (taken from English, social studie mathematics, or science)	spring term 1992 MODAL GRADE = SENIOR **Dropouts, Students:** Questionnaire, Tests, H.S. Transcripts **Parents:** Questionnaire **Principals:** Questionnaire **One Teacher per student:** (taken from mathematics or science)	spring 1994 MODE = H.S. + 2 YEARS **All Individuals:** Questionnaire **FOURTH FOLLOW-UP** spring 1997 HS + 5 YEARS **All Individuals:** Questionnaire

* Reading, social studies, math and science tests are administered in the three in-school rounds.

1.3 *Analysis potential - longitudinal vs. cross-sectional applications.*

Analytic Levels. The NELS:88 design enables researchers to conduct analyses on three principal levels: (1) within-wave (or cross-sectional) analysis at a single time point, (2) cross-cohort analysis (by comparing cross-sectional NELS:88 findings to those of comparable populations studied earlier in HS&B and NLS-72) and (3) cross-wave (or longitudinal[4]) analysis.

The first analytic level within NELS:88 is cross-sectional. By beginning with a cross-section of 1988 eighth graders, following a substantial subsample of these students at two-year intervals, and freshening the 1990 and 1992 samples to obtain representative national cross-sections of tenth and twelfth graders, the study also provides a statistical profile of America's eighth graders, high school sophomores, and high school seniors.

A second analytic level extends representative cross-sections to intercohort comparisons. NELS:88 provides researchers with data for drawing comparisons with previous NCES longitudinal studies. After the release of NELS:88 first follow-up data, researchers were able to conduct trend analyses with the 1980 sophomore cohort of HS&B. With completion of the NELS:88 second follow-up, comparisons may be made among NELS:88, HS&B, and NLS-72 senior cohorts. To facilitate cross-cohort comparisons, some of the questionnaire items used in the NLS-72 and HS&B high school surveys were repeated in NELS:88, and data processing and file conventions were kept consistent, to the maximum possible extent, with HS&B and NLS-72.[5]

The third analytic level is longitudinal, and utilizes repeated measurements on the same individuals over time. However, because NELS:88 comprises three nationally representative grade- and year-defined cross-sections, it supports multiple panels:

1988	eighth graders two, four, six, and nine years later
1990	sophomores two, four, and seven years later
1992	seniors two and five years later[6]

Change Analysis. Cross-sectional analysis provides a snapshot at a single point in time. Repeated cross-sectional analysis, and longitudinal analysis, permit the measurement of change over time. Change (and stability) over time can be measured at the group or individual level:

(1) *At the group level*, change can be measured across the successive cross-sections--eighth graders in 1988, sophomores in 1990, and seniors in 1992. In the same way, multicohort assessments such as NAEP can estimate overall and subgroup gains in specific subject matter proficiency across selected points in the school career (e.g., between fourth, eighth, and twelfth grade). In addition, NELS:88 and comparable studies (e.g., NLS-72 and HS&B) can be analyzed as repeated cross-sections (e.g. of seniors in 1972, 1980/82, and 1992) to measure trends. A cross-sectional time-series such as NAEP also measures trends (e.g. in math achievement for 17 year olds from 1973 to 1990 for the nation and subgroups).

A principal weakness of change measurement at the group level--whether one is looking at rolling (e.g., eighth graders in 1988, sophomores in 1990, seniors in 1992) or repeated (e.g., eighth graders in 1988, 1990, and 1992) static cross-sections[7] is that it sometimes masks individual change; high levels of individual change are not incompatible with stability at the aggregate level. Thus, for example, looking at the proportion of 1988 eighth graders in 1988 who were out of school in 1990 (6.8%) and comparing this to the proportion out of school in 1992 (11.6%) masks the cumulative number of individuals who were

1990 or 1992 spring term dropouts, since some 1990 dropouts had returned to school by 1992.[8] A *locus classicus* of this phenomenon is found in studies of poverty and welfare recipience. While the proportion of adolescent mothers receiving AFDC over time is relatively constant, the AFDC population is not. Mobility onto, and off, the AFDC's rolls is demonstrated by longitudinal data provided by the NLSY, but would not be apparent from repeated cross-sectional results. Likewise, PSID data show that while poverty rates may be roughly stable over time, poverty spells for individuals and households tend to be relatively brief.[9]

(2) Change can also be analyzed *at the individual level* over time. The latter possibility—true longitudinal measurement—represents, for most purposes, the unique strength of the NELS:88 design. Following individual educational histories generally provides the best basis for drawing causal inferences about educational processes and their effects. Two broad kinds of analysis scenarios are possible. Longitudinal analysis can involve repeated measures of the same outcome—for example, test data can be used to measure growth in academic achievement over time. Or longitudinal analysis can show how conditions at an earlier time point are predictive of outcomes at a later time point. For example, one might examine how eighth graders with single or clustered "risk factors" (for example, such status risk factors as coming from a low-income home, having parents who did not finish high school, and so on; or such behavioral risk factors as cutting classes, lack of participation in extracurricular activities, and so on) fared two years later (for example, what proportion had dropped out, repeated a grade, and so on).

While longitudinal studies are prospective, in that they offer the opportunity to record new events, longitudinal analysis may be either retrospective or prospective. In NELS:88, priority in the baseline was given to questions predictive of future behavior. However, while questions that asked for *reasons* for past behavior were deliberately avoided, some retrospective questions were posed, when their focus was on simple descriptions of salient past events. For example, parents were asked whether their eighth grader had attended a Head Start program[10] or kindergarten or preschool, whether other of their children (respondent's elder siblings) had dropped out of school, and so on.

2. Sample Representativeness

This section discusses three key issues. *First*, eligibility and exclusion rules, particularly as applied in the NELS:88 base year, and the measures taken in later rounds of the study to deal with the potential for undercoverage biases that might result from these exclusions. *Second*, the need for sample freshening to ensure representative sophomore and senior cohorts in 1990 and 1992, and the procedures undertaken to bring that sample freshening about. And *third*, attempts to minimize sample attrition and nonresponse error.

2.1 Eligibility: Excluded Students and Undercoverage Bias.

In the base year of NELS:88, students were sampled through a two-stage process. First, stratified random sampling and school contacting resulted in the identification of the school sample; second, students were randomly selected (with oversampling of Hispanics and Asians) from within cooperating schools.

The target population for the base year comprised all public and private schools containing eighth grades in the fifty states and the District of Columbia. Excluded from the NELS:88 school sample are Bureau of Indian Affairs (BIA) schools, special education schools for the handicapped, area vocational schools that do not enroll students directly, and schools for dependents of U.S. personnel overseas.[11] The student population excludes students with severe mental handicaps, students whose command of

the English language was not sufficient for understanding the survey materials (especially the cognitive tests), and students with physical or emotional problems that would make it unduly difficult for them to participate in the survey. This chapter discusses (1) the consequences of student exclusion for the research design and results, and (2) the special measures that have been undertaken in NELS:88 to compensate or correct for the effects of exclusion. Before either of these two topics is pursued in detail, however, it will be desirable to say more about student exclusion in the NELS:88 base year–the 1987-88 school year during which the eighth grade cohort was selected and surveyed.

To better understand how excluding students with mental handicaps, language barriers, and severe physical and emotional problems affects population inferences, data were obtained on the numbers of students excluded as a result of these restrictions.

Seven ineligibility codes defining categories of excluded students were employed at the time of student sample selection:

A - attended sampled school only on a part-time basis, primary enrollment at another school.

B - physical disability precluded student from filling out questionnaires and taking tests.

C - mental disability precluded student from filling out questionnaires and taking tests.

D - dropout: absent or truant for 20 consecutive days, and was not expected to return to school.

E - did not have English as the mother tongue AND had insufficient command of English to complete the NELS:88 questionnaires and tests.

F - transferred out of the school since roster was compiled.

G - was deceased.

Before sampling, school coordinators–members of the school staff, typically an assistant principal or guidance counselor who acted as liaison between the school and the study–were asked to examine the school sampling roster and annotate each excluded student's entry by assigning one of the exclusion codes. Because eligibility decisions were to be made on an individual basis, special education and Limited English Proficiency (LEP) students were not to be excluded categorically. Rather, each student's case was to be reviewed to determine the extent of limitation in relation to the prospect for meaningful survey participation. Each individual student, including LEPs and physically or mentally handicapped students, was to be designated eligible for the survey if school staff deemed the student capable of completing the NELS:88 instruments, and excluded if school staff judged the student to be incapable of doing so. School coordinators were told that when there was doubt, they should consider the student capable of participation in the survey. Exclusion of students after sampling ("post-roster ineligibles") occurred either during the sample update just prior to survey day, or on survey day itself. Such exclusion after sampling normally occurred because of a change in student status (for example, transfer, death). However, in very rare instances such exclusions reflected belated recognition of a student's pre-existing ineligibility–that is, if an annotation error was made and an ineligible student selected for the sample in consequence of such an error, ineligibility became apparent later in the survey, whereupon the student was excluded.

Excluded students were divided into those who were full-time students at the school (categories B, C, and E) and those who were not (categories A, D, F, & G). Our main concern here is with students who were full-time students at the school but who were excluded from the sample. Excluding these students will affect estimates made from the sample.

Students in categories A (n=329), D (n=733), F (n=3,325), and G (n=6) were either not at the school or were present only part time (with primary registration at another school, hence a chance of selection into NELS:88 at another school). Thus excluding students in these categories has no implications for making estimates to the population of eighth grade students.

It should be noted that students in category F, those who had transferred out of the sampled school, had some chance of being selected into the sample if they transferred into another NELS:88 sampled school just as transfers into NELS:88 schools from non-NELS:88 schools had a chance of selection at the time of the sample update. The sampling of transfer-in students associated with the sample update allowed NORC to represent transfer students in the NELS:88 sample.

The total eighth grade enrollment for the NELS:88 sample of schools was 202,996. Of these students, 10,853 were excluded owing to limitations in their language proficiency or to mental or physical disabilities. Thus 5.37 percent of the potential student sample (the students enrolled in the eighth grade in the 1,052 NELS:88 schools from which usable student data were obtained) were excluded. Less than one half of one percent of the potential sample was excluded for reasons of physical or emotional disability (.41 percent), but 3.04 percent was excluded for reasons of mental disability, and 1.90 percent because of limitations in English proficiency.

Put another way, of the 10,853 excluded students, about 57 percent were excluded for mental disability, about 35 percent owing to language problems, and less than 8 percent because of physical or emotional disabilities. Because current characteristics and probable future educational outcomes for these groups may depart from the national norm, the exclusion factor should be taken into consideration in generalizing from the NELS:88 sample to eighth graders in the nation as a whole. This implication for estimation carries to future waves. For example, if the overall propensity to drop out between the eighth and tenth grades is twice as high for excluded students as for non-excluded students, the dropout figures derivable from the NELS:88 first follow-up (1990) study would underestimate early dropouts by about ten percent. (In point of fact, the 1988-90 status dropout rate derivable from the eligible NELS:88 sample representing about 94.6 percent of the cohort is between 6.0 and 6.1 percent, and from the expanded—eligible + ineligible—1988 sample representing [virtually] 100 percent of the cohort, 6.8 percent.)

Undercoverage of course affects the power of a study both to produce national estimates, and, yet more dramatically, to produce estimates for the particular group that is not fully covered.[12] Undercoverage, moreover, poses some special difficulties for the representativeness of a multi-cohort longitudinal study such as NELS:88.

In a school-based longitudinal survey such as NELS:88, baseline excluded students affect the representativeness of freshened grade cohorts in future waves. To achieve a thoroughly representative tenth grade (1990) and twelfth grade (1992) sample comparable to the High School and Beyond 1980 sophomore cohort (or, for 1992, the HS&B 1980 senior cohort and the base year of NLS-72), the NELS:88 follow-up samples must approximate those which would have come into being had a new baseline sample independently been drawn at either of the later

time points. In 1990 (and 1992) one must therefore freshen, to give "out of sequence" students (for example, in 1990, those tenth graders who were not in eighth grade in the spring of 1988) a chance of selection into the study. One must also accommodate excluded students whose eligibility status has changed, for they too (with the exception of those who fell out of sequence in the progression through grades) would potentially have been selected had a sample been independently drawn two years later, and must have a chance of selection if the representativeness and cross-cohort comparability of the follow-up sample is to be maintained. Thus, for example, if a base year student excluded because of a language barrier achieves the level of proficiency in English that is required for completing the NELS:88 instruments in 1990 or 1992, that student should have some chance of re-entering the sample.

A substantial subsample of the base year ineligibles was, accordingly, followed in 1990 and 1992, to reassess eligibility status and gather information about excluded students' demographic characteristics, educational paths, and life outcomes. Data on persistence in school to be obtained from this subsample has been used to derive an adjustment factor for national estimates of the eighth grade cohort's dropout rates between spring of 1988 and spring of 1990, and from 1988 and 1990 to 1992.

The base year ineligibles study largely compensates for population undercoverage. Small populations who remain outside the baseline sampling frame include students who are educated at home or in private tutorial settings, those who are in excluded categories of schools[13] and those who have dropped out of school before reaching the eighth grade.

Table 1 shows that by 1992, a substantial portion of the sample of base year ineligible students had been reclassified as eligible. Excluded students who were later classified as eligible were included in NELS:88 follow-up surveys.

Reclassifications reported in Table 1 reflect multiple phenomena. In some cases--and presumably this is particularly the case for the language exclusions--reclassification reflects change in the eligibility status of the sample member over time. In other cases, change represents the unreliability of exclusion judgments, particularly for exclusion reasons that are more open to interpretation (e.g., mental as opposed to physical handicaps) or that apply to individuals at the margin of the classification--different individuals were asked to assess eligibility at different points in time. Finally, some of the change registered in Table 1 reflects the fact that in the follow-ups we provided more detailed interpretation for the guidelines, so that the validity of exclusion judgments would be enhanced. All in all, however, if any individuals in the target population are to be subject to exclusion from the baseline of a longitudinal study, it is of some importance to reassess their eligibility over time, particularly, in a school-based survey, if the panel is to represent additional grade cohorts.

Table 1: 1992 Status Ns of 1988 Excluded Students

1988 reason for exclusion:	ELIG.	INELIG.	OUT OF SCOPE	N.A.	SAMPLING ERROR
language	125	22	25	30	23
physical	13	9	0	1	1
mental	166	140	5	25	16
unknown	30	15	2	10	16
TOTAL	334	186	32	66	56

* N.A. = status not ascertained.

2.2 Representativeness and New Grade Cohorts: Sample Freshening.

Pearson (1989) notes that a potential limitation of longitudinal samples is that they may provide estimates of the population from which they were originally drawn, but not of the current population. It is of interest to follow a sample of 1988 eighth graders. Nevertheless, an eighth grade panel two years later will not by itself provide a representative sample of the nation's high school sophomores, nor four years later a representative sample of seniors. Representative sophomore and senior samples are analytically desirable at all three levels of NELS:88 analysis *First*, it is desirable to be able to make cross-sectional generalizations about the nation's sophomores in 1990 and seniors in 1992. *Second*, it is desirable to be able to make intercohort comparisons between HS&B 1980 sophomores and 1990 NELS:88 sophomores; between NLS-72 (1972) and HS&B (1980) seniors and NELS:88 (1992) seniors; and between the transcript records of HS&B (1982), NAEP (1987 and 1990), and NELS:88 (1992) seniors. *Third*, it is desirable to be able to conduct longitudinal analyses of 1990 sophomores two, four, and more years later, and of 1992 seniors two and more years later.

Hence a major sampling objective of NELS:88 was to create a valid probability sample of students enrolled in tenth grade in the spring term of the 1989-1990 school year and of students enrolled in the twelfth grade in the spring term of the 1991-92 school year. This goal was achieved by a process we have termed "freshening." The 1990 freshening procedure was carried out in four steps:

1. For each school that contained at least one base year 10th grade student selected for interview in 1990, a complete alphabetical roster of all 10th grade students was obtained.

2. For each base year sample member, we examined the next student on the list; if the base year student was the last one listed on the roster, we examined the first student on the roster (that is, the roster was "circularized").

3. If the student who was examined was enrolled in the 8th grade in the U.S. in 1988, then the freshening process terminated. If the designated student was not enrolled in the 8th grade in the U.S. in 1988, then that student was selected into the freshened sample.

4. Whenever a student was added to the freshened sample in step 3, the next student on the roster was examined and step 3 was repeated. The sequence of steps 3 and 4 was repeated (adding more students to the freshened sample) until a student who was in the 8th grade in the U.S. in 1988 was reached on the roster.

At a given first follow-up school, the freshening process could yield zero, one, or more than one new sample member. Altogether, 1,229 new students were added to the tenth grade sample—on average, just less than one student per school.[14] This procedure was repeated in 1992, to generate a probability sample of the nation's high school seniors.

This freshening procedure is an essentially unbiased method[15] for producing a probability sample of students who were enrolled in the tenth grade in 1990 (or twelfth grade in 1992) but were not enrolled in the eighth grade in the U.S. in 1988. There is a very small bias introduced by the omission of eligible tenth (or twelfth) graders attending schools that included *no* students who were eighth graders in 1988. There is an additional small bias introduced by not freshening on the members of the sample of base year ineligibles. All other 1990 sophomores (or 1992 seniors) who qualify for the freshening sample have some chance of selection. This is because every student who was in the tenth grade in 1990 (or twelfth grade in 1992) but not in the eighth grade in 1988 is linked to exactly one student who was a 1988 eighth grader—this is the 1988 eighth grader who would immediately precede the candidate for the freshening sample on a circularized, alphabetical roster of tenth graders at the school. Because each 1988 eighth grader had a calculable, non-zero probability of selection into the base year and first follow-up samples, we can calculate the selection probabilities for all students eligible for the freshening sample. Thus, the freshening procedure produces a student sample that meets the criterion for a probability sample.

The NELS:88 school sample in 1990 and 1992—the schools to which 1988 eighth graders matriculated—was of course not a nationally representative sample of schools. However, for a select subset of schools, in order to provide a basis for studying school effects, feeder pattern information was collected so that tenth grade school selection probabilities could be approximated, and student samples augmented to make them robust and representative of the school's tenth grade class.[16]

2.3 Nonresponse Error as a Potential Source of Bias: Measures to Maximize Response Rates.

Cumulative nonresponse poses a special threat to longitudinal studies. Some individuals are missed in the baseline measurement, and may enter the study late. Other individuals may be lost, through mobility and the inability to locate them at a later date, or may cease to participate in the study. Still others may participate in the baseline, become temporarily out of scope by leaving the country or become nonrespondents by refusing to participate in the initial follow-up, then re-enter the study in a later follow-up. A longitudinal study must maximize the number of individuals who have data at all data points. Although weighting may help to adjust for nonresponse, the representativeness of the panel depends, in the final analysis, on maintaining high participation rates.

NELS:88 Response Rates. High response rates have been achieved by the study. In the NELS:88 base year (1988) 93.1 percent of selected eighth graders participated. In the NELS:88 first follow-up (1990), 93.9 percent of student and dropout sample members (19,264 of

20,524) took part. In the second follow-up, 90.7 percent of student and dropout sample members took part.

However, from the point of view of longitudinal analysis, a more critical statistic is the proportion of the sample with data at all time points (or, the proportion of baseline participants with data for all follow-ups). Of the 18,261 base year participants retained in the first follow-up, 17,424--or 95.4 percent--were successfully resurveyed. From this base of eighth grade cohort members with both (1988 and 1990) data points, 95.1 percent were resurveyed in the second follow-up.

Table 2A shows overall and subgroup results for the base year-first follow-up respondents for whom a reinterview was attempted in 1992. While, as noted above, around 95 percent were successfully resurveyed (that is, completed a student or dropout questionnaire) in 1992 and thus have data for all three waves, far fewer (72 percent) completed the cognitive test in all three rounds. Table 2B depicts the across-round questionnaire completion status of base year-first follow-up participants who were second follow-up *students*, and the likelihood that school contextual data was available for them for all three rounds. These tables show that completion rates were very similar across different school control types, urbanicity, region, and high and low minority enrollment, and that similar response rates were obtained for members of different racial and ethnic groups.

However, even with these high rates of success in baseline and follow-up data collection, the proportion of the 1988 eighth grade cohort in 1992 with all three data points drops to 84 percent (16,489 of 19,645) when all students missing one or more data point owing to base year, first or second follow-up nonresponse or any other source of sample attrition--being deceased, sample members who suffered grave impairments in the course of the study that did not permit them to be surveyed, individuals out of scope for either follow-up round by virtue of being outside the country--are factored in.

Overall, then, NELS:88 has achieved reasonably high student panel response rates. In addition, final weights have been adjusted for nonresponse, using nonresponse adjustment cells based upon combinations of classification values reflecting race, gender, and data collection status (e.g., dropout; in school in expected grade; in school in another grade; and so on).[17]

Means of Achieving High Response Rates. The means by which these high response rates were achieved may be concisely summarized. Most individuals changed schools, and many changed home addresses, between the base year and the follow-up surveys. About 99 percent of students were successfully traced between the base year and first follow-up, whereupon clusters of students were subsampled to reduce, for cost reasons, the number of high schools to be included in the study. The ability to successfully trace individuals was based upon extensive locating information collected in the base year from both students and parents. This locating information included name, address and telephone number for the student, each parent, and the family's closest relative or friend who did not live in the household. Eighth grade students were also asked to indicate what school they expected to be attending two years later. Tracing was carried out at two levels: first, it was ascertained if the sample member was at the expected school. If not, household information was used to locate the individual. In order to find base year nonrespondents (about 7 percent of the sample did not complete a 1988 student questionnaire and hence did not provide locating information), in addition to conventional survey locating sources, information about the schools matriculated to by the eighth grader's classmates was also utilized. Tracing procedures were repeated in the second follow-up, though between tenth and twelfth grade there is less dispersion to new schools and it was not necessary to further subsample students.

TABLE 2A

NELS:88 Second Follow-Up student survey results for Base Year– First Follow-Up panel participants

	Student/Dropout questionnaire (BY, F1 and F2) Completion rates		Student/Dropout cognitive test[a] (BY, F1 and F2) Completion rates	
	Weighted	Unweighted	Weighted	Unweighted
Total				
Participated	94.7	95.1	69.6	72.2
Selected	16,489[b]		11,902	
School type[c]	17,337		16,489	
Public				
Catholic	94.3	94.7	69.0	71.4
Other private	97.9	97.0	74.1	78.6
Urbanicity[c]	97.4	97.0	73.0	73.7
Urban				
Suburban	93.5	95.1	64.3	69.5
Rural	95.5	95.3	69.1	70.1
Region[c]	94.8	94.9	74.6	77.2
Northeast				
South	94.8	95.1	70.3	71.3
Midwest	94.1	94.5	68.2	73.1
West	95.7	96.0	74.9	76.4
Ethnicity	94.6	95.1	63.7	65.7
Asian/PI				
Hispanic	93.3	95.0	71.5	71.9
Black	93.1	94.4	63.9	65.5
White	92.4	92.6	59.6	67.0
Am. Indian	95.5	95.7	72.1	74.2
Refused/Missing[d]	94.1	91.3	64.8	64.0
Minority schools[c]	81.1	75.0	38.3	55.6
Schools with more than 19% minority students				
Schools with less than 19% minority students	92.2	93.5	55.1	59.3
	95.0	95.3	71.0	73.5

[a] Cognitive test coverage rate for each sample member who has completed a BY student questionnaire, F1 and F2 student/dropout questionnaire.

[b] Sample members who participated in the BY, F1 and F2.

[c] Refers to 8th grade schools.

[d] Refused/Missing refers only to the status of a sample member's ethnicity. It does not refer to student/dropout nonparticipants.

TABLE 2B

NELS:88 Second Follow-Up data collection results for Base Year – First Follow-Up panel participants

	Student questionnaire (BY, F1 and F2) Completion rates		School questionnaire[a] (BY, F1 and F2) Completion rates	
	Weighted	Unweighted	Weighted	Unweighted
Total	95.7	96.1	95.5	95.6
Participated	14,674[b]		13,182	
Selected	15,269		13,783	
School type[c]				
Public	95.4	95.8	95.8	95.7
Catholic	98.2	97.3	94.3	94.8
Other private	97.5	97.1	93.5	95.8
Urbanicity[c]				
Urban	94.4	96.4	93.7	94.7
Suburban	96.2	96.1	94.4	94.3
Rural	95.8	95.9	98.4	98.2
Region[c]				
Northeast	95.2	95.5	94.9	94.6
South	95.8	96.2	95.6	95.9
Midwest	96.2	96.5	97.3	97.8
West	95.5	96.0	93.1	93.2
Ethnicity				
Asian/PI	94.9	95.8	90.2	93.9
Hispanic	94.2	95.8	89.8	91.3
Black	94.3	95.0	95.1	95.3
White	96.2	96.4	96.5	96.5
Am. Indian	93.8	90.9	97.6	97.3
Refused/Missing[d]	74.2	72.7	100.0	100.0
Minority schools[c]				
Schools with more than 19% minority students	92.5	96.3	90.7	90.0
Schools with less than 19% minority students	96.0	94.4	96.0	96.2

[a] School questionnaire coverage rate for each student who completed a BY, F1, and F2 student questionnaire.

[b] Panel *students* only.

[c] Refers to 8th grade schools.

[d] Refused/Missing refers only to the status of a sample member's ethnicity. It does not refer to student nonparticipants.

In order to survey students, contractor (NORC) staff administered the survey forms at a date agreeable to the school. Make-up sessions were conducted for students who missed the initial survey session. Dropouts and chronic absentees were pursued outside school. Such individuals were invited to group sessions and provided reimbursement for their travel expenses, or were interviewed in their households, over the telephone or in person.

In rare instances, NELS:88 has made use of respondent fees. For example, some dropouts received a monetary incentive, as did some high burden teachers (teachers who had to rate an unusually high number of NELS:88 students such that their burden of questionnaire completion might be two hours or more). School coordinators were given a modest honorarium (normally $25) for assisting with survey activities (for example, supplying annotated rosters, arranging space, and so on), but neither schools nor students were ever paid for their participation

3. Other Sources of Survey Error

When all is said and done, it is the total variable error and bias of a survey estimate that is critical (see Kish, 1965; Andersen, Kasper, and Frankel, 1979; Groves, 1989). From the point of view of total survey error, our discussion thus far is incomplete. It may be useful to identify additional sources of survey error, though space limitations do not permit us to address them.

There are various "repeated measurement" problems in longitudinal surveys. One of these problems is that of panel effects.[18] We do not believe that problems associated with repeated measurements (such as remembering past responses to individual items) are likely to be a difficulty, both because of the sheer number of test and questionnaire items asked, and the two year intervals between data collections. However, participation in a longitudinal study in theory may influence the survey member's subsequent behavior or attitudes.

There are many sources of measurement error. The validity of responses to the NELS:88 eighth grade questionnaire items has been examined in Kaufman, Rasinski, Lee and West (1991), which compares parent and student reports. Transcript and student reports were compared for the HS&B data by Fetters, Stowe and Owings (1984). Psychometric issues in the base year tests are addressed in Rock and Pollack (1991) and in a forthcoming second follow-up psychometric report.

Our earlier discussion dealt with unit nonresponse as a problem of maintaining individual participation across rounds. However, school nonresponse in the base year, and item nonresponse across the survey instruments, also are important nonresponse issues. To the extent that students at noncooperating base year schools may have differed from students at cooperating schools, student level bias is introduced that persists through subsequent waves of observation. Base year school nonresponse is documented and analyzed in the *NELS:88 Base Year Sample Design Report*.

Item nonresponse rates and patterns are documented in the various NELS:88 user's manuals. In general, missing data have not been imputed in the NELS:88 dataset. Although item response rates in NELS:88 are generally high, item nonresponse propensities vary with student characteristics (e.g., race, gender, test quartile), and hence may be a source of bias.

Finally, our discussion has not dealt with the important consideration of sampling error. Design effects for NELS:88 are documented in the various user's manuals. In this respect, dispersion of the student sample after eighth grade has been both a blessing and a curse for NELS:88. The high costs of following dispersed students required that we subsample students in the first follow-up; subsampling increases design effects. At the same time, the general tendency in a longitudinal study is for design effects to decrease over time, as dispersion reduces the original clustering.

Bibliography

Andersen, R., Kasper, J., and Frankel, M.R. (1979). *Total Survey Error*. San Francisco: Jossey-Bass.

Babbie, E.R. (1973). *Survey Research Methods*. Belmont, CA: Wadsworth.

Bradburn, N.M., Rips, L.J., and Shevell, S.K. (1987). "Answering Autobiographical Questions: The impact of Memory and Inference on Surveys." *Science*, 239:157-161.

CBO (Congress of the United States, Congressional Budget Office). (1990). *Services of Support for Adolescent Mothers*. Washington, D.C.: Author.

Duncan, G.J., Hill, M.S., and Hoffman, S.D. (1988). "Welfare Dependence Within and Across Generations." *Science*. 239:467-471.

Fetters, W.B, Stowe, P., and Owings, J.A. (1984). *Quality of Responses of High School Students to Questionnaire Items*. Washington, D.C.: National Center for Education Statistics.

Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York: Wiley.

Ingels, S.J., et al. *NELS:88 Data File User's Manuals*. Base Year: Student, Parent, Teacher, School (NCES, 1990); First Follow Up: Student, Dropout, Teacher, School (NCES 1992); Second Follow-Up: Student, Dropout, Parent, Teacher, School, Transcript (NCES 1994).

Kasprzyk, D., Duncan, G., Kalton, G., & Singh, M.P., eds. (1989). *Panel Surveys*. New York: Wiley.

Kaufman, P., Rasinski, K.A., Lee, R., and West, J. (1991). *Quality of Responses of Eighth-Grade Students in NELS:88*. Washington, D.C.: National Center for Education Statistics.

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

Kish, L. (1987). *Statistical Design for Research*. New York: Wiley.

Lee, V.E., and Loeb, S. (1994). "Where Do Head Start Attendees End Up? One Reason Why Preschool Effects Are Not Sustained." Paper presented to the American Educational Research Association, Annual Meeting, New Orleans, Louisiana, April 7, 1994.

McGrew, K.S., Thurlow, M.L., and Spiegel, A.N. (1993). "An Investigation of the Exclusion of Students With Disabilities in National Data Collection Programs." vol.15. *Educational Evaluation and Policy Analysis*.

Pearson, R. W. (1989). "The Advantages and Disadvantages of Longitudinal Surveys." *Research in the Sociology of Education and Socialization*. vol. 8.

Rock, D.A. and Pollack, J.M. (1991). *Psychometric Report for the NELS:88 Base Year Test Battery*. Washington, D.C.: National Center for Education Statistics.

Rosenzweig, M.R., and Wolpin, K.I. (1993). "Maternal Expectations and Ex Post Rationalizations: The Usefulness of Survey Information on the Wantedness of Children." *Journal of Human Resources*, vol. XXVIII.

Spencer, B.D., Frankel, M.R., Ingels, S.J., Rasinski, K.A., and Tourangeau, R. (1990). *NELS:88 Base Year Sample Design Report*. Washington, D.C.: National Center for Education Statistics.

Spencer, B.D. (1991). "Eligibility/Exclusion Issues in the 1990 Trial State Assessment." In G. Bohrnstedt et al., *Assessing Student Achievement in the States: Background Studies for the Evaluation of the NAEP Trial State Assessment*. National Academy of Education. Stanford, California: Stanford University.

Spencer, B.D., and Foran, W. (1991). "Sampling Probabilities for Aggregations, With Application to NELS:88 and Other Education Longitudinal Surveys." *Journal of Educational Statistics*, 16(1).

END NOTES

1Longitudinal studies are prospective, in that they record new events for *individual* units under observation more or less as they occur. A repeated cross-sectional study can also be prospective, and measure change over time at the *group* level. A single (not repeated) cross-sectional study can measure change in *individual* units over time by assuming a retrospective focus--for example, by relying on individual memories to reconstruct an historical record of events and statuses. While retrospective studies are appropriate for many purposes, when used for other purposes there may be significant reliability problems. For a concise summary of issues concerning the reliability of retrospective reports, see Bradburn, Rips and Shevell (1987). For a useful comparison of prospective and retrospective studies, see Kish (1987) pp. 178-181. For a recent example (*ex post facto* reports of wantedness of children) of an analysis of the degree to which retrospectively-obtained survey data provides unbiased estimates, see Rosenzweig and Wolpin (1993).

2We take a representative sample to be a probability sample drawn, with known selection probabilities for sample units, from the target population.

3Groves (1989, chapter 3) provides a useful discussion of coverage error and its consequences.

4There are many ways to characterize sample designs that measure change over time, and the term "longitudinal" has both strict and looser usages. Kish (1987, Chapter 6) presents a useful typology of designs for covering time spans across populations, and Babbie (1973, pp.62-66) some standard terminology.

5For individuals interested in conducting trend analyses of NLS-72, HS&B and NELS:88 data, further information on content and design similarities and differences between these three studies is presented in the second follow-up student component data file user's manual. Comparison of sophomore cohort dropouts across HS&B and NELS:88 is discussed in the dropout component user's manual, while high school transcript comparisons (HS&B, NAEP 1987, NAEP 1990, NELS:88) are discussed in the transcript user's manual.

6For each cohort, the timing of the last follow-up assumes that the tentatively scheduled date for the fourth follow-up – 1997 – will hold.

7Repeated cross-sections compound sampling error. This is the case because a repeated cross-section is drawn two or more times; change measurement must contend with the fact that differences in multiple sample means will in part be a function of the sampling errors associated with each independent sample. In contrast, a longitudinal sample is drawn but once. However, for a freshened cohort study such as NELS:88, some sampling error may be associated with the freshening process. Hence when NELS:88 data are analyzed in the aggregate as a rolling cross-section, some of this advantage of a longitudinal design is lost.

8The 1988-90 dropout rate for the expanded (eligible + ineligible) NELS:88 eighth grade cohort was 6.8 percent for 1988-90. Excluding students who dropped out between 1988 and 1990 (or left the country), the dropout rate between 1990 and 1992 was 7.6 percent. However, the proportion of 1988 eighth graders who were dropouts in the spring of 1992 was 11.6 percent. (Of course, the number of sample members experiencing brief duration dropout spells or dropout *events* is even further undercounted by virtue of using a cohort status [spring to spring across two years] measurement.)

9On NLSY (the BLS National Longitudinal Survey of Youth which began in 1979), see CBO, 1990. On PSID (Panel Study of Income Dynamics, a nationally-representative sample of families, begun in 1968) results, see Duncan, Hill, and Hoffman, 1988.

10Researchers (see Lee and Loeb, 1994) have used the response to this retrospective item in conjunction with NELS:88 measures of school quality to inquire into whether Head Start participants are more likely than their peers to attend lower quality elementary/middle schools, a possibility that could in part explain why academic gains from Head Start may fade out over time.

11For further details of school-level exclusion, see Spencer, Frankel, Ingels, Rasinski, & Tourangeau, 1990, p.10.

12Recent investigations of the extent to which students with disabilities are allowed to participate in major national data collection programs suggest that 40-50 percent of students with disabilities are typically excluded from major assessments, though students with disabilities are included to a greater degree in data collections that do not require the completion of cognitive tests (McGrew, Thurlow, & Spiegel, 1993). Additional numbers of students are excluded from assessments or other state and national education data collection programs owing to language barriers to participation. For a parallel discussion based on the NAEP trial state assessments, see Spencer in Bohrnstedt, ed., 1991.

13According to Office of Special Education figures reported in the *Digest of Education Statistics, 1992*, Table 51, 5.5 percent of special education students receive services in separate schools or residential facilities, while .8 percent are in a homebound or hospital environment. Not all of these individuals are in graded programs. Separate facilities tend in particular to be available for comparatively rare populations such as individuals with severe visual or hearing impairments, and for emotionally disturbed students whose presence might impede regular classroom activities. Most students who are doubly physically disabled by being both deaf and blind are educated in special facilities.

14Some of these freshened students were dropped in the subsampling process either because they themselves were not included in the subsample or because the base year student to whom they were linked was not included. Some 1,043 students selected through the freshening procedure remained in the final first follow-up sample. In the second follow-up (1992), 244 students were added through freshening.

15See Kish (1965) for a discussion of the half-open interval procedure that underpins this approach.

16A strategy for estimating a school's selection probabilities under these circumstances is sketched in Spencer and Foran, 1991.

17Again, however, while weights can compensate for nonresponse by correcting errors in the population estimates for particular subgroups, they do not correct nonresponse *bias* within subgroups. For example, weighting can adjust for the fact that male eighth graders responded to NELS:88 at a lower rate than did their female classmates, but do not address bias that may be present if male responders and nonresponders differed in the very characteristics inquired into by the base year student questionnaire.

18Discussions of longitudinal conditioning or panel effects (also known as "time in sample bias" or "panel conditioning")–for example, whether strong effects potentially exist or could affect data quality–may be found in Kasprzyk, D., Duncan, G., Kalton, G., & Singh, M.P., eds. *Panel Surveys*, 1989 (New York: Wiley). See especially contributions by B. Bailar; D. Cantor; D. Holt; A. Silberstein and C. Jacobs; L. Corder and D. Horvitz; and J. Waterton and D. Lievesley.

DISCUSSION

Gary M. Shapiro
Abt Associates

The two papers in this session discuss the methodology of two of the most important longitudinal surveys ever conducted, and thus are of considerable interest. This session is on longitudinal surveys, but in fact the papers relate to two topics that are in the Statistical Policy Working Paper Series in addition to longitudinal surveys: nonresponse and survey coverage.

Both of these surveys have been very well designed. The sample design issues for the surveys have been carefully investigated, and are well documented both here and in earlier publications. Specific comments on each paper follow.

I. National Medical Expenditure Survey Paper by Steven Cohen

I will separately discuss the main topics covered in the paper. The first topic was the screening procedure and differential sample selection by demographic characteristics. The interviewers returned to screened-in addresses rather than follow individual persons. The paper provides a careful, detailed discussion of the consequences of following addresses, and contains a good discussion of the pro's and con's of following addresses vs. persons.

I have a couple of side observation on this topic. First, there were 722 units that were occupied at time of screening that became vacant by time of interview. However, there were only 448 vacant units that became occupied by time of interview. In theory, one would expect these two figures to be equal. The difference is probably statistically significant. This most likely signifies some coverage loss - either some of the 722 were not really vacant, or more than 448 of the vacants actually became occupied. I do not find this discrepancy to be alarmingly large, but it is evidence of a minor problem.

My second observation is on the use of screening. There were about 28,000 occupied addresses that were screened, with about 15,000 interviewed the next year. The smallest sampling rate in any demographic group was about 40%. This makes me wonder how cost effective the screening was. Fewer demographic categories and a simpler differential sampling scheme would quite possibly have been about as effective. It might have been preferable to undertake a very simple sub-sampling scheme at the time of interview instead of having entirely separate screening interviews. This has obvious implications for NMES-3.

I have no comments on the paper's second major topic, the return visits to households that were nonresponse for screener interviews, in an effort to improve response rates.

The third topic dealt with persons who were nonrespondents for only some of the waves. The paper discusses the characteristics of these people and estimation considerations for them. About 6.2 % of all participants missed at least one wave but not all waves, and about half of them were missing more than 1/3 of the data. The original paper states "When the level of partial response is low, it is often preferable to treat respondents as complete nonrespondents", as opposed to using imputation for missing waves. I invited audience members to discuss under what circumstances nonresponse adjustment should be done as opposed to imputation. One or two audience members expressed the view that imputation might have been preferable in this survey, and indicated some level of disagreement with the quotation above.

The fourth major topic was on the estimation for the institutional sample, when people move in and out of institutions during the year. I have no significant comments on this topic.

II. National Education Longitudinal Study of 1988(NELS:88) Paper by Steven Ingels and Jeffrey Owings

There are three main topics in this paper, two dealing with coverage issues and one dealing with nonresponse. I'm extremely pleased to see the emphasis here on coverage. Coverage is a major problem in many surveys, but it usually receives much less attention than nonresponse and many other less important topics.

The paper first discussed students that were excluded from the sample, with emphasis on disabled and non-English speaking student exclusions. These are reasonable exclusions that I think many survey practitioners would not have worried at all about excluding - I applaud the authors for their concern about the effects of the exclusions. I very much liked the paper's discussion of the effects of this undercoverage on survey results - more analyses of this type are needed.

The decision to follow students who were ineligible in 1988 in the 1990 and 1992 interviews is exemplary. I think most survey practitioners would probably not have undertaken the expense. It's unfortunate that except for the language exclusion category, the follow-up was not very effective.

The second topic of the paper was on sample freshening, to assure that all tenth graders in 1990 were covered in the survey. This again is a coverage issue that many surveys would not have worried about, and it is a real pleasure to see this level of care in survey implementation. As a side note, the survey estimated the number of schools attended by tenth graders in 1990 using complex methodology in Spencer and Foran(1991). I believe a better approach to this type of estimation issue is one used in the Survey of Income and Program Participation for estimating the number of households when following individuals (see Huang,1984, and Ernst,1989.)

The third topic dealt with two nonresponse issues. I have no comments on this section.

The last section of the paper briefly mentions several other sources of error, including a statement that item nonresponse was rarely imputed for. I note that even when there is not explicit imputation in a published table, there is implicit imputation in the sense that a reader generally assumes that nonrespondents are distributed like respondents. Thus, even a very crude explicit imputation is usually better than no imputation at all.

REFERENCES

Ernst, L. (1989), "Weighting for Longitudinal Household and Family Estimates", in "Panel Surveys", edited by D. Kasprzyk, G. Duncan, G. Kalton, and M.P. Singh, pp. 139-159, Wiley and Sons.

Huang, H. (1984), "Obtaining Cross-sectional Estimates for a Longitudinal Survey: Experience of ISDP", Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 670-675.

Spencer, B. and Foran, W. (1991), "Sampling Probabilities for Aggregations, With Application to NES:88 and Other Educational Longitudinal Surveys", Journal of Educational Statistics, Vol. 16, No.1, pp. 21-33.

DISCUSSION

James M. Lepkowski
University of Michigan

Panel surveys are used to address many scientific and policy issues. The primary focus of many panel surveys is to assess change over time and examine causal mechanisms in various phenomenon. A wide range of designs are employed in these panel surveys, reflecting the diverse uses to which panel survey data may be put.

The presentation of papers on the design of two panel surveys in this session provides an opportunity to compare and contrast features of panel survey design. The two panel surveys described in these papers have very different topics, populations, sampling frames, and sources of survey error. The National Education Longitudinal Study of 1988 (NELS 88) is designed to provide data on education for a cohort of eighth graders followed over a six year period, sampling students from schools. The National Medical Expenditure Survey (NMES) is a survey of health and health care utilization and coverage among the U.S. civilian non-institutionalized population and those residing in nursing homes. The NMES uses a national area sample frame supplemented with a national nursing home list. There are similarities in design features between these surveys, but there are also many interesting differences. Comparing and contrasting these surveys illustrates the value and flexibility of panel survey design.

The two surveys, for example, have different statistical estimation purposes. The NELS 88 provides data that can be used to create cross-sectional estimates at each data collection time period, inter-cohort comparisons over time, and individual change through the repeated measurement of the same individuals. These are typical goals of many panel survey designs, attempting to answer questions about a single time period through cross-sectional estimates and about changes over time in a group or among individuals of the group. The NMES uses the panel survey design for a different purpose, to provide precise and accurate measurement of health and health care utilization and coverage through the accumulation of retrospective reports from a panel of respondents. By limiting the time between observations, NMES reduces the size of recall errors. There is little if, any interest, in assessing change at an individual or group level in the NMES.

Different purposes have lead to different design features. NELS 88 uses a longer period between interviews, following the cohort of eighth graders every two years. Characteristics measured in the NELS 88 are not expected to change substantially in a shorter interval, and thus more frequent data collection does not appear necessary. NMES interviews reporting units (essentially households) every three or four months over a 16 month period.

This frequent interviewing schedule allows NMES to measure continuously the rapidly changing health characteristics of the sample. The longer period between interviews would pose a larger problem for the NELS 88 if it were not for the fact that a large share of the NELS 88 sample each interview is students in school being followed. Information on contact persons is collected to assist in the tracking of students who leave school or move to another location and school. The NMES also collects contact information to assist tracking efforts, but NMES reporting unit members appear to be more difficult to track than the NELS 88 student sample.

An important methodological difference between the two surveys is the need for screening and sampling at varying rates subgroups of the population. NELS 88 does not need to sample any particular subgroups at higher rates than others. NMES attempted to sample poorer persons at higher rates through an interview conducted before the primary data collection began. A substantial share of the NMES paper is devoted to an examination of the screening procedures and their effectiveness.

The two surveys address the issue of the changing nature of the population that the sample represents quite differently. In panel survey design, "representation" of the population by the sample is hindered by loss due to non-response and failure to recruit into the sample persons who are recently joined the population such as immigrants. The NELS 88 sample is "refreshed" every interview period by a selection of additional students who could not have been selected at any of the previous rounds of data collection because they were not members of the U.S. eighth grade student population in 1988. The refreshing process is designed to improve the coverage of the NELS 88 sample of students who were eighth graders in 1988 at any given future interview round. The NMES covers a much shorter time period during which fewer changes to the population that could affect the quality of the NMES data occur. For instance, immigration changes to the U.S. civilian population is not expected to be an important departure from complete coverage during the time period of the NMES.

The NELS 88 and NMES are both subject to non-response, and both surveys make considerable effort to reduce non-response rates. The NELS 88 faced difficulty recruiting school districts and schools, the primary sampling units in the selection, into the sample. When a sample school district or school refused to participate, NELS 88 substituted a school district or school as similar as possible to the non-participating school. Student participation within cooperating schools was quite high, with very low non-response at the initial round and all subsequent rounds of data collection. High response rates were maintained despite the difficulties of tracking school drop-outs. NMES did not lose any primary sampling units to non-response in the household sample portion of the sample, although it did experience losses of reporting units and reporting unit members at first and subsequent interviews. NMES also experienced loss of primary sampling units

in the nursing home survey portion of the sample, as well as difficulty obtaining interview data for persons who were extremely ill.

Many panel surveys compensate for non-response losses through weighting for unit non-response and imputation for item non-response. The NELS 88 analyzed first and subsequent round non-response, but it is not clear from the paper if non-response weights are employed. No imputation has been done to compensate for item non-response rates, which are reported to be low. The NMES discards persons from the sample who provided reports for less than one-third of the target year. Non-response compensation weights are employed to adjust for the these and all other persons who failed to respond to the survey adequately. As in the NELS 88, imputation is not employed to compensate for item non-response. As a result, both surveys implicitly impute for missing items by using available data, effectively an imputation of the mean values of the responses for the missing items.

The NELS 88 and NMES are examples of well-designed and carefully executed panel surveys. Since both have different goals and topics, the panel designs employed by each are quite distinct. Those interested in understanding more fully the features and complexities of panel survey design can learn much from a reading of these two papers.

Reports Available in the Statistical Policy Working Paper Series

1. Report on Statistics for Allocation of Funds (Available through NTIS Document Sales, PB86-211521/AS)
2. Report on Statistical Disclosure and Disclosure-Avoidance Techniques (NTIS Document sales, PB86-211539/AS)
3. An Error Profile: Employment as Measures by the Current Population Survey (NTIS Document Sales PB86-214269/AS)
4. Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics (NTIS Document Sales, PB86-211547/AS)
5. Report on Exact and Statistical Matching Techniques (NTIS Document Sales, PB86-215829/AS)
6. Report on Statistical Uses of Administrative Records (NTIS Document Sales, PB86-214285/AS)
7. An Interagency review of Time-Series Revision Policies (NTIS Document Sales, PB86-232451/AS)
8. Statistical Interagency Agreements (NTIS Documents Sales, PB86-230570/AS)
9. Contracting for Surveys (NTIS Documents Sales, PB83-233148)
10. Approaches to Developing Questionnaires (NTIS Document Sales, PB84-105055/AS)
11. A Review of Industry Coding Systems (NTIS Document Sales, PB84-135276)
12. The Role of Telephone Data Collection in Federal Statistics (NTIS Document Sales, PB85-105971)
13. Federal Longitudinal Surveys (NTIS Documents Sales, PB86-139730)
14. Workshop on Statistical Uses of Microcomputers in Federal Agencies (NTIS Document Sales, PB87-166393)
15. Quality on Establishment Surveys (NTIS Document Sales, PB88-232921)
16. A Comparative Study of Reporting Units in Selected Employer Data Systems (NTIS Document Sales, PB90-205238)
17. Survey Coverage (NTIS Document Sales, PB90-205246)
18. Data Editing in Federal Statistical Agencies (NTIS Document Sales, PB90-205253)
19. Computer Assisted Survey Information Collection (NTIS Document Sales, PB90-205261)
20. Seminar on the Quality of Federal Data (NTIS Document Sales, PB91-142414)
21. Indirect Estimators in Federal Programs (NTIS Document Sales, PB93-209294)
22. Report on Statistical Disclosure Limitation Methodology (NTIS Document Sales, PB94-165305)
23. Seminar on New Directions in Statistical Methodology (NTIS Document Sales, PB95-182978)

Copies of these working papers may be ordered from NTIS Document Sales, 5285 Port Royal Road, Springfield, VA 22161 (703)487-4650

Part 3

Session 9

COGNITIVE TESTING AND SELF-ADMINISTERED QUESTIONNAIRES

Literacy Limitations and Solutions for Self-Administered Questionnaires

Judith T. Lessler and James M. O'Reilly
Battelle Memorial Institute

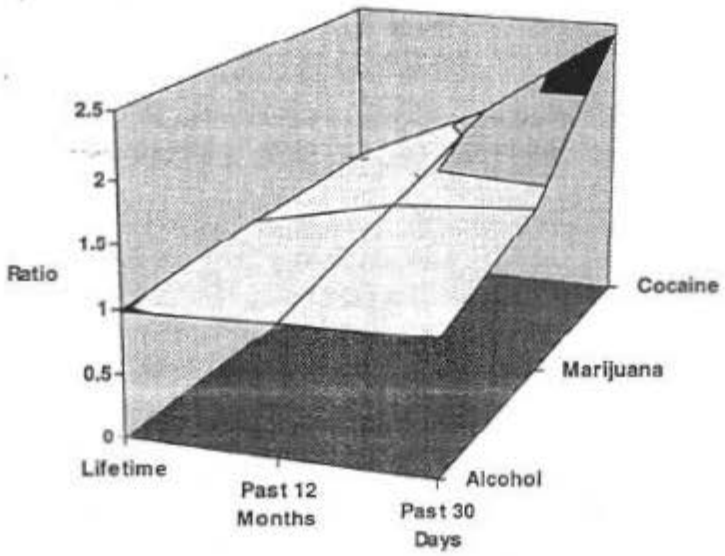
Introduction. The self-administered paper questionnaire is a standard method for asking questions on sensitive subjects. Yet as much as one-fifth of the adult population of the U.S. has levels of literacy which may make using the typical SAQ futile. This paper reviews the research on the efficacy of self-administered questioning on sensitive subjects. Then we will discuss how low literacy or other cognitive burdens can limit the effectiveness of self-administration which require reading.

New computer technology, called Audio Computer Assisted Self-Interviewing (ACASI), now makes is possible to conduct self-administered interviews in which the text on the screen is accompanied by a high quality voice recording played over headphones. The first major field test of the technology was in the 500-person 1993 National Survey of Family Growth Pretest. We describe how ACASI was implemented in the study and the impact on abortion reporting and respondents' reactions to the technology.

Methods of Interviewing on Sensitive Subjects. Research has generally shown that more private methods of interviewing yield higher reports of sensitive behaviors (Bradburn, 1983; Miller, Turner, and Moses, 1990, Ch. 6; Catania, et al., 1990, and Schwarz et al., 1991). For example, Hay (1990) found differences in reported consumption of alcoholic beverages and cigarette use in a study of some 1500 students in grades 2 through 12 who were randomly assigned to receive either a SAQ or a personal interview. The differences were 74 versus 63 percent for over use of alcohol and 38 versus 30 percent for use of cigarettes. Turner, Lessler, and Devore (1992) in a large-scale field experiment, in which 3,200 respondents were randomly assigned to either an interviewer or self-administered questionnaire found that the difference between the two modes of data collection increased as the sensitivity of the behavior increased. Exhibit A shows the ratio between the proportion of SAQ respondents reporting a given behavior to the proportion of respondents reporting that behavior when the interviewer administered the questions. The exhibit displays the results for three time periods (lifetime, last 12 months, and last 30 days) and three types of drugs (alcohol, cocaine, and marijuana).

Exhibit A. Ratio of Prevalence Estimates from SAQ and Interviewer-Administered Items

Drug Type	Lifetime	Past 12 Months	Past 30 Days
Alcohol	**0.99**	**1.04**	**1.06**
Marijuana	**1.05**	**1.30**	**1.38**
Cocaine	**1.06**	**1.58**	**2.40**



Examining this table, we note that as the sensitivity of the characteristic increases from alcohol to marijuana to cocaine, there is a concomitant increase in the superiority of the self-administered format relative to the interviewer-administered questions.

Cognitive Aspects of Conventual SAQs. Thus, self-administered questionnaires can have a positive impact on data quality because of the increased privacy. In addition, SAQs allow respondents to control the pace of the interview, and no additional variance is introduced by the interviewers.[1] However, conventional SAQs do have drawbacks. First and foremost, they require that the respondent can read.[2] In addition, the respondents must complete a number of the questionnaire administration tasks such as finding and reading instructions, implementing skip patterns, and marking answers. They are prone to the same types of errors that are seen in interviewer administered questionnaires—missing, out-of-range, and inconsistent answers. Lessler and Holt (1987) found that some respondents who could read the questions had difficulty understanding the conventions concerning recording of answers and movement through forms.

Cognitive testing of self-administered has noted problems in each of these areas:

Reading problems:

- Complete inability to read the questions
- Failure to understand specific terms or phrases

---

1 The presence of high levels of interviewer variance in the decennial census was one of the motivations for adopting a mailout-mailback self administered method for the census beginning in 1960.

2 The National Adult Literacy Survey (NALS) was conducted in 1992 using a nationally represented sample of 13,600 persons aged 16 and older. Literacy was measured in terms of five proficiency levels on three scales—prose, document and quantitative. The survey found that the percentage of adults in the lowest level of proficiency was 21 percent for prose literacy, 23% in document literacy, and 22% in quantitative literacy.

The lowest level of prose literacy is described as "Most of the tasks in this level require the reader to read relatively short text to locate a single piece of information which is identical or synonymous with the information given in the question or directive" (pg. 74). For document literacy the lowest level means: "Tasks in this level tend to require the reader either to locate a piece of information based on a literal match or to enter information from personal knowledge onto a document. Little, if any distracting information is present" (pg. 85). Quantitative literacy at the lowest level means: "Tasks in this level require readers to perform single, relatively simple arithmetic operations, such as addition. The numbers to be used are provided and the arithmetic operation to be performed is specified" (pg. 94).

National Center for Education Statistics, 1993. Adult Literacy in America: A First Look at the Results of the National Adult Literacy Survey. Washington, DC: U.S. Department of Education.

- Use of a time consuming two-step strategy for reading questions
  First the questions are read to decode the words and then they are read a second time to get the meaning of the sentences
  Complete inability to read the questions
- Not reading all of the questions or response categories in order to reduce the reading task

Questionnaire administration problems:

- Ignoring or neglecting to read instructions
- Difficulty finding the instructions
- Physical difficulties with marking answers that require filling circles for mark-sense forms or writing in small spaces
- Difficulty understanding or failure to follow skip instruc
- Missing questions
- Writing in illegible or out-of-range responses
- Failure to follow marking instructions
- Idiosyncratic response or marking conventions

The result of these difficulties is that researchers using SAQs typically simplify the questionnaires and avoid contingent questioning. Contingent questioning is avoided for two reasons. One is to reduce the chance that data is lost because of the errors that respondents make through incorrect implementation of skip patterns and the second is to increase the privacy of responses. For example, the National Household Survey on Drug Abuse uses SAQs to ask questions on use of alcohol, misuse of prescription drugs, and use of illegal drugs. In that survey, interviews are conducted in the respondent's home. Some of the questions are interviewer administered. On the more sensitive topics respondents are instructed how to complete the SAQs. Respondents are required to mark an answer for every question in these sections in order to (1) increase their privacy, (2) prevent errors in implementing skip instructions, and (3) eliminate the tendency for respondents to mark no on gate questions on use of a particular substance in order to reduce the response burden of answering detailed questions about the drug. It is believed that if respondents are allowed to skip answers, they will realize that interviewers are able to distinguish those who were and were not drug users and, as a consequence, be less truthful.

ACASI Technology. Audio computer assisted self-interviewing (ACASI) has been developed to overcome some of the difficulties associated with the response to self-administered questionnaires. When a computer-assisted self-administered interview (CASI) is used, the computer can take care of the "housekeeping" or administrative tasks for the respondent. By adding simultaneous audio renditions of each question and instruction aloud, ACASI can remove the literacy barriers to self-administration.

In CASI respondents read the questions as they appear on the screen and enter their

answers with the keyboard (or some other device). In Audio-CASI, an audio box is attached to the computer and respondents put on headphones and listen to the answers as they are displayed on the screen. Respondents have the option of turning off the screen so that people coming into the room cannot read the questions, or turning off the sound if they can read faster than the questions are spoken, or keeping both the sound and video on as they answer the questions. Respondents can interrupt the question while it is in process.

Comparisons of CASI with personal interviews have noted findings similar to those cited above for the comparison of SAQs to interviewer administered questionnaires. Waterton and Duffy (1984) compared reports of alcohol consumption under CASI and personal interviews. Overall, reports of alcohol consumption were 30 percent higher under the CASI procedure, and reports of liquor consumption were 58 percent higher. This may understate the potential gains because in this study respondents were first asked if they had consumed any alcoholic beverages in the past seven days by an interviewer. Only those respondents who indicated that they had done so received the CASI interview.[3] Several recent studies comparing CASI to personal interviews in clinical settings have also noted the superiority of this method. Locke (1992) found significant differences between the reporting of HIV-risk behaviors when CASI was used to administered questions to donors at an American Red Cross donor center (4.4 percent versus 0.3 percent in the traditional interview procedure). Robinson and West (1992) compared reporting of symptoms in a genito-urinary clinic using CASI, SAQs, and physician interviews. They found that more symptoms were reported by computer than by paper, and both found more symptoms in physician interviews. Levine, Ancill, and Roberts (1992) found that patients who had been admitted to a hospital after harming themselves were more likely to report suicidal ideation in a computer interview than to a physician. The CASI version of the Diagnostic Interview Schedule (DIS) yielded diagnostic information consistent with the traditional interviewer administered DIS, and patients considered the computer contact to be less embarrassing (Erdman et al., 1992). A computer interview with sex offenders yielded large numbers of previously undetected crimes (Weinrott and Saylor, 1991), and a comparison of clinician and computer interviews directed at identifying obsessive compulsive disorders found that the two methods were equally good at distinguishing those with the disorder and that patients showed no preference of the clinician interviews (Rosenfield et al., 1992).

Formal comparisons of Audio-CASI versus other modes are just now being conducted. O'Reilly, et al. (in press) compared paper SAQs, CASI, and Audio-CASI in a small scale experiment designed to assess the potential for the technology. Subjects answered questions on drug use, sexual behaviors, and income. A greco-

---

3 In the literature, this study is often reported as a CAPI study. It was actually a CASI study in which computers were taken into the homes of respondents and asked to enter their responses on selected questions while the interviewer stood in a part of the room that did not permit observation of the respondent's answers.

latin square design was used to assign subjects to one of three interviewing modes for each topic producing an experiment that was fully balanced across mode and content. For eight of nine rating scales comparing these modes, respondents reported a preference for one of the two CASI methods. Sample sizes were quite small in this study (n=40); however, O'Reilly, et al. found that the CASI methods tended to produce significantly more reports of marijuana and cocaine use. Few differences in sexual behaviors were found.

Respondents were asked which method they thought was better for nine facets:

(1) Liked best
(2) Best for asking sensitive questions
(3) Easiest to change answers
(4) Most interest
(5) Easiest to use
(6) Best for getting honest answers
(7) Best for privacy after interview
(8) Best for privacy interview, and
(9) Overall preference

For all but number 3, the two CASI methods, audio and video-only, were rated significantly better. ACASI was rated consistently higher than video-only CASI. However the difference was significant for three items: overall preference, interest, and ease of use.

National Survey of Family Growth (NSFG) Pretest. Under funding from the National Center for Health Statistics (NCHS), scientists at NCHS, Battelle and the Research Triangle Institute (RTI) collaborated in a formal field experiment that compared abortion reporting under three different interviewing conditions. Respondents were randomly assigned to receive either an in-home CAPI interview only, an in-home CAPI interview followed by a Audio-CASI interview that asked additional questions about abortions, or an interview at a neutral site away from the respondent's home. The respondents in the Audio-CASI treatment were first asked to report their abortions to the interviewer during a section of the CAPI interview that asked about the outcome of each pregnancy that they ever had. The question asked:

Now I'd like to ask some questions about your Nth pregnancy.

Please look at Card B-1. Thinking about your Nth pregnancy, in which of the ways shown on Card B-1 did the pregnancy end?

(READ LIST. CODE ALL THAT APPLY.)

Miscarriage?	(Occurs naturally, during the first 6 months of pregnancy)
Stillbirth?	(Baby born dead after 7 or more months of pregnancy)
Abortion?	(Induced during the first 6 months of pregnancy; include D&C, vacuum extraction, suction, and saline injections)
Ectopic Pregnancy?	(Occurs outside the uterus or womb)

Livebirth by Cesarean section?
Livebirth by vaginal delivery? (Includes delivery through natural or induced labor)

At the end of the interview, respondents were trained in the Audio-CASI procedures and were asked additional questions on abortion.

Prior to the field experiment, the Audio-CASI interview was tested in the cognitive laboratory. Respondents were brought into the laboratory and were first asked to completely answer a series of questions from the NSFG. Following this, the use of the Audio-CASI implementation was explained to the respondent, and she answered questions using the computer herself. Since the questions on abortion were considered to be highly sensitive, we did not ask respondents to think-aloud during laboratory testing. Instead, the interviewer stood across the room from the respondent and was asked to describe what the respondent was doing as the respondent listened and answered questions. Thus, the respondent reported things like:

- She is reading the first question[4]
- I am putting in my answer
- I made a mistake, and I am backing up
- She is reading the next question
- I interrupted her since I already read the question
- I do not know what to do now

The first round of testing revealed that some respondents needed help learning how to enter their answers. Thus, a training interview was constructed that contained questions that were not on the interview and were not sensitive. The interviewer went over these questions with the respondent who then completed the rest of the interview on her own. The field experiment included a comparison of ACASI, in-home CAPI, and out-of-home CAPI. It was hypothesized that the willingness of women to report sensitive informaton would be increased if they were interviewed outside of their homes because in prior rounds of the survey, respondents had indicated that one of their concerns was that family members would overhear their responses.[5]

Abortion Reporting. Prior rounds of the NSFG identified significant underreporting of abortion (Jones and Forrest, 1992). Exhibit B compares the results from the Audio-CASI question on whether or not the woman had ever had an abortion and both the pilot questions and pregnancy outcome questions in Section B. There was one refusal to the Audio-CASI question on whether the woman had ever had an abortion in her

---

4 Respondents refereed to the computer as "she" because the recordings were done by a woman.

5 An incentive experiment was also included. The out of home respondents were paid $40.00 and the in home respondents received either no incentive or a $20.00 incentive.

lifetime so that there are 177 rather than 178 respondents in this second set of Audio-CASI tables. We note that six additional women reported having had an abortion at some time in their life in the Audio-CASI interview. The six additional women who reported an abortion represents a 14 percent increase in the number of women reporting ever having had an abortion.

Exhibit B. Relationship of Abortion Reporting in the Pregnancy Outcome Section and to Abortion Reporting in the ACASI Interview National Survey of Family Growth -- Cycle V Pretest

	Abortion reported as a birth outcome		
ACASI: Ever had an abortion	Yes	No	
Yes	42	6	48
No	0	129	129
	42	135	177

Exhibit C shows detailed information on abortion reporting by site of interview, incentive, and type of interview. Two series of numbers are shown for the ACASI respondents—the number of abortions that they reported in answer to interviewer questions in Section B of the interview and the number reported in subsequent ACASI interview. Finally, in Exhibit D we show the results on the number of abortions reported in Section B and the Audio-CASI interview for those 178 respondents who completed the Audio-CASI interview. Note, in this table the following:

> Women who had reported an abortion in Section B reported additional abortions in the Audio-CASI interview.
>
> All of the differences in numbers of abortions reported are above the main diagonal indicating that the different numbers of abortions reported in the Audio-CASI are probably not due to random error.

Exhibit C: Distribution of the Number of Abortions Reported in Section By Treatment and Incentive

			In Home						A-CASI						Neutral	
			$0		$20		Total		$0		$20		Total		$40/Total	
# of abortions	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%
0	380	77.1	83	86.5	56	77.8	139	82.7	78	79.6	58	72.5	136	76.4	105	71.4
1	79	16.0	10	10.4	12	16.7	22	13.1	16	16.3	11	13.8	27	15.2	30	20.4
2	24	4.9	1	1.0	4	5.6	5	3.0	3	3.1	9	11.3	12	6.7	7	4.8
3	9	1.8	2	2.1	0	0.0	2	1.2	1	1.0	1	1.3	2	1.1	5	3.4
4	1	0.2	0	0.0	0	0.0	0	0.0	0	0.0	1	1.3	1	0.6	0	0.0
Total	493	100	96	100	72	100	168	100	98	100	80	100	178	100	147	100
One or more abortions reported in B			13	13.5	16	22.2	29	17.3	20	20.4	22	27.5	42	23.6	42	28.6
One or more abortions reported in A-CASI									24	24.5	24	30.4*	49	27.1*		

* The proportions of A-CASI $20 incentives is based on 79 cases and total for A-CASI is based on 177 cases.

Exhibit D: Number of Abortions Reported in Section B and in ACASI

	ACASI Section						
Section B	0	1	2	3	4	5	DK
0	129	4	1	0	0	0	1
1	0	24	2	1	0	0	0
2	0	0	11	1	0	0	0
3	0	0	0	1	0	1	0
4	0	0	0	0	1	0	0

We also fit a series of logistic regression models to determine if there were significant differences due to interviewing conditions. As independent variables, we included the type of interview (CAPI only, Audio-CASI, or neutral site), incentive for in-home interviews (none or $20), race/ethnicity (Hispanic, black, non-Hispanic-non-black), marital status (married, not married), income (unknown, greater than $20,000, or other), and age. We used a stepwise selection procedure in which an independent variable that was significant at the 0.15 level was added to the model. Exhibit E summarizes the results.

Exhibit E. Analysis of the Impact of Characteristics of Women and Interview Conditions on Abortion Reporting
National Survey of Family Growth -- Cycle V Pretest

	Parameter estimate	Standard error	Probability (Chi-square)	Odds Ratio
Intercept	-2.52	0.49	0.0001	1.081
Incentive - 20	0.38	0.27	0.1348	1.488
Married	-0.34	0.23	0.1428	0.714
Age	0.03	0.01	0.0264	1.033
Audio-CASI	0.54	0.27	0.0419	1.723
Neutral site	0.83	0.31	0.00672	2.294

Based on these results, we concluded that both the neutral site and the Audio-CASI increases the number of women who report that they ever had an abortions.[6]

Respondent Attitudes. We also asked respondents who received the Audio-CASI interview their attitudes toward the alternative methods of reporting abortion. Exhibit F presents the results.

[6] We also examined the reporting of the number of abortions and found that given a woman had reported an abortion, there were not significant differences in the number of abortions reported.

Exhibit F. Respondents' Attitudes Toward Methods of Reporting Abortion Among Women who Received the Audio-CASI Interview National Survey of Family Growth -- Cycle V Pretest

Response	Percent respondents
How do you rate telling the interviewers your answers to questions on abortion?	
poor	15.2
fair	20.3
good	30.5
very good	17.5
excellent	16.4
How do you rate using the computer and earphones to answer questions on abortion?	
poor	2.8
fair	8.5
good	17.5
very good	26.0
excellent	45.2
Which method of answering questions on abortion is the most private?	
earphones and computer	62.7
no difference	32.2
telling the interviewer	4.5
Don't know	0.6
Which method do you recommend for the main study?	
Interviewer	16.9
Computer	58.2
Do not ask about abortion	2.8
Does not matter	22.0

In general, these women recommended the Audio-CASI procedure for abortion reporting.

Description and Demonstration of the Audio-CASI System. The ACASI system used has the following features:

- Implements a full range of audio functions so that audio self-interviewing can offer as many capabilities as interviewer-administered systems
- Runs on a powerful, existing CAI development platform
- Uses MS-DOS operating system

From an implementation and operational point of view, the key requirement is the second—that the audio system be built as an extension of an existing CAI

development system. It is not difficult to build PC systems which can generate sound through digital audio devices, display questions, and record answers. However, to have a system which can conduct a complex questionnaire with integrated audio is much more difficult. One especially important requirement is the ability to allow the user to backup easily, correct a previous entry, and be directed forward following a route appropriate for the latest set of responses. If the underlying CAI platform is not robust and widely used in complex applications, then the stability of the ACASI application during interviews is likely to be problematic.

The system used the Blaise CAI system as its base. The Blaise system is a product of the state statistical agency of the Netherlands—Statistics Netherlands and is widely used across Europe by government statistical agencies for computer-assisted personal interviewing, telephone interviewing, data entry and date editing. Questionnaires are programmed in the Blaise CAI language by defining the questions, their answer choices, and the logic of the questioning, including tailored text fills and consistency checks. Blaise then compiles the questionnaire code into a executable DOS application, automatically handling the question administration, screen and keyboard control, range and consistency checks, data management and navigation through the questions.

The audio capability is implemented through a background DOS process which the Blaise instrument triggers as each question is displayed on the screen. This process interprets commands specifying the recorded digital audio files to play in order to duplicate in audio just what is displayed on the screen.

The hardware for the audio system is a small, one-pound external analog-digital box that is connected to the notebook PC by two cables and headphones. The audio quality in the system is quite high. The items are a digitally recorded human voice—not synthesized. The system is both very flexible and fast. It has the capability of rendering questions with variable components. For example, in the NSFG ACASI instrument, when a women said she had had an abortion, she was asked when it happened and how many weeks pregnant she was at the time. Then the following series of questions were asked:

> Based on this, this pregnancy began around [MONTH AND YEAR OF CONCEPTION].
>
> So we can understand how well birth control methods work, I would like for you to tell me what methods of birth control you were using—if you used any—during the three months before this pregnancy began.
>
> Were you using any method of birth control in [MONTH AND YEAR PRECEDING CONCEPTION]?
>
> As I read the methods, please press 1 for YES if you used that method in [MONTH AND YEAR PRECEDING CONCEPTION].

Then the woman is asked for each of the three months preceding conception, whether she used each of the birth control methods she had reported ever having used earlier in the CAPI section. In this question series, the ACASI system must be able to generate audio questions which can vary on the month and year and among 19 possible contraceptive methods. The system was able to instantly concatenate and play the appropriate audio files to duplicate the screen text properly.

DEMONSTRATION OF THE A-CASI SYSTEM WILL OCCUR HERE

REFERENCES

Bradburn, Norman M. (1983) "Response Effects" in P. Rossi, J. Wright, and A. Anderson (eds.) Handbook of Survey Research, New York: Academic Press.

Catania, J.A., Gibson, D.R., Chitwood, D.D., and Coates, T.J. (1990). *Methodological Problems in AIDS Behavioral Research: Influences on Measurement Error and Participation Bias in Studies of Sexual Behavior*. Psychological bulletin (1990), vol. 108, No. 3, 339-362 by American Psychological Associates, Inc.

Erdman, H.P., Klein, M.H., Greist, J.H., Skare, S.S., Husted, J.J., Robins, L.N., Helzer, J.E., Goldring, E., Hamburger, M., & Miller, T.P. (1992). *A Comparison of two computer-administered versions of NIMH Diagnostic Interview Schedule*. Journal of Psychiatric Research, 26(1), 85-95.

Forsyth, Barbara, Lessler, J.T., and Hubbard, Michael, (1992). *Cognitive Evaluation of the Questionnaire*, Survey measurement of Drug Use: Methodological Studies, National Institution Drug Abuse.

Hay, David A. (1990) "Does the Method Matter on Sensitive Survey Topics," Survey Methodology, Vol. 16, No. 1.

Jones, Elise F. and Jacqueline Darroch Forrest (1992) "Underreporting of Abortion in Surveys of US Women: 1976 to 1988," Demography, Vol.29, No.1.

Lessler, Judith and Mimi Holt (1987) "Using Response Protocols to Identify Problems in the US. Census Long Form." Proceedings of the Survey Research Methods Section, American Statistical Association.

Levine, S.; Ancill, R.J.; Robert, A.P. (1989). *Assessment of Suicide Risk by Computer-Delivered Self-Rating Questionnaire: Preliminary Findings*. Acta Psychiatric Scandinavia (1989) Sep Vol 80(3) 216-220.

Locke, S.E., Kowaloff, H.B., Hoff, R.G., Safran, C., Popovsky, M.A., Cotton, D.J., Finkelstein, D.M., Page, P.L., & Slack, W.V. (1992). *Computer-Based Interview for Screening Blood-Donors for Risk of HIV Transmission*. Journal of the American Medical Association, 268 (10), 1301-1305.

Miller, H.G., Turner, C.F., and Moses, L.E., eds (1990). Chapter 6: Methodological Issues in AIDS Surveys. In AIDS: The Second Decade. Washington, DC: National Academy Press.

National Center for Education Statistics, 1993. Adult Literacy in America: A First Look at the Results of the National Adult Literacy Survey. Washington, DC: U.S. Department of Education.

O'Reilly, J.M, Hubbard, M., Lessler, J.T., Biemer, P. & Turner, C.F. "Audio and Video Computer-Assisted Self-Interviewing: Preliminary Tests of New Technologies for Data Collection." 1994. Journal of Official Statistics (In Press).

Schwarz, N., Strack, F., Hippler, H., and Bishop, G. (1991). *The Impact of Administration Mode on Response Effect in Survey Measurement*. Applied Cognitive Psychology, 5, 193-212.

Turner, C.F., Lessler, J.T., and Devore, J.W. (1992). *Effects of Mode of Administration and Wording on Reporting of Drug Use*. In Survey Measurement of Drug Use: Methodological Studies, eds. C.F. Turner, J.T. Lessler, and J.G. Gfroerer, DHHS Publication (ADM) 92-1929, Washington, DC: Government Printing Office.

COPAFS2.DOC

Waterton, J.J., and Duffy, J.C. (1984) "A Comparison of Computer Interviewing Techniques and Traditional Methods in the Collection of Self-Report Alcohol Consumption Data in a Field Survey." International Statistical Review 52:173-182.

Weinrott, Mark R.; Saylor, Maureen (1991). *Self-Report of Crimes Committed by Sex Offenders.* Journal of Interpersonal Violence (1991) Sep Vol 6(3) 286-300.

THE LANGUAGE OF SELF-ADMINISTERED QUESTIONNAIRES AS SEEN THROUGH THE EYES OF RESPONDENTS[1]

by

Cleo R. Jenkins
U.S. Bureau of the Census
and
Don A. Dillman
U.S. Bureau of the Census and Washington State University

---

[1]This paper draws heavily from a previous paper of ours (see Jenkins and Dillman, 1993). Cleo R. Jenkins is a Survey Statistician, Center for Survey Methods Research, U.S. Bureau of the Census, and Don A. Dillman is Senior Survey Methodologist, Office of the Director, U.S. Bureau of the Census, and Professor and Director of the Social and Economic Sciences Research Center at Washington State University. The views expressed in this paper are those of the authors and are not necessarily those of the U.S. Bureau of the Census or Washington State University. Questions concerning this paper should be addressed to Cleo Jenkins, U.S. Bureau of the Census, 433 Washington Plaza, Building 1, Washington DC, 20233-4700 (301-763-7331).

Acknowledgments: The authors would like to thank Elizabeth Martin and Theresa J. DeMaio of the Center for Survey Methods Research, U.S. Bureau of the Census, for reviewing this paper and providing insightful comments.

The Language of Self-Administered Questionnaires As Seen Through the Eyes of Respondents

by
Cleo R. Jenkins and Don A. Dillman

"... we must recall that language includes much more than oral and written speech. Gestures, pictures, monuments, visual images, finger movements--anything consciously employed as a sign is, logically, language."

John Dewey in Newell, A and Simon, H.A.
Human Problem Solving, *1972, pp. 65-66*

I. INTRODUCTION

Much survey research has been directed at studying interviewer-administered questionnaires. As a result, we have learned a great deal about question wording and sequencing effects in surveys and about the effect of memory on data quality (e.g., Jobe et al, 1993; Jobe et al., 1990; Lessler, 1989; Converse and Presser, 1986; Belson, W., 1981). Although it is equally important to understand these sources of error in self-administered questionnaires, it is not sufficient. The graphical presentation of information is every bit as important because it too has something to say to the respondent.

In Tourangeau's model (1984), as well as other models of the survey interview process, the first step is specified as "comprehending the question." Depending on the model, different steps follow, but generally, "retrieval of the relevant facts, judgment, and finally, response" are mentioned.

Although "comprehending the question" is the first step in an interviewer-administered survey, the task is different in a self-administered survey. In a self-administered survey, respondents must first "perceive the information" before they can comprehend it. Once they perceive it, they must "comprehend the layout of the information" as well as "the wording of the information." Furthermore, respondents must comprehend much more than just the wording of the survey questions and response categories. In a self-administered survey, respondents are often given introductory material and instructions. Also, they must comprehend directions that are meant to guide them through the form.

In an interviewer-administered questionnaire, the interviewer plays a critical role in what information the respondent perceives. In a self-administered format, the entire onus of perception is on the respondent, and we have not developed procedures for controlling errors that might arise as a result of their not perceiving information as we intend. In fact, we have not studied this much at all.

In addition, we need to pay attention to what motivates respondents to answer surveys. Cialdini (1988) has argued that people decide whether to perform a requested task on the basis of the inherent attractiveness of that task <u>and</u> other social or psychological influences, including

- reciprocation (the tendency to favor requests from those who have previously given something to you),
- commitment and consistency (the tendency to behave in a similar

way in situations that resemble one another),

- social proof (the tendency to behave in ways similar to those like us),
- liking (the tendency to comply with attractive requests),
- authority (the tendency to comply with requests given by those in positions of power), and
- scarcity (the tendency for rare opportunities to be more highly valued).

Groves and others (1992) provide examples of how each of these can be utilized to encourage survey participation. Although most of the examples refer to interviewer behavior or the implementation process, some can be applied to questionnaire design. For example, the fact that people tend to comply with attractive requests suggests that respondents will be more likely to answer an attractive questionnaire than an unattractive one.

Groves and others also argue that the helping tendencies of people can be utilized to encourage response. Three emotional states have been found to be associated with decisions to help another: anger, happiness, and sadness. One would expect that people are likely to become angry and therefore less likely to respond to a mail survey when the questions or the instructions are not easily understood.

Finally, the literature on opinion change (Petty and Capioppo, 1986) suggests that when a topic is of high personal relevance, subjects will change their opinion based on an in-depth review of a message. However, when the topic is not important to the subject, they will rely on a heuristic review, such as the credibility of the source. This literature suggests that if a questionnaire is not really important to a respondent, then we probably aren't going to persuade them to complete it by presenting them with an in-depth, highly logical, persuasive discussion of why they should complete it. Instead, we should rely on other means.

II. GRAPHICAL DESIGN PRINCIPLES

In the remainder of this paper, we present questionnaire design principles that struggle with how to best present information to (1) motivate respondents to attempt the tasks presented to them and to (2) aid respondents to accurately answer the questionnaire once they are motivated to do so. Most of the principles have not been tested carefully on controlled designs, although we try to cite those occasions in which they have been. Principle 20 (structuring and organizing a questionnaire) is an example of a principle drawn from experimental evidence with the Decennial Census questionnaires. Most of the other principles are drawn from the results of cognitive interviews with both the Schools and Staffing Survey and the Census of Construction Industries and from the redesign of the Survey of College Graduates. The response effects of the redesigned examples remain generally untested at this time. Therefore, the principles should be viewed as reasonable hypotheses for improving response, lowering item non-response, and improving accuracy. A major reason for writing this paper is to encourage experimental research on these issues.

1. **Present information in a format that respondents are accustomed to reading.**

We consider this the most important principle and one that is constantly and inadvertently violated. Jenkins et al. (1992a) present the results of cognitive interviews with the Schools and Staffing Survey. Example 1 shows the cover page of the School Questionnaire from the Schools and Staffing Survey. Jenkins et al. conclude that the readers, persistent as they were, usually did a pretty good job of following this page until they reached the end of the first column. These respondents read through the title information, then the first two paragraphs on the left-hand side of the page. Because these paragraphs refer to the label, they turned the questionnaire sideways to look at the label. When done, they returned to where they had left off on the left-hand column, and continued to read down the column. Instead of continuing to the top of the second column, however, generally they turned the page.

Example 2 presents a diagrammatical representation of the cover page's reading structure. It reveals the eye's necessary movement across the page. As can be seen, the current format requires respondents to make some pretty large unexpected leaps across the page, unexpected in the sense that a person anticipates reading a line of information from left to right, starting at the top of the page and moving down it.

It is not surprising, therefore, that the skimmers didn't bother to read this page at all. Generally, they glanced at it and turned the page. Because the skimmers never read the school named on the label, they often reported for the wrong school. In fact, this error was so great in 1991 that data for 10 states needed to be suppressed at first.

We present a redesigned version of the cover page using a natural reading format in conjunction with the next principle.

2. **Present only the most relevant information using graphical design features and composition.**

Another problem with the School Questionnaire's cover page is that it presents too much information. Skimmers quickly dismissed this information, probably because nothing was made particularly salient to them and they were not willing to look for that which was important. This suggests that the most important information needs to be made easily perceptible.

Example 3 shows a "user-friendly" cover page adapted from Jenkins and Ciochetto (1993). This cover page presents only the information the respondent needs to begin completing the questionnaire and it does so using a natural reading format and graphical design features. Jenkins and Ciochetto deliberately used a box that contains an unshaded area within a shaded one to showcase the very important instruction that was overlooked on the original questionnaire.

Example 4 exhibits the straight forward reading structure of this page. No need for the eye to do anything out of the ordinary, which we are more and more convinced is critical to designing good self-administered questionnaires.

3. Pique respondents' interests early in the questionnaire.

A third problem with the School Questionnaire was that respondents found themselves being asked to passively read a lot of material: the cover page, the cover letter (which was placed on the inside cover of the questionnaire booklet), and instructions (see Example 5). Skimmers skipped over this material entirely and went directly to question (a) in the middle of page 4 to begin completing the questionnaire.

Based on this, the third principle is: pique respondents' interest early in the questionnaire. Don't begin the questionnaire with a lot of prose, begin by asking a question or two. We hypothesize that respondents are more likely to read information once they have become actively engaged in answering the questionnaire.

As shown in Example 6, Jenkins and Ciochetto (1993) suggest beginning the Student Records Questionnaire by asking respondents to record the current time followed by a screener question. It is only after they ask these questions that they present a condensed version of the cover page and letter information. Even when they present this information, they deliberately used a question-and-answer format to keep the respondent actively engaged.

4. Dominantly feature questions over additional explanatory information.

This principle is violated in the Census of Construction Industries Questionnaire. Example 7 presents the "dollar value of business" item from that questionnaire. As can be seen, this item, like all of the items on the questionnaire, begins with a brief capitalized heading in bold that is meant to quickly convey the nature of the item. DeMaio and Jenkins (1991) conclude that respondents often neglected to read beyond the heading because the heading provided them with just enough information to formulate their own question. And, of course, they formulated the wrong question.

Example 8 presents a revised version of the item. In this version, the item heading is replaced with a bold-faced, comprehensive question. Bold-faced type was used to convey the importance of the question. Also, it serves as a road map for questions like this that have a leading phrase followed by several parts that are interrupted with other information. The other information is put in light-faced type.

5. Include in each question all of the relevant information necessary for respondents to answer it, rather than specifying information in a subsequent instruction.

This principle is demonstrated using another item from the Census of Construction Industries Questionnaire, the "number of employees" item (see Example 9). Respondents would read the heading here, and sometimes the question, then they would turn their attention to the answer boxes at the right (DeMaio and Jenkins, 1991). At this point, their eyes were drawn immediately to the column headings rather than the header. The header reads "Number of employees of this establishment during the pay period including the 12th of--," and because "pay period" is not mentioned in either the heading or the question, several respondents mistakenly thought this item was referring to monthly or quarterly time periods.

A revised version of this item is presented in Example 10. Besides

removing the heading, the pay period header was made the leading phrase in this version. This question now contains all of the information the respondent needs to answer it. And just to make sure respondents don't misunderstand, the pay periods are also repeated in each of the column headings.

6. Vertically align the questions and response categories.

As can be seen in the top view of Example 11, the Public School Questionnaire uses a question-on-the-left-answer-on-the-right format. Jenkins et al. (1992a) conclude that respondents often did not read instructions in this format. This is because they generally began to search for the answer once they read the question. As a result, their thoughts and consequently, their eyes were drawn away from the left-hand side of the page, where the instructions lay, to the right-hand side, where they knew the answer categories were.

The second view in Example 11 uses a vertical alignment. This places the instructions directly before the answer category, where respondents are more likely to perceive them. However, this may not solve the problem of respondents either overlooking or ignoring instructions. As already mentioned, respondents have a tendency to read only as much as they think is necessary to answer a question. Therefore, even if they perceive the instructions, they may still ignore them. If the instruction is relatively simple to begin with, a better solution is to incorporate it into the body of the question, as demonstrated in the last view.

7. If incorporating needed information into the question makes it too complicated to understand, then provide accompanying instructions at the place where they are needed.

If an instruction is long and/or complicated, incorporating it into the body of the question is likely to fail. As can be seen in Example 12, Item 2 of the Public School Questionnaire asks how many students were enrolled in the school on or about October 1 of this school year. Jenkins et al. (1992a) conclude that this item was difficult for respondents to read and understand because the flow of the question is interrupted by two parenthetical phrases and a lengthy two-sentence instruction. This leads us to conclude that one should never try to insert a stand-alone instruction between phrases of a continuous question.

Question 13a(2) in Example 8 illustrates the use of include and exclude statements that are too lengthy to incorporate into the body of the question. Here, the instructions are placed directly after the question. This places them as close to the question as possible without disrupting its flow. Still, further research is needed to determine the best method for coaxing respondents to read information that is not easily incorporated into the question.

8. Utilize single-task formats rather than multi-task formats.

Item 30 from the Public School Questionnaire (shown in Example 13) asks respondents to cross classify their employees by full- or part-time status and assignment. Jenkins et al. (1992a) conclude that quite a few respondents

seemed able to process only one aspect of this item--the categorization of employees by job description--and were unable to simultaneously deal with the additional request to report these employees by both full- and part-time status.

In Example 14, the multi-tasked format is replaced with a simpler one. This is accomplished by focusing on only one request at a time--first, respondents are asked to classify part-time employees by job description in part a. Then they are asked to classify full-time employees by job description. Because this format repeats information at the point it is needed, we hypothesize that it will make the respondent's task easier. Of course, the disadvantage is that it lengthens the questionnaire, which may decrease a respondent's motivation to complete it. This example illustrates the fact that there can be competing forces at work when we design a questionnaire and that we clearly need to learn more about these forces.

9. Utilize single-question formats rather than matrix-question formats.

Question 1 in Example 15 asks respondents if they currently have the students for class listed down the left-hand side of the page. If the respondents do, then they are to answer three follow-up questions. Jenkins and Ciochetto (1993) conclude that this format presents respondents with too many tasks at once. Furthermore, it presents them with a choice, but provides little guidance for making the choice. They can choose to answer a full set of questions about one student at a time. In this case, they work across the rows. Or they can answer the same question for each of the students. In this case, they work down the columns.

In Example 16, the matrix format is replaced with a single array of questions pertaining to one student at a time, with the questions running down the page rather than across it. In this version, respondents need only be concerned with answering one question about one student at a time, and they need not deviate from moving down the page in search of the next question. Although the researchers recommend further work in this area, a small number of cognitive interviews showed that this is a more manageable task from the respondent's point of view. (This example also used a new skip instruction. We discuss skip instructions later.)

Additional research supporting the single-question format comes from both focus group and experimental research on the 1990 Decennial Census Questionnaire. A focus group examination of the Census Questionnaire in which respondents were asked to answer a series of questions for each member of their household in a matrix format (questions in left-hand column to be answered for household members listed across the top of the page) identified the matrix format as a barrier to response (Dillman et al., 1991). Furthermore, a revised questionnaire, which used a single-question rather than a matrix-question format, attained an improved response rate (Dillman et al., 1992).

10. Make headings and instructions at the top of a page more prominent than those in the middle of a page.

Respondents find transitions between topics helpful. A transition need not be complicated, it simply needs to be enough to warn the respondent the topic is about to change. For instance, in Example 17, the heading "SECTION

2--STAFFING PATTERNS" in the middle of the page was enough to convey to respondents that the topic was about to change (Jenkins et al., 1992a).

In contrast, Example 18 shows a transitional heading followed by an instruction that comes at the top of a page. Contrary to expectations, respondents tended to read transitional headings and instructions that came in the middle of a page, but few read information that fell at the top of a page.

We hypothesize that respondents may be exhibiting a similar kind of behavior here as they exhibited with the introductory information. Many respondents skipped over the introductory material in an effort to get to the questions, but once they were actively engaged in answering the questionnaire, they were more likely to read information put into their path. Perhaps respondents skip over information at the top of a page in an effort to get to the next question, but once they become involved in answering the questions, they are more likely to see other information.

11. Provide directions in a natural reading format and utilize graphical design features and composition to make the directions more salient.

In order to efficiently and accurately answer a self-administered questionnaire, respondents must be able to maneuver their way through the questionnaire. One very important instruction for doing this is the skip instruction. The problem with skip instructions, however, is that respondents commonly overlook them (Jenkins and Ciochetto, 1993; Turner et al., 1992; Gower, 1989).

Jenkins and Ciochetto (1993) conclude that respondents overlook skip instructions for two reasons, one of which is derived from the other. The primary reason respondents overlook the skip instruction is because they do not perceive it, but the reason they do not perceive it is because of the convoluted reading structure presented by the skip instruction. Item 29a shown in Example 19 illustrates this. A respondent begins to answer this item by first reading the question "Were there any teaching vacancies in this school for this school year, i.e., teaching positions for which teachers were recruited and interviewed?" Then they will move to the right-hand side of the page to answer the question, see the answer boxes, and continue to the right of these to read the answer choices "yes" and "no." The next step in the process is to choose one of these, say the "no" response, and to move back to the left of this to mark the answer box. Note what is happening at this moment--the respondents are moving away from the skip instruction. <u>If the skip instruction has not been in some way made salient to respondents before they begin their journey back to the left, chances are they are never going to see it.</u> Once they mark the answer box they are likely to conclude they are done answering this question and are going to begin to look for the next question.

Experimental data presented by Turner et al. (1992) confirm the hypothesis that respondents only see information to the right of an answer category if it is in some way made salient. Among other questionnaire design issues, Turner et al. studied the extent to which respondents and interviewers correctly executed branching instructions embedded in alternative versions of the 1990 National Household Survey on Drug Abuse (NHSDA) questionnaire. Turner et al. conclude that respondents were more likely to overlook a visually obscured branching instruction, as shown in question 1 of Example 20, than a visually salient one, as shown in question 5. They conclude that both

the length and visual salience of questions a through e in question 5 assisted respondents in correctly following the branching instruction in question 5.

These studies suggest that directions need to be presented in a more natural reading format and graphical design features and composition should be used to make the directions more salient. Given this information, three alternative skip instruction formats are presented below, each of which appears to have advantages and disadvantages: (1) the salient skip instruction, (2) the intermediate skip instruction, and (3) the natural reading sequence skip instruction.

Salient Skip Instruction. Example 21 shows the salient skip instruction. Rather than having information placed to the right of the answer categories, directional arrows are placed to the left of both answer boxes. These arrows extend horizontally from each answer box towards the left-hand margin of the page and then turn vertically downwards. One of these arrows proceeds to the next question and the other ends with a verbal instruction within a shaded box.

This format was designed to overcome the conventional skip instruction's highly convoluted reading format (moving from left to right, right to left, back again to the right and finally, back to the left) and to replace a more-difficult-to-perceive verbal instruction with a more-easily-perceived combination visual/verbal instruction. In our judgement, the advantage of this format is that respondents may visually take in, if only briefly, the skip instruction information while moving from left to right in search of the answer categories.

The disadvantage, however, is that whereas respondents may be more likely to see this information, they also may be more likely to misunderstand it. It is possible that the wrong respondents (those who are supposed to continue to the next question) may mistakenly execute the skip instruction because of its visual salience, leading to a serious error--the omission of data. Another disadvantage with this format is that a question with a complicated skip instruction may become visually cluttered.

Intermediate Skip Instruction. Example 22 presents the intermediate skip instruction. This format relies on two features: (1) graphical instructions (an arrow) for going to the next question and (2) words to direct other respondents through a skip pattern. The two paths are further distinguished by originating the arrow from the left of the answer choice, and placing the words to the right.

In our judgment, the advantage that this format may have over the salient skip instruction is that respondents are unlikely to make the serious error of incorrectly executing the skip instruction. Another advantage is that it may not appear as visually cluttered to respondents. However, a small number of cognitive interviews suggest that it may not be as efficiently executed as the salient skip instruction. Just as with the conventional skip instruction, the word instruction to the right of the answer category may be overlooked at first. However, it is likely to be more efficiently executed than the conventional skip instruction, in which nothing but words are used off to the right of the answer choices. Therefore, this skip instruction format is a deliberate compromise between the conventional and salient skip instruction.

Natural Reading Sequence Skip Instruction. In contrast to the above skip instructions in which a respondent must move from left to right in search of the answer categories and then reverse this direction and move from right

to left to answer the question, another possibility might be to establish a skip instruction format with a more efficient, natural, and logical flow. This format is shown in Example 23. As can be seen in this example, the answer boxes are placed to the right of the answer categories and the skip instructions to the right of the answer boxes. To maintain the vertical alignment of the answer boxes, the answer categories are right-justified rather than left-justified. Also, if the answer categories need to be double or triple-lined, as is the case with the category "Dropout/Chronic Truant (See Definition Below)" in question 1 of Example 23, then the answer box should follow the last of these lines. This is to help maintain the respondent's natural reading structure, for which we have been arguing all along.

This skip instruction seems to have several advantages over the preceding skip instructions. The first and probably best advantage is that the information is presented to respondents in the sequence they need it: first the answer categories, then the answer boxes, and finally, the skip instruction. Example 24 compares the reading format of the natural reading sequence skip instruction with that of the conventional skip instruction. As can be seen, respondents need not ever reverse their direction with the natural reading sequence skip format. Another advantage is that the natural reading sequence format is not cluttered looking.

A disadvantage, however, is that respondents may overlook bracketed skip instructions using this format. Although these instructions will be closer to the answer boxes in this format than they are in the conventional skip instruction format (that is, if the answer categories come between the answer boxes and the skip instruction), they be just far enough away from the answer boxes as to be out of the respondent's view.

Another disadvantage is that from an overall perspective, the questionnaire's vertical alignment is disrupted. In the previous formats, the questions, answer boxes, and categories are all left justified and begin in the same horizontal position on the page. Although vertical alignment of the questions can be maintained using the natural reading sequence skip instruction format, the answer categories will certainly not be vertically aligned. The answer boxes can be made to maintain vertical alignment within a question; however, they may not be able to maintain alignment from question to question, further disrupting the overall look of the questionnaire.

A final disadvantage with the natural reading sequence skip instruction is related to data processing. In this format, the location of the keycodes is problematic. One possibility is to place the keycodes before the answer category, but this puts them quite a distance from the answer box from the keyer's perspective. This may slow down production and/or increase keyer error. Another possibility is to place them either directly before or after the answer box, but this may confuse the respondent.

We have described skip instructions at some length because it is an area which is exceedingly important, but now lacks ideal solutions. The alternatives presented need extensive testing in large samples.

12. Utilize graphical design techniques to establish a clear path through the questionnaire for the respondent to follow.

Many questionnaires mix questions and information, utilizing space wherever it is available and thinking that so long as the information is presented, it will get read. As can be seen in Example 25, it is unclear to

the respondent where to begin, and most important in what order the information is to be read.

Example 22 is a redesigned page from the Survey of College Graduates. Here the white answer spaces contrast with the light blue background. The message intended, and communicated by graphical layout rather than words, is to establish a visual path through the questionnaire by associating the white spaces with the "need to provide an answer."

13. Avoid using the same design feature to request different respondent actions.

The essence of this principle is to associate particular design features with what the respondent is being asked to do, and to be completely consistent with their use. For example:

13a. Use dark type for question stems and light type for response category options.

13b. Write all definitions and special instructions for a particular question in italics placed within parentheses.

13c. Use capital letters for words to be emphasized to the respondent in both questions and answers.

The important point here is not that capitals must be reserved for emphasis, and italics for instructions, or that bold type is better for questions than light type. Doing the opposite may work just as well--the issue is consistency, so that as a respondent gets into a questionnaire they begin to associate the chosen procedure with a particular piece of information or request for action.

14. Utilize variability in design features judiciously.

Closely associated with the need to be consistent is the need to limit variability. One would never consider writing a paragraph in which every word is written in different type fonts and sizes. Doing so would slow down the reader's comprehension. Instead, one should select a limited number of design elements and use them consistently.

15. Visually emphasize information the respondent needs to see and de-emphasize information the respondent does not need to see.

Coding information is a good example of this principle. In Example 26, the codes are bold and made even more prominent by encasing them in boxes. Not only that, but they are placed directly in the respondent's reading path. One result is that respondents may mentally process information irrelevant to them, thus making the task of responding more time consuming and difficult than necessary.

In Example 22 the light blue background is a 10 percent screen, and the coding information is printed in small numbers without boxes in 100 percent color. The respondent, who is already being guided "towards" the white answer spaces by black type of questions and answers and "away from" the blue background seems less likely to see or be confused by the dark blue lettering. Furthermore, the codes are placed outside the respondent's reading path. Yet, for a person who is searching for the blue code numbers, they are easily visible.

16. Utilize graphical layout of questions on the page to distinguish among different types of question structures; maintain consistency within types.

If a questionnaire begins by listing answer categories below the stem of a question vertically, like the "yes/no" answer categories in the first question in Example 27, it is undesirable to occasionally present answer choices horizontally, or even to sometimes use a second or third column of answer choices. Once a format is selected it needs to be followed consistently.

If one has a question like C9 and C10, where several items in a series are to be evaluated and the answer categories are the same for each item in the series, these answer categories should be placed to the right of the items and the respondent should be instructed to choose from among horizontally arranged categories in this case. Respondents should learn to choose from among vertical choices when the boxes are on the left and from among horizontal choices when boxes are to the right.

17. Provide descriptive captions either above, beneath, or to the right of blank answer spaces and utilize appropriate signs or symbols whenever numbers are requested.

When people are asked to report income, number of weeks worked, or other data by filling in blank spaces, inaccuracies may result from utilizing the wrong units or from not remembering exactly what was asked. Therefore, the answer spaces in Example 27 have captions to remind people what is being requested. For instance, C13 has the caption "Total 1991 earned income." In addition, the blank answer space has a dollar sign and ".00" in it to keep people from reporting cents, since they weren't wanted.

18. Utilize dominant graphical markings to provide the most important information needed by the respondent to guide them through the answering process.

This principle is violated in Example 25, where the "return to" instruction is predominant. It is also violated in Example 28 where the black marks used to optically scan the questionnaire are quite dominant. In neither case do the dominant marks effectively guide the answer process.

In Example 29, the dominant markings are the questionnaire's title, **THE 1992 NATIONAL CENSUS TEST**, followed by the **ARROW**, and the **PERSON 1** and **PERSON 2** headings. These dominant markings are meant to guide the respondent through the form.

19. Avoid the separation of questions through the use of lines and rectangles in favor of an open format in which the respondent's answering path is clearly shown.

Frequently designers of questionnaires utilize lines and rectangles to separate questions from one another. In general this practice makes questionnaires more, rather than less, difficult to answer. The use of rectangles, as shown in Examples 25 and 30, gives no clear indication of where to go next; the lines function in much the same way as "stop" signs, requiring

one to stop and contemplate the next steps. This is especially the case in these two examples, where it is not readily apparent which box comes next. Consequently, the boxes require additional information, that is, the prominent section numbers. Also, the use of lines is one additional use of ink on a page which must then be cognitively processed by the reader, in contrast to white (or other background color) space which one can pass over without pausing to think about what it means.

In contrast, the formats used in Examples 22, 27, and 29 are open, using lines mostly to identify the page space in which answers are to be provided. These pages are easier for respondents to follow. Also, the respondent path is easily recognized, following the cultural norm of left to right within the defined answering space, and top to bottom on the page.

20. Structure and organize the questionnaire in such a way that it, first, makes sense to respondents and, second, avoids leaving the choice of the order in which questions get answered up to the respondent.

On the surface, this principal seems obvious and easy to implement. However, this may not be so, if the Census long and short forms are any indication. The Census long form is probably one of the most complex questionnaires in existence. It has a fold-out flap which asks for a listing of household members, followed by a matrix of short-form information and finally two pages of sample population questions for each person. This form involves a complex sequence of tasks, the order of which was traditionally dictated by Census needs to provide Congress with mandated information by the end of the census year.

A split-panel experiment with the long form, known as the 1990 Alternative Questionnaire Experiment (AQE), showed that the form's structure was not properly organized from the respondent's point of view (DeMaio et al., 1992). Along with the control form, which was identical to the 1990 Census long form, five experimental questionnaires were tested in the 1990 AQE. Two of the experimental questionnaires (Panels 2 and 3) incorporated many of the principles we have discussed concerning color, the consistent use of typeface and answer spaces, etc. However, three of the experimental questionnaires (Panels 4, 5, and 6) incorporated dramatic changes to the structure and organization of the form. Panel 4 became a matrix booklet in which the flap was eliminated and all of the person items were placed together. Panels 5 and 6 became "kits" in which individual questionnaires for each person in the household were placed in a folder.

The main finding was that "small" format changes alone (as incorporated in Panels 2 and 3) did very little to improve either item or overall response rates, but it took changes to the structure and organization of the questionnaire (as incorporated in Panels 4, 5, and 6) to make improvements. This suggests that "small" format changes are not enough to overcome the difficulty of completing a questionnaire that is not properly organized from the respondent's point of view. In addition, the Simplified Questionnaire Test (Dillman et al, 1992) and the Appeals and Long Form Experiment (Bates, 1993) confirmed this finding.

III. CONCLUSION

Little information on the design of self-administered questionnaires existed until relatively recently. That which did was based primarily on common sense and individual experience. Instead, it was the verbal language of interviewer-administered questionnaires that predominately captured the attention of researchers.

The evidence presented in this paper demonstrates, however, that we need to pay serious attention to the visual language of self-administered questionnaires in addition to the verbal. Toward this end, we need to develop a set of scientifically derived and experimentally proven graphic design principles to guide us in our quest to improve both response rates and the accuracy of responses. We hope that the principles we've developed are a first step in that direction. We have little doubt that the problems we've uncovered exist. However, because many of the solutions have not been tested, we openly admit that they are subject to challenge. Some of the solutions we've offered will stand the test of time; others will not. Undoubtedly, this is an area in need of further study and creative insight.

Finally, we also hope to expand upon our work here by exploring literature that has remained outside the domain of survey methodology to date-- most notably, the eye-movement and the graphical design literature. Knowing what we do now, it certainly seems that this literature may offer further insight into the self-administered question-response process.

REFERENCES

Bates, N. and DeMaio, T. 1989. "Using Cognitive Research Methods to Improve the Design of the Decennial Census Form." *Proceedings of the Bureau of the Census Fifth Annual Research Conference*, March 1989, pp. 267-290.

Bates, N. 1993. Personal Communication. Preliminary Results from Appeals and Long Form Experiment. U.S. Bureau of the Census. Washington, DC.

Belson, W. 1981. *The Design and Understanding of Survey Questions*. Gower.

Cialdini, R.B. 1984. *Influence: The New Psychology of Modern Persuasion*. New York: Quill.

Converse, J.M. and Presser, S. 1986. *SURVEY QUESTIONS Handcrafting the Standardized Questionnaire*. Newbury Park: Sage Publications, Inc.

DeMaio, T. and Jenkins, C. 1991. "Questionnaire Research in the Census of Construction Industries." *Proceedings of the Section on Survey Methods Research*, American Statistical Association, August 1991.

DeMaio, T.J. and Bates, N.A. 1992. "Redesigning the Census Long Form: Results from the 1990 Alternative Questionnaire Experiment" *Proceedings of the Section on Survey Methods Research*, American Statistical Association, August 1992.

Dillman, D.A., Reynolds, R.W., and Rockwood, T.H. 1991. *Focus Group Tests of Two Simplified Decennial Census Forms*. Technical Report 91-39. The Social and Economic Sciences Research Center. Washington State University. Pullman.

Dillman, D.A., Sinclair, M.D., and Clark, J.R. 1992. "Mail-back Response Rates for Simplified Decennial Census Questionnaire Designs." *Proceedings of the Section on Survey Methods Research*, American Statistical Association, August 1992.

Groves, R.M., Cialdini, R.B., and Couper, M.P. 1992. "Understanding the Decision to Participate in a Survey." *Public Opinion Quarterly*, 56: 475-495.

Jenkins, C. 1992. "Questionnaire Research in the Schools and Staffing Survey: A Cognitive Approach." *Proceedings of the Section on Survey Methods Research*, American Statistical Association, August 1992.

Jenkins, C., Ciochetto, S., and Davis, W. 1992a. "Results of Cognitive Research on the Public School 1991-92 Field Test Questionnaire for the Schools and Staffing Survey." Internal Census Bureau Memorandum, June 15, 1992.

Jenkins, C., Ciochetto, S., and Davis, W. 1992b. "Recommendations for Changes to the Public School 1991-92 Field Test Questionnaire for the Schools and Staffing Survey (SASS-3A) Based on Cognitive Interviews." Internal Census Bureau Memorandum, September 24, 1992.

Jenkins, C. and Ciochetto, S. 1993. "Results of Cognitive Research on the Multiplicity Question from the 1991 Schools and Staffing Survey Student Records Questionnaire." A report submitted to the National Center for Education Statistics, Census Bureau Memorandum, February 10. 1993.

Jenkins, C.R. and Dillman, D.A. 1993. "Combining Cognitive and Motivational Research Perspectives for the Design of Respondent-Friendly Self-Administered Questionnaires. " Revision of a Paper Prepared for Presentation at the Annual Meetings of the American Association for Public Opinion Research. May, 1993.m

Jobe, J.B., White, A.A., Kelley, C.L., Mingay, D.J., Sanchez, M.J. and Loftus, E.F. 1990. "Recall Strategies and Memory for Health Care Visits." *Milbank Memorial Fund Quarterly/Health and Society*, 68: 171-189.

Jobe, J.B., Tourangeau, R., and Smith, A.F. 1993. "Contributions of survey research to the understanding of memory. *Applied Cognitive Psychology*, 7: 567-584.

Lessler, J.T. 1989. "Reduction of Memory Errors in Survey Research: A Research Agenda." *Proceedings of the 47th Session of the International Statistical Institute*, Paris, France, pp. 303-322.

Newell, A. and Simon, H.A. 1972. *Human Problem Solving*. New Jersey: Prentice-Hall, Inc., pp. 65-66.

Petty, R.E. and Cacioppo, J.T.. 1986. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. New York: Springer-Verlag.

Tourangeau, R. 1984. "Cognitive Science and Survey Methods" in T.B. Jabine, M.L. Straf, J.M. Tanur, and R. Tourangeau (eds), *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, Washington, DC: National Academy Press, pp. 73-100.

Turner, C.F., Lessler, J.T., George, B.J., Hubbard, M.L. and Witt, M.B. 1992. "Effects of Mode of Administration and Wording on Data Quality" in C.F. Turner, J.T. Lessler, and J.C. Gfroerer, (eds), *Survey Measurement of Drug Use Methodological Studies*, Washington, DC: National Institute of Drug Abuse, U.S. Department of Health and Human Services, pp. 221-243.

FORM SASS-3A

U.S. DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS
ACTING AS COLLECTING AGENT FOR
U.S. DEPARTMENT OF EDUCATION
NATIONAL CENTER FOR EDUCATION STATISTICS

OMB No. 1850-0588
Approval Expires 12/31/92

This report is authorized by law (20 U.S.C. 1221e). Your answers will be kept strictly **confidential.** Results from this survey will appear in summary or statistical form only, so that individuals cannot be identified.

PUBLIC SCHOOL QUESTIONNAIRE

SCHOOLS AND STAFFING SURVEY

1991-92 Field Test

RETURN TO

Bureau of the Census
Current Projects Branch
1201 East 10th Street
Jeffersonville, IN 47132-0001

Please complete this questionnaire with information about the SCHOOL named on the label, and return it to the Bureau of the Census in the enclosed preaddressed envelope. Please return it within 3 weeks.

If the school's name or grade level is different from that indicated on the label or if you have any questions, please call the Bureau of the Census at 1-800-221-1204.

If the school named on the label is no longer in operation, mark (X) the box below and return this questionnaire to the Bureau of the Census in the enclosed envelope.

001 1 ☐ School no longer in operation

THIS SURVEY HAS BEEN ENDORSED BY:

- American Association for Counseling and Development
- American Association of School Administrators
- American Federation of Teachers
- Council of Chief State School Officers
- National Association of Elementary School Principals
- National Association of Secondary School Principals
- National Education Association

(Please correct any error in name, address, and ZIP Code)

SKU: 374
TEA: 31.4

ID #: 8768420512

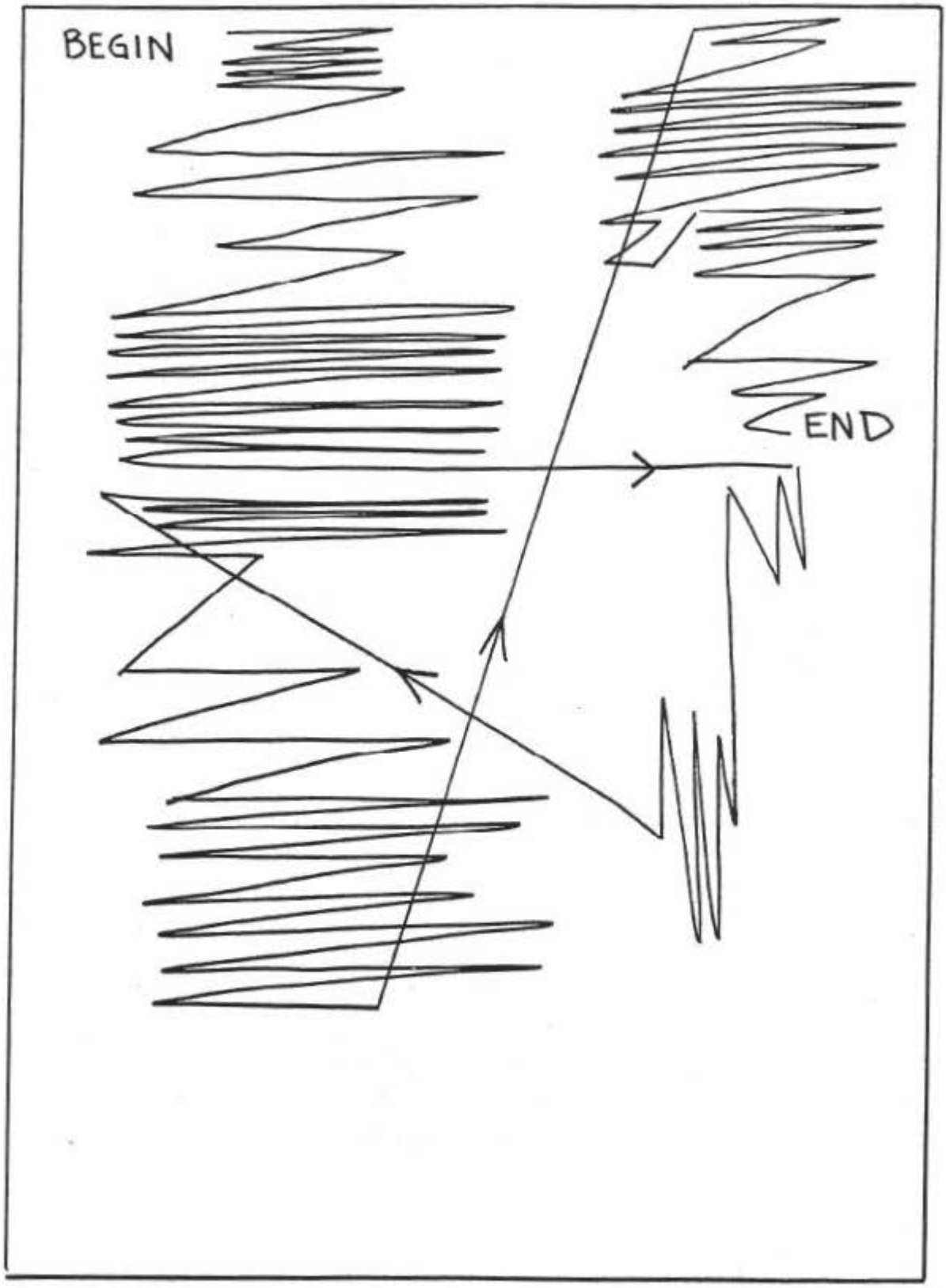
0 41 1 092 501 3 1

PRINCIPAL
Martin High School
1001 4th Street
Nevada, IA 50201

09200

3A 27

Example 1. Cover Page of the Public School Questionnaire (Jenkins et al., 1992b).



Example 2. Diagrammatical Representation of the Public School Questionnaire Cover Page's Reading Structure.

OMB No. 1850-0598: Approval Expires 12/31/93

FORM SASS-3X
(1-22-93)

U.S. Department of Education
National Center for Education Statistics

SCHOOL QUESTIONNAIRE
SCHOOLS AND STAFFING SURVEY

1993 FIELD TEST

Conducted by:

U.S. Department of Commerce
Bureau of the Census

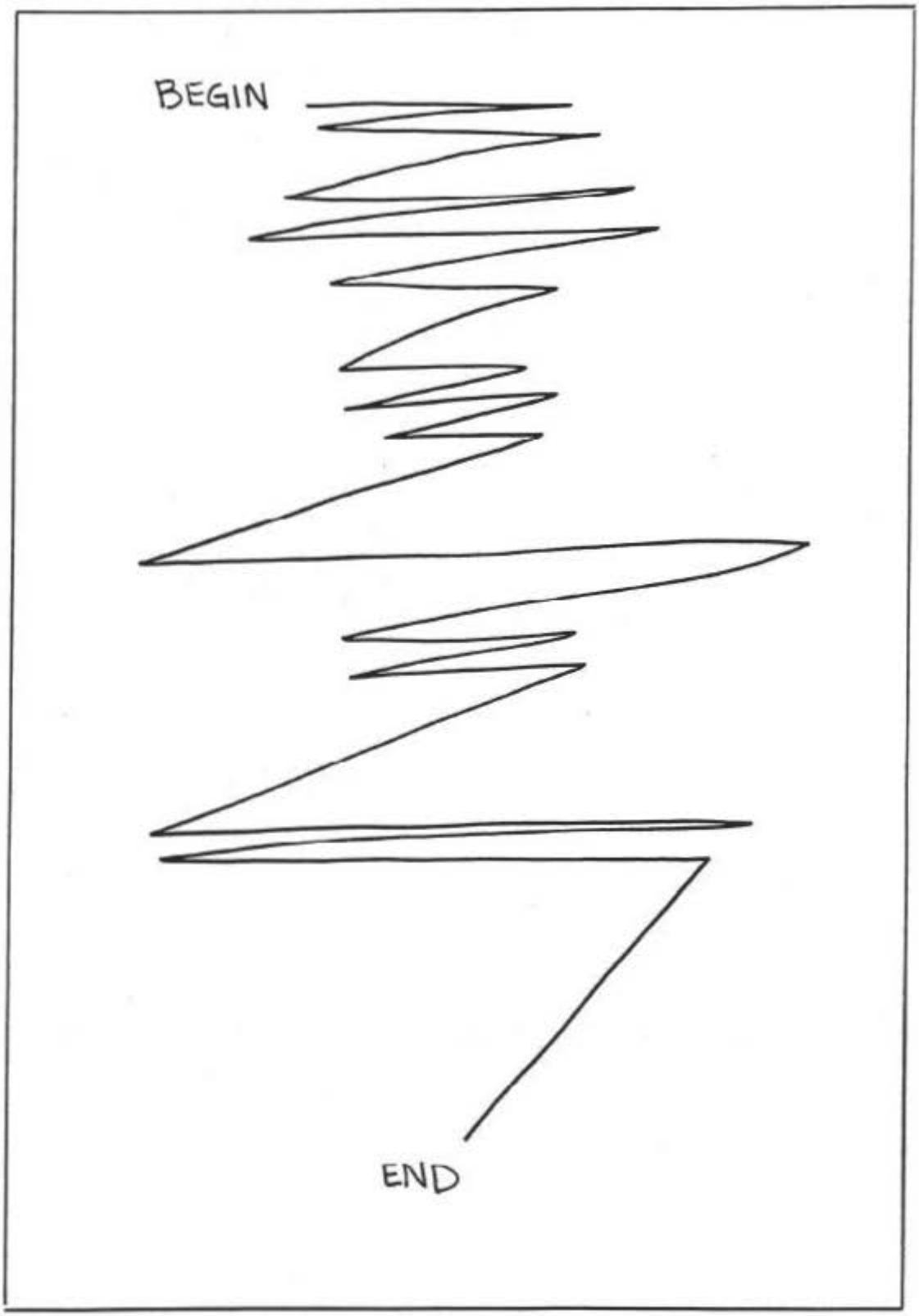
PLEASE COMPLETE THIS QUESTIONNAIRE WITH INFORMATION ABOUT:

MARTIN HIGH SCHOOL
GRADES 9-12

If you have any questions, call the Bureau of the Census at 1-800-221-1204.

After completing this form, mail it to the Bureau of the Census in the preaddressed envelope provided. Please return it within 2 weeks.

Example 3. Redesigned Version of the Public School Questionnaire's Cover Page (adapted from Jenkins and Ciochetto, 1993).



Example 4. Diagrammatical Representation of the Redesigned Public School Questionnaire Cover Page's Reading Structure.

Dear Principal:

The National Center for Education Statistics (NCES) of the U.S. Department of Education requests your participation in the field test of the 1992–93 Schools and Staffing Survey. Your school is one of 900 public and private schools across the Nation selected to be in the sample.

The Schools and Staffing Survey, first conducted in school year 1987–88, and again in 1990–91, is an integrated set of surveys consisting of the Teacher Demand and Shortage Survey, the School Survey, the School Administrator Survey, and the Teacher Survey. These surveys are being conducted periodically to measure critical aspects of teacher supply and demand, the composition of the administrator and teacher work force, and the general status of teaching and schooling. The purpose of the School Survey is to obtain information about schools such as staff-pupil ratio, student characteristics, staffing patterns, and teacher turnover.

The U.S. Bureau of the Census is conducting the survey for the National Center for Education Statistics by the authority of Section 406(b) of the General Education Provisions Act, as amended (20 U.S.C. 1221e). The data will be treated as confidential and will be reported only in statistical summaries that preclude the identification of any individual participating in the surveys.

We are conducting this field test with a sample of schools. While this minimizes overall response burden, the value of each individual survey response is greatly increased because it represents many other schools. I, therefore, encourage you to participate in this voluntary survey by completing this questionnaire and returning it within 3 weeks to the **Bureau of the Census, Current Projects Branch, 1201 East 10th Street, Jeffersonville, IN 47132-0001,** in the preaddressed envelope enclosed for your convenience.

Thank you for your cooperation in this very important effort.

Sincerely,

Emerson J. Elliott
Acting Commissioner
National Center for Education Statistics

INFORMATION ABOUT YOUR PARTICIPATION

Public reporting burden for this collection of information is estimated to average one hour, including the time for reviewing instructions, gathering the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to the U.S. Department of Education, Information Management and Compliance Division, Washington, DC 20202-4651; and to the Office of Management and Budget, Paperwork Reduction Project 1850-0598, Washington, DC 20503.

Example 5. Introductory Information from the Public School Questionnaire (Jenkins et al., 1992b).

INSTRUCTIONS

Most of the items on this questionnaire are arranged so that the questions are on the left side of the page and the answer categories or spaces for written answers are on the right. Please answer the questions by marking the appropriate answer category with an X, or recording your answer in the space provided. We suggest that you use a pencil or a typewriter, rather than a pen or marker.

Notice that at the end of some answer categories and answer spaces, there are instructions to skip to a later item or to continue with the next item on the questionnaire.

Unless otherwise indicated, all questions refer to the 1991–92 school year.

If you are unsure about how to answer a question, give the best answer you can and make a comment in the "Remarks" space. Please include the item number.

If you have any questions, call the Bureau of the Census at 1–800–221–1704.

Return your completed questionnaire to the Bureau of the Census in the enclosed preaddressed envelope. Please return it within 3 weeks.

Please keep count of the time required to complete this questionnaire. At the end of the survey, you are asked to record the amount of time spent.

a. Please give your name, title, telephone number, and the most convenient days/times to reach you. This information will be used only if it is necessary to clarify any of your responses.

Name			
Title			
Telephone	Area code	Number	
If necessary to reach you – *Specify*	Days	Time	a.m. p.m.

b. Does this school serve students in ANY of grades 1 through 12 or comparable ungraded levels?

010 1☐ Yes – *Continue with c.*
2☐ No – ***Stop now*** *and return this questionnaire to the Bureau of the Census in the enclosed envelope.*

Thank you for your time.

c. Please check the identification number on address label – Is this your School State Identification Number?

011 1☐ Yes
2☐ No – *Provide the correct number below.*
012 _____

Remarks

Example 5 Continued.

A. Please record the current time. At the end of the questionnaire you are asked to record the amount of time required to complete this questionnaire.

Current time: _____

B. Does this school provide instruction for grade 9 or above?

009 1 ☐ Yes → 2 ☐ No →

What grading system is used to compute a student's grade point average (GPA)?

010
1 ☐ 0.0 to 4.0
2 ☐ 0 to 100
3 ☐ −1 to 3
4 ☐ Other specify _____

Skip to information below

WHY ARE WE CONDUCTING THIS SURVEY?

This questionnaire is the last in a series of surveys designed to obtain nationwide information on schools, staffing patterns, and student characteristics. We will treat your data as confidential and only use it to prepare statistical summaries.

WHO IS CONDUCTING THIS SURVEY?

The National Center for Education Statistics of the U.S. Department of Education requests your participation in this voluntary survey. The Bureau of the Census is conducting this survey by the authority of Section 406(b) of the General Education Provisions Act, as amended (20 USC 1221e).

INFORMATION ABOUT YOUR PARTICIPATION

Public reporting burden for this collection of information is estimated to average thirty minutes, including the time for reviewing instructions, gathering the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to the U.S. Department of Education, Information Management and Compliance Division, Washington, DC 20202–4651; and to the Office of Management and Budget, Paperwork Reduction Project 1850–0598, Washington, DC 20503.

Example 6. Revised Introductory Material Beginning with Questions Rather Than Prose (Jenkins and Ciochetto, 1993).

Item 15 — DOLLAR VALUE OF BUSINESS DONE IN 1989

		Mil.	Thou.	
a. What was the value of all construction work done by this establishment in 1989? *Report your best estimate using either billings, revenues, receipts, or other estimate of value of construction work done. Refer to instructions for further explanation.*				
Reported data based on — *Mark (X) in appropriate box.* 136 1 ☐ Billings 3 ☐ Receipts 2 ☐ Revenues 4 ☐ Other — *Specify* _____				
INCLUDE • New construction • Additions, alterations, or reconstruction • Special trades contracting work • Maintenance and repair work • Land development and improvement • Installation and service of equipment • Buildings and other structures built for sale, excluding the value of the land • Construction work on own account				
EXCLUDE • The cost of industrial and other specialized machinery which are not an integral part of a structure	137			☐0
b. What were your receipts from kinds of business other than those reported in line a above in 1989?				
INCLUDE • Architectural and engineering services • Construction management/ consulting services • Wholesale trade • Retail trade • Realty services • Rental of machinery, equipment, or buildings to others • Manufacturing • Transportation • Materials sold to contractors • Sale of land	139			☐0
c. TOTAL dollar value of business done in 1989 — ***Sum of lines a and b*** →	140			☐0

Example 7. The Original Version of the Dollar Value of Business Item from the Census of Construction Industries Questionnaire (DeMaio and Jenkins, 1991).

	Key	Dollars in thousands ($)	Mark (X) if "0"
13a. For this establishment in 1989, —			
(1) what were the receipts or billings for all contract construction work done for others? Exclude the dollar value of items purchased by this establishment that were installed in a building but were not part of its structure, such as production machinery, furniture, etc.	141		☐ 0
(2) what was the estimated dollar value of speculative construction work done on residential and other building projects which you sold or intended to sell? **Include the estimated dollar value of —** • work actually done in 1989, whether buildings were sold or not. • work in progress, that was actually done in 1989. • all improvements to land associated with these building projects done by or for you in 1989. **Exclude the estimated dollar value of —** • work done before January 1989 and after December 1989. • land. Even though land would generally be included in the value of your building projects, the value of the land is not considered construction work done. • work done in 1989, for rent or lease.	142		☐ 0
(3) what was the estimated dollar value of construction work done for this establishment's own use, i.e., not intended for sale, rent, or done under contract for others?	143		☐ 0
(4) what was the dollar value of construction work done? Sum lines (1) through (3).	137		☐ 0
b. What was the dollar value of receipts or billings for all other business activities done by this establishment in 1989? **Include —** • architectural services • building on your own land for rent or lease • construction management services • engineering services • manufacturing • mining • real estate agents and managers • rental of construction machinery or equipment to others • retail trade • subdividing and preparing your own land into lots, for sale, rent, or lease • transportation • wholesale trade • other business activities	139		☐ 0
c. What was the total dollar value of all business done by this establishment in 1989? Sum lines 13a(4) and 13b.	140		☐ 0

Example 8. The Revised Version of the Dollar Value of Business Item from the Census of Construction Industries Questionnaire (DeMaio and Jenkins, 1991).

Item 5 — NUMBER OF EMPLOYEES

How many paid employees, permanent or temporary, full-time or part-time, were on the payroll of THIS ESTABLISHMENT during the pay periods shown at the right? *Include those on paid sick leave, paid holidays, and paid vacations as well as those actually working. If a corporation, include salaried officers and executives of this establishment. If unincorporated, exclude proprietors or partners. Include all employees for whom you file quarterly withholding statements. Do not include your subcontractors or their employees.*

	Number of employees of this establishment during the pay period including the 12th of —			
	March 1989	May 1989	August 1989	November 1989
a. Construction workers — ***Include*** — • Laborers • Apprentices • Journeymen • Craftsmen • Equipment operators and mechanics • Truck drivers and helpers • On-site record keepers • Others engaged directly in construction operations, including supervisors up through the working foreman level	101	102	103	104
b. All other employees — ***Include*** — • Executives • Professionals • Technicians • Purchasing • Accounting • Personnel • Office staff • Supervisors above the working foreman level	105	106	107	108
c. TOTAL NUMBER OF EMPLOYEES — ***Sum of lines a and b above*** →	109	110	111	112

Example 9. The Original Version of the Number of Employees Item from the Census of Construction Industries Questionnaire (DeMaio and Jenkins, 1991).

Your answers to 5 through 9 should be based on all employees for whom you filed withholding statements (Federal Tax Return Form 941). Do NOT include your subcontractors or their employees.

5. During the pay periods which include the 12th of March, May, August, and November 1989, —	Number of employees			
	Pay period including the 12th of March 1989	Pay period including the 12th of May 1989	Pay period including the 12th of August 1989	Pay period including the 12th of November 1989
a. how many construction workers were on the payroll of this establishment? **Include —** • Working foremen • Job-site record keepers • Laborers • Apprentices • Journeymen • Craftsmen • Equipment operators and mechanics • Truck drivers and helpers • Others engaged directly in construction	101	102	103	104
b. how many other employees were on the payroll of this establishment? **Include —** • Supervisors • Job-site and home office clerical and maintenance staff • Personnel staff • Purchasing agents • Accounting staff • Technicians • Architects • Engineers • Professionals • Executives • Others engaged in non-construction activities	105	106	107	108
c. how many total employees were on the payroll of this establishment? Sum lines a and b.	109	110	111	112

Example 10. The Revised Version of the Number of Employees Item from the Census of Construction Industries Questionnaire (DeMaio and Jenkins, 1991).

4. What percent of K-12 students enrolled in this school are male?

Record the percent in whole numbers, not tenths; do not enter a decimal point. Do NOT include prekindergarten or postsecondary students.

055 _____ %

---

4. What percent of the students enrolled in this school are male?

Record the percent in whole numbers, not tenths; do not enter a decimal point. Do NOT include prekindergarten or postsecondary students.

055 _____ %

---

4. Excluding prekindergarten or postsecondary students, what percent (in whole numbers) of the students enrolled in this school are male?

055 _____ %

Example 11. Horizontally Aligned Question Format (Top View), Vertically Aligned Question Format with Stand Alone Instructions (Middle View), and Vertically Aligned Question Format with Incorporated Instruction (Bottom View).

2. How many students (in head counts) were enrolled in THIS SCHOOL (the school named on the questionnaire label) in grades K–12 or comparable ungraded levels – *Include only students enrolled in the school named on the questionnaire label. Do NOT include prekindergarten or postsecondary students.* **a. On or about October 1 of THIS SCHOOL YEAR?**	048 _____ Students
b. On or about October 1 of LAST SCHOOL YEAR?	049 _____ Students o☐ School not operating in Fall 1990

Example 12. The Student Enrollment Item from the Public School Questionnaire (Jenkins et al., 1992b).

SECTION 2 — STAFFING PATTERNS — Continued

30. How many employees hold full- or part-time positions in this school in each of the following categories?

If an employee holds a position in more than one of the categories, count that person as part-time in each category that applies.

	FULL-TIME	PART-TIME
INSTRUCTIONAL STAFF		
a. Instructional staff – Instructional aides (paraprofessionals who assist classroom teachers)	255 _____ o☐ None	256 _____ o☐ None
INSTRUCTIONAL SUPPORT STAFF		
b. Instructional coordinators and supervisors (including curriculum specialists)	257 _____ o☐ None	258 _____ o☐ None
c. Librarians/Media specialists	259 _____ o☐ None	260 _____ o☐ None
d. Library/Media center aides	261 _____ o☐ None	262 _____ o☐ None
e. Guidance counselors	263 _____ o☐ None	264 _____ o☐ None
f. Vocational-technical counselors	265 _____ o☐ None	266 _____ o☐ None
SUPPORT SERVICES STAFF		
g. Administrators: (1) Principal(s)	267 _____ o☐ None	268 _____ o☐ None
(2) Vice Principal(s)	269 _____ o☐ None	270 _____ o☐ None
(3) Other managers – e.g., business	271 _____ o☐ None	272 _____ o☐ None
h. Administrative support staff – Clerical and nonmanagerial support staff	273 _____ o☐ None	274 _____ o☐ None
i. Student support services staff – Professionals and supervisory staff providing noninstructional services to students, including health, psychology, social work, or attendance	275 _____ o☐ None	276 _____ o☐ None
j. All other support staff (not reported in other categories, such as health aides, maintenance, bus drivers, security, and cafeteria workers)	277 _____ o☐ None	278 _____ o☐ None

Example 13. The Classification of Employees Item from the Public School Questionnaire (Jenkins et al., 1992b).

PART-TIME EMPLOYEES

26a. How many employees hold part-time positions in this school in each of the following categories? Please read through all of the categories listed below before starting to answer.

INCLUDE AS PART TIME

- o Employees who work part-time at this school only.
- o Employees you share with other schools within or outside of the school district.
- o Employees who perform more than one function at this school, e.g., a teaching principal would be counted once as a part-time teacher and again as a part-time principal.

	PART-TIME
1. Administrators:	
(a) Principals	☐ None or _____
(b) Vice Principal(s)	☐ None or _____
(c) Other managers, such as business	☐ None or _____
2. Instructional coordinators and supervisors, such as curriculum specialists	☐ None or _____
3. Guidance counselors-- needs a definition	☐ None or _____
4. Vocational-technical counselors-- needs a definition	☐ None or _____
5. Librarians/Media specialists	☐ None or _____
6. Teachers	
Do not include as teachers-- Other employees listed in this item, unless they also teach. Teachers who teach only prekindergarten students	☐ None or _____
7. Student support services staff, such as school psychologists, social workers, occupational therapists, speech therapists, nurses, and truant officers	☐ None or _____
8. Library/media center aides	☐ None or _____
9. Teacher aides	☐ None or _____
10. Student teachers	☐ None or _____
11. Clerical and nonmanagerial support staff	☐ None or _____

Example 14a. Redesigned Version of the Classification of Employees Item from the Public School Questionnaire, Part A, Part-Time Status (Jenkins et al., 1992b).

FULL-TIME EMPLOYEES

26b. **How many employees hold full-time positions in this school in each of the following categories?** Please read through all of the categories listed below before starting to answer.

		FULL-TIME
1.	Administrators:	
	(a) Principals	☐ None or ____
	(b) Vice Principal(s)	☐ None or ____
	(c) Other managers, such as business	☐ None or ____
2.	Instructional coordinators and supervisors, such as curriculum specialists	☐ None or ____
3.	Guidance counselors-- needs a definition	☐ None or ____
4.	Vocational-technical counselors-- needs a definition	☐ None or ____
5.	Librarians/Media specialists	☐ None or ____
6.	Teachers Do not include as teachers-- Other employees listed in this item. Teachers who teach only prekindergarten students	☐ None or ____
7.	Student support services staff, such as school psychologists, social workers, occupational therapists, speech therapists, nurses, and and truant officers	☐ None or ____
8.	Library/media center aides	☐ None or ____
9.	Teacher aides	☐ None or ____
10.	Student teachers	☐ None or ____
11.	Clerical and nonmanagerial support staff	☐ None or ____
12.	Cafeteria workers	☐ None or

Example 14b. Redesigned Version of the Classification of Employees Item from the Public School Questionnaire, Part B, Full-time Status (Jenkins et al., 1992b).

STUDENT CLASS SCHEDULES WITH SELECTED TEACHERS

JOHN JONES

	1. Excluding study halls and free periods, does John Jones currently have this student for class?	2. Excluding study halls and free periods, please list the classes in which John Jones currently has this student.	3. How many times per week does this class meet?	4. How many total students are enrolled in this class?
KAYE STEWART	[ ] yes	1.		
		2.		
		3.		
	[ ] no			
MARIE LEARY	[ ] yes	1.		
		2.		
		3.		
	[ ] no			
SUZANNE FLANIGAN	[ ] yes	1.		
		2.		
		3.		
	[ ] no			

Example 15. Questions Developed for the Student Records Questionnaire Using a Matrix Format (Jenkins and Ciochetto, 1993).

STUDENT 1's NAME

1a. Excluding homeroom, study halls, and free periods, do you currently teach this student?

010 1 ☐ Yes → 2 ☐ No → Skip to Item 2a

b. Do you teach multiple subjects to this student all or most of the day?

011 1 ☐ Yes → Skip to Item 2a 2 ☐ No →

c. Excluding homeroom, study halls, and free periods, please list the classes that you teach this student and the number of times per week that each class meets.

Class name	Meetings per week

STUDENT 2's NAME

2a. Excluding homeroom, study halls, and free periods, do you currently teach this student?

012 1 ☐ Yes → 2 ☐ No → Skip to Item 3a

b. Do you teach multiple subjects to this student all or most of the day?

013 1 ☐ Yes → Skip to Item 3a 2 ☐ No →

c. Excluding homeroom, study halls, and free periods, please list the classes that you teach this student and the number of times per week that each class meets.

Example 16. Redesigned Version of the Questions Developed for the Student Records Questionnaire Using a Single Question Format (Jenkins and Ciochetto, 1993).

SECTION 1 — SCHOOL CHARACTERISTICS — Continued

23a. Does this school offer a general program for students who do not plan to attend college?	214	1☐Yes – *Continue with b* 2☐No – *Skip to item 24a*
b. How many students in grades 10–12 are enrolled in this program?	215	_____ Students 0☐None
24a. LAST SCHOOL YEAR, how many students were enrolled in 12th grade?	216	_____ Students – *Continue with b* 0☐No 12th graders in 1990–91 – *Skip to the note above item 26a*
b. How many students were graduated from the 12th grade last year? Include 1991 summer graduates.	217	_____ Students – *Continue with c* 0☐None – *Skip to the note above item 26a*
c. How many of last year's graduates applied to two- or four-year colleges?	218	_____ Graduates 0☐None
25a. Does this school offer job placement services for graduating seniors?	219	1☐Yes 2☐No
b. Does this school have a "Tech-Prep" program, i.e., vocational-technical instruction in the last two years of high school designed to prepare students for two years of vocational instruction at the postsecondary level?	220	1☐Yes 2☐No

SECTION 2 — STAFFING PATTERNS

NOTE: *For items 26–28, INCLUDE itinerant teachers and long-term substitutes. DO NOT INCLUDE student teachers, teacher aides, short-term substitutes, teachers who teach ONLY prekindergarten or postsecondary students, and other non-teaching staff (administrators, other professionals such as counselors and librarians, and support staff) unless they also teach part-time. Report in head counts, not FTEs.*

26a. How many K–12 teachers have FULL-TIME teaching positions at THIS school?	221	_____ Full-time teachers 0☐None
b. How many K–12 teachers have PART-TIME teaching positions at THIS school?	222	_____ Part-time teachers 0☐None
c. What is the total number of K–12 teachers at this		

Example 17. Transitional Heading "Section 2 -- Staffing Patterns" Falling in the Middle of a Page on the Public School Questionnaire (Jenkins et al., 1992b).

SECTION 1 — SCHOOL CHARACTERISTICS — Continued

NOTE – Items 2–7 refer to grades K through 12. Do not include prekindergarten or postsecondary students or grades in answering these questions.

2. How many students (in head counts) were enrolled in THIS SCHOOL (the school named on the questionnaire label) in grades K–12 or comparable ungraded levels – *Include only students enrolled in the school named on the questionnaire label. Do NOT include prekindergarten or postsecondary students.*		
a. On or about October 1 of THIS SCHOOL YEAR?	048	_____ Students
b. On or about October 1 of LAST SCHOOL YEAR?	049	_____ Students o☐ School not operating in Fall 1990
3. How many K–12 students in this school are – *Do NOT include prekindergarten or postsecondary students.*		
a. American Indian, Alaskan Native *(Aleut, Alaskan Indian, Yupik, Inupiat)*?	050	_____ Students o☐ None
b. Asian or Pacific Islander *(Japanese, Chinese, Filipino, Korean, Asian Indian, Vietnamese, Hawaiian, Guamanian, Samoan, other Asian)*?	051	_____ Students o☐ None
c. Hispanic, regardless of race *(Mexican, Puerto Rican, Cuban, Central or South American, or other Hispanic culture or origin)*?	052	_____ Students o☐ None
d. Black (not of Hispanic origin)?	053	_____ Students o☐ None
e. White (not of Hispanic origin)?	054	_____ Students o☐ None
4. What percent of K–12 students enrolled in this school are male? *Record the percent in whole numbers, not tenths; do not enter a decimal point. Do NOT include prekindergarten or postsecondary students.*	055	_____ .0 %
5. How many K–12 students were absent the most recent school day? *Include both excused and unexcused absences. Do NOT include prekindergarten or postsecondary students.*	056	_____ Students o☐ None
6. How many days are in the school year for students in this school?		

Example 18. Transitional Heading and Advanced Instruction Coming at the Top of a Page on the Public School Questionnaire (Jenkins et al., 1992b).

29a. Were there teaching vacancies in this school for this school year, i.e., teaching positions for which teachers were recruited and interviewed?	230	1 ☐ Yes – *Continue with b* 2 ☐ No – *Skip to item 30*
b. Did this school have any teaching vacancies this school year that could not be filled with a teacher qualified in the course or grade level to be taught?	231	1 ☐ Yes – *Continue with c* 2 ☐ No – *Skip to item 29d*
c. Which of these methods did this school use to cover the vacancy(ies)? *Mark (X) all that apply.*	232	1 ☐ Cancelled planned course offerings
	233	2 ☐ Expanded some class sizes
	234	3 ☐ Added sections to other teachers' normal teaching loads
	235	4 ☐ Assigned a teacher of another subject or grade level to teach those classes
	236	5 ☐ Used long-term and/or short-term substitutes
	237	6 ☐ Used part-time or itinerant teachers
	238	7 ☐ Hired a less qualified teacher
	239	8 ☐ Other – *Specify* _____

Example 19. Conventional Skip Instruction from the Public School Questionnaire (Jenkins et al., 1992b).

1. Have you ever drunk a beer, a glass of wine or a wine cooler, a shot of liquor, or a mixed drink with liquor in it? If you have only had sips from another person's drink, answer "no."
MARK ONE BOX.

1 ☐ Yes → a. Think about the very first time you drank a beer, glass of wine or wine cooler, a shot of liquor, or a mixed drink. Not counting sips you might have had from someone else's drink, how old were you the first time you drank an alcoholic beverage?

_____ years old

CONTINUE WITH QUESTION 2.

2 ☐ No

GO TO QUESTION 2.

YOUR 12-MONTH REFERENCE DATE IS: _____

5. The next questions are about the past 12 months—the period from the date written above up to and including today.

Did you smoke a cigarette during the past 12 months?
MARK ONE BOX.

1 ☐ Yes → a. During the past 12 months, have you smoked cigarettes every day or almost every day for two or more weeks in a row?
MARK ONE BOX.

1 ☐ Yes 2 ☐ No

b. During the past 12 months, have you felt that you needed or were dependent on cigarettes?
MARK ONE BOX.

1 ☐ Yes 2 ☐ No

c. During the past 12 months, have you needed larger numbers of cigarettes to get the same effect?
MARK ONE BOX.

1 ☐ Yes 2 ☐ No

d. During the past 12 months, have you have you tried to cut down on your use of cigarettes?
MARK ONE BOX.

1 ☐ Yes 2 ☐ No

e. During the past 12 months, have you felt sick or had withdrawal symptoms because you stopped or cut down on cigarettes?
MARK ONE BOX.

1 ☐ Yes 2 ☐ No

2 ☐ No

GO TO QUESTION 6 ON PAGE 6.

CONTINUE WITH QUESTION 6 ON PAGE 6.

Example 20. Branching Instructions in Questions 1 and 5 of the 1990 National Household Survey and Drug Abuse Questionnaire (Turner et al., 1992).

SECTION 1 – SCHOOL CHARACTERISTICS – CONTINUED

11. Please indicate whether each of the following programs or services is currently available at this school **to the students included in item 6**, either during or outside of regular school hours and regardless of funding source.

a. English as a Second Language – Students with limited English proficiency are provided with intensive instruction in English.

051 1 ☐ Yes
2 ☐ No

SKIP to b.

How many students participate in this program?

052 0 ☐ None or _____ Students

b. Bilingual education – Native language is used to varying degrees in instructing students with limited English proficiency. *For example, transitional bilingual education and structured immersion. Do not include foreign language classes or foreign language immersion programs.*

053 1 ☐ Yes
2 ☐ No

SKIP to c.

How many students participate in this program?

064 0 ☐ None or _____ Students

Example 21. "Salient" Skip Instruction (adapted from Jenkins and Ciochetto, 1993).

PART A – Employment Status During the Week of April 12–18, 1992

A1. Were you working for pay or profit during the week of April 12–18, 1992? This includes being self-employed or temporarily absent from a job (e.g., illness, vacation or parental leave), even if unpaid.

PGM2

001
1 ☐ Yes – *Skip to A6*
2 ☐ No

A2. Did you look for work at any time during the 5 weeks between March 8 and April 12, 1992?

002
1 ☐ Yes
2 ☐ No

A3. What was your MAIN reason for not working during the week of April 12–18?

Mark (X) One

003
1 ☐ Retired – *Skip to A5*
2 ☐ On layoff from a job
3 ☐ My work is seasonal
4 ☐ Student
5 ☐ Family responsibilities
6 ☐ Chronic illness or permanent disability
7 ☐ Could not find work or believed no suitable jobs available in my field
8 ☐ Waiting for new job to begin within 30 days
9 ☐ Waiting for school to begin
10 ☐ Did not need or want to work
11 ☐ Other *Specify* ✗

A4. Had you previously RETIRED from any position (e.g., mandatory retirement or early retirement)?

004
1 ☐ Yes
2 ☐ No – *Skip to Part B on Page 4*

A5. When did you retire?

Month Year

005 _____ 19 _____

NOW SKIP TO PART B ON PAGE 4

A6. During the week of April 12–18, 1992 were you working full time or part time?

009
1 ☐ Full time (usually worked a total of 35 or more hours per week) – *Skip to A9*
2 ☐ Part time (usually worked less than 35 hours per week)

A7. Were you seeking full-time work during the week of April 12–18, 1992?

010
1 ☐ Yes
2 ☐ No

A8. What was your MOST important reason for holding a part-time position during the week of April 12–18, 1992?

Mark (X) One

011
1 ☐ Full-time position not available
2 ☐ Worked part time to accommodate spouse's/partner's job or career
3 ☐ Worked part time for other family-related reasons
4 ☐ Preferred part-time position for other reason *Specify* ✗

A9. Although you were working during the week of April 12–18, 1992, had you previously RETIRED from any position (e.g., mandatory retirement, early retirement)?

012
1 ☐ Yes
2 ☐ No – *Skip to A11*

A10. When did you retire?

Month Year

013 _____ 19 _____

A11. For whom did you work during the week of April 12–18, 1992? (IF YOU HAD MORE THAN ONE JOB THAT WEEK: Please answer for the job you considered your principal employment.)

Employer Name

014 _____

Street

_____

City/Town

_____

State/Foreign Country

_____

015 x ☐ *MARK (X) HERE IF YOU WERE SELF-EMPLOYED*

Example 22. "Intermediate" Skip Instruction from the Survey of College Graduates

Fill in the first student's name from the cover page on the line below.

STUDENT 1's NAME

1. What is this student's current status at this school?

Enrolled ☐1
Suspended ☐2
} Skip to 2

Expelled ☐3
Transferred ☐4
Dropout/Chronic Truant (See definition below) ☐5
Deceased ☐6
} Skip to NEXT STUDENT ON PAGE ...

Other ☐7 Specify _____ Skip to 2

2. Is this student male or female?

Male ☐1
Female ☐2

3. What is this student's race/ethnicity?

American Indian or Alaskan Native ☐1
Asian or Pacific Islander ☐2
Hispanic, regardless of race ☐3
Black (not of Hispanic origin) ☐4
White (not of Hispanic origin) ☐5

4a. Excluding homeroom, study halls, and free periods, is this student currently taught by [Teacher 1]?

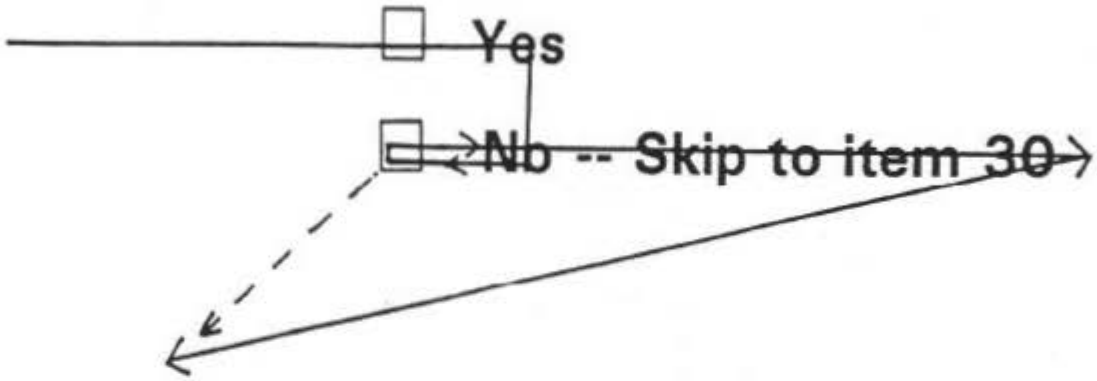
Yes ☐1 Skip to 4b
No ☐2 Skip to 5a

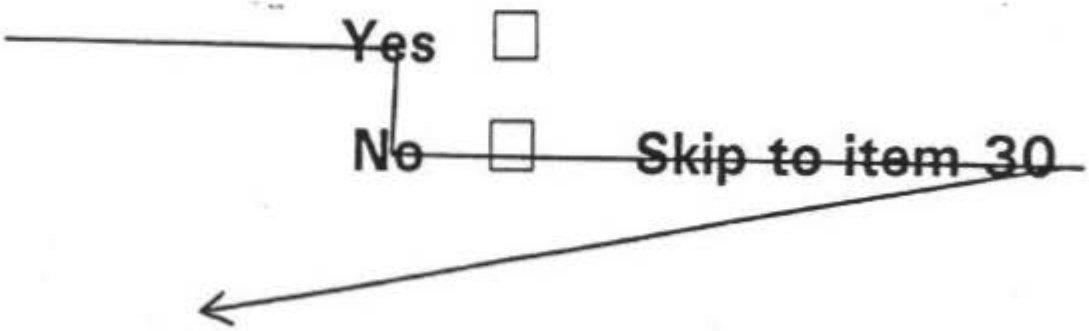
4b. Does [Teacher 1] teach this student all or most of the day?

Yes ☐1 Skip to 5a
No ☐2 Skip to 4c

4c. ...

Example 23. "Natural Reading Sequence" Skip Instruction (Jenkins and Ciochetto, 1993).





Example 24. Diagrammatical Representation of the Conventional (Top View) and "Natural Reading Sequence" (Bottom View) Skip Instruction Reading Structures.

FORM IPEDS-IC-2
(9-1-91)

U.S. DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS
ACTING AS COLLECTING AGENT FOR THE
U.S. DEPARTMENT OF EDUCATION
NATIONAL CENTER FOR EDUCATION STATISTICS

INTEGRATED POSTSECONDARY EDUCATION DATA SYSTEM

INSTITUTIONAL CHARACTERISTICS SURVEY 1991-92

OMB No. 1850-0582: Approval Expires 2/29/92

NOTE — This form is authorized by law (20 U.S.C. 1221e-1). While you are not required to respond, your cooperation is needed to make the results of this survey comprehensive, accurate, and timely.

Public reporting burden for this collection of information is estimated to average 1.0 hours per response but may range from 30 minutes to 2.0 hours depending on whether the information is readily accessible to machine readable files. The burden estimate includes the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to the U.S. Department of Education, Information Management and Compliance Division, Washington, DC 20202-4651, and to the Office of Management and Budget, Paperwork Reduction Project 1850-0582, Washington, DC 20503.

Please read the accompanying instructions before completing this survey form. Respond to each item on this report in the space provided. Certain responses are preprinted. These responses were provided by your institution on the previous IPEDS Institutional Characteristics Survey form. If a response is preprinted, verify that it is correct. If a preprinted response is incorrect, cross out the existing, incorrect response with a single line and clearly indicate the correct response. ***MAKE YOUR CHANGES IN RED*** *so they are easily identified. Be sure to update the enrollment and tuition questions. Certain terms in the various questions are defined in the Glossary which begins on page 5 of the instructions.*

RETURN TO

If there are any questions about this form, contact Elaine Kroe, NCES, at (202) 219-1361 or the Bureau of the Census IPEDS representative at (301) 763-4947.

Date due: September 1, 1991

INSTITUTIONAL IDENTIFICATION 1991-92

Please correct errors in the name, address, and ZIP Code listed below.

1. Name of institution covered by this report 001

2. UNITID 002

3. Address *(Number and street name)* 003

4. Name of county or independent city 004

City 005 | State 006 | ZIP Code 007 | **5.** Congressional district 008

6. Name of chief administrator 009 | Title 010

7. Name of respondent 011 | Telephone number 012

8. Telephone numbers ▶ General information 013 | Financial aid office 014 | Admissions office 015

Part A — TYPES OF EDUCATIONAL OFFERINGS

1. Which of the following types of instruction/programs does your institution offer?
Mark (X) all that apply.

Example 25. Cover Page of the 1991-92 Integrated Postsecondary Education Data System Questionnaire.

Part II — EDUCATION AND TRAINING — Continued

11. If you are a student attending a college or university, mark your status.

132
1 ☐ Student, full-time
2 ☐ Student, part-time
3 ☐ Not currently a student

12a. Which of these kinds of training did you participate in during 1980 or 1981?

Mark (X) the appropriate year for each type of training you received.

1980	1981	KIND OF TRAINING
133 1 ☐	134 1 ☐	Military training applicable to present civilian occupation
2 ☐	2 ☐	Extension or correspondence courses applicable to present civilian occupation
3 ☐	3 ☐	Courses at employer's training facility
4 ☐	4 ☐	Courses at adult education center
5 ☐	5 ☐	Courses presented in conjunction with professional meetings
6 ☐	6 ☐	Courses presented by professional training organizations (commercial or non-profit)
7 ☐	7 ☐	Other training
0 ☐	0 ☐	None

b. Were continuing education units (CEU's) or other forms of recognized credit units earned as a result of the above training (in Item 12a)?

1980	1981
135 1 ☐ Yes	136 1 ☐ Yes
2 ☐ No	2 ☐ No

Part III — EMPLOYMENT STATUS

13. During the week of May 9, 1982, were you —

137
1 ☐ Working full time (35 hours or more per week in at least one position) — *SKIP to 17a*
2 ☐ Working part time — *GO to 14*
3 ☐ Not working, but seeking work — *SKIP to Part IV*
4 ☐ Not working and not seeking work — *SKIP to 15*

14. Were you seeking full-time work?

138
1 ☐ Yes
2 ☐ No
} *SKIP to 17a*

15. Did you look for work at any time during the 3 weeks PRIOR to the week of May 9, 1982?

139
1 ☐ Yes
2 ☐ No

16. What was the MAIN reason you were not working or not seeking work during the week of May 9, 1982?

Mark (X) only one box.

140
1 ☐ On layoff from a job
2 ☐ On vacation or otherwise temporarily absent from a job for health or personal reasons
} *GO to 17a*

3 ☐ Retired
4 ☐ Student
5 ☐ Family responsibilities
6 ☐ Chronic illness or permanent disability
7 ☐ Could not find work or believed no jobs available in my particular field
8 ☐ Did not want to work
9 ☐ New job to begin within 30 days
10 ☐ Waiting for school to begin
11 ☐ Other — *Specify* _____
} *SKIP to Part IV*

17a. During the week of May 9, 1982, were you working at (or on layoff from) a position related to the natural sciences, social sciences, or engineering?

141
1 ☐ Yes — *SKIP to Part IV*
2 ☐ No — *GO to b*

Example 26. A Page from a Conventionally Designed Questionnaire.

C7. Do you currently belong to any national professional societies or associations?

119 1☐Yes

2☐No

C8. In the 2 years between April 1990 and April 1992, did you attend any work-related workshops, seminars, or other work-related training activities?

- *Do **not** include college courses – these will be discussed in Part D.*
- *Do **not** include professional meetings unless you attended a special training session conducted at the meeting/conference.*

120 1☐Yes

2☐No – *Skip to C12*

C9. In which of the following areas did you attend work- related workshops, seminars, or other work-related training activities?

Mark (X) Yes or No for each	Yes	No
121 **a.** Management or supervisor training. .	1☐	2☐
122 **b.** Technical training in my occupational field	1☐	2☐
123 **c.** General professional training (e.g., public speaking, business writing). . .	1☐	2☐
124 **d.** Other work-related training *Specify*	1☐	2☐

C10. For which of the following reasons did you attend training activities between April 1990 and April 1992?

Mark (X) Yes or No for each	Yes	No
125 **a.** To acquire further skills or knowledge in my current occupational field.	1☐	2☐
126 **b.** To acquire skills or knowledge in a different field	1☐	2☐
127 **c.** For licensure/certification.	1☐	2☐
128 **d.** To increase opportunities for promotion/advancement/higher salary	1☐	2☐
129 **e.** To learn skills or knowledge needed for a recently acquired position (e.g., orientation).	1☐	2☐

C11. What were your TWO most important reasons for attending training activities?

132 ___ Most IMPORTANT REASON from Question C10 – *(Enter LETTER)*

133 ___ SECOND most IMPORTANT REASON from Question C10 – *(Enter LETTER)*

C12. During the 52 weeks of the 1991 calendar year, how many weeks were you –

NUMBER OF WEEKS

134 ___ **Working,** including weeks of **paid** vacation, **paid** sick leave, and military service?

135 ___ **On an unpaid** leave of absence from a job?

136 ___ **Not working** but seeking work?

137 ___ **Not working** and **not** seeking work?

5 2 **TOTAL**

C13. What was your TOTAL EARNED income, BEFORE deductions for 1991?

Include all wages, salaries, bonuses, overtime, commissions, consulting fees, net income from businesses, summertime teaching or research, post-doctoral appointment, or other work associated with scholarships.

138 $ _____.00

Total 1991 earned income

Example 27. A Page from a Redesigned Questionnaire.

PLEASE ALSO ANSWER HOUSING QUESTIONS ON PA

Please fill one column → for each person listed in Question 1a on page 1.	PERSON 1	PERSON 2	
	Last name First name — Middle initial	Last name First name — Middle initial	Last name First name
2. How is this person related to PERSON 1? Fill ONE circle for each person. If **Other relative** of person in column 1, fill circle and print exact relationship, such as mother-in-law, grandparent, son-in-law, niece, cousin, and so on.	START in this column with the household member (or one of the members) in whose name the home is owned, being bought, or rented. If there is no such person, start in this column with any adult household member.	If a RELATIVE of Person 1: ○ Husband/wife ○ Brother/sister ○ Natural-born or adopted son/daughter ○ Father/mother ○ Grandchild ○ Other relative ○ Stepson/ stepdaughter If NOT RELATED to Person 1: ○ Roomer, boarder, or foster child ○ Unmarried partner ○ Housemate, roommate ○ Other nonrelative	If a REL If NOT
3. Sex Fill ONE circle for each person.	○ Male ○ Female	○ Male ○ Female	
4. Race Fill ONE circle for the race that the person considers himself/herself to be. If **Indian (Amer.)**, print the name of the enrolled or principal tribe. If **Other Asian or Pacific Islander (API)**, print one group, for example: Hmong, Fijian, Laotian, Thai, Tongan, Pakistani, Cambodian, and so on. If **Other race**, print race.	○ White ○ Black or Negro ○ Indian (Amer.) (Print the name of the enrolled or principal tribe.) ○ Eskimo ○ Aleut Asian or Pacific Islander (API) ○ Chinese ○ Japanese ○ Filipino ○ Asian Indian ○ Hawaiian ○ Samoan ○ Korean ○ Guamanian ○ Vietnamese ○ Other API ○ Other race (Print race)	○ White ○ Black or Negro ○ Indian (Amer.) (Print the name of the enrolled or principal tribe.) ○ Eskimo ○ Aleut Asian or Pacific Islander (API) ○ Chinese ○ Japanese ○ Filipino ○ Asian Indian ○ Hawaiian ○ Samoan ○ Korean ○ Guamanian ○ Vietnamese ○ Other API ○ Other race (Print race)	
5. Age and year of birth a. Print each person's age at last birthday. Fill in the matching circle below each box. b. Print each person's year of birth and fill the matching circle below each box.	a. Age: 0 0 0 0 0 / 1 0 1 0 1 0 2 0 2 0 3 0 3 0 4 0 4 0 5 0 5 0 6 0 6 0 7 0 7 0 8 0 8 0 9 0 9 0 b. Year of birth: 1 ● 8 0 0 0 0 / 9 0 1 0 1 0 2 0 2 0 3 0 3 0 4 0 4 0 5 0 5 0 6 0 6 0 7 0 7 0 8 0 8 0 9 0 9 0	a. Age: 0 0 0 0 0 / 1 0 1 0 1 0 2 0 2 0 3 0 3 0 4 0 4 0 5 0 5 0 6 0 6 0 7 0 7 0 8 0 8 0 9 0 9 0 b. Year of birth: 1 ● 8 0 0 0 0 / 9 0 1 0 1 0 2 0 2 0 3 0 3 0 4 0 4 0 5 0 5 0 6 0 6 0 7 0 7 0 8 0 8 0 9 0 9 0	a. Age: 0 1
6. Marital status Fill ONE circle for each person.	○ Now married ○ Separated ○ Widowed ○ Never married ○ Divorced	○ Now married ○ Separated ○ Widowed ○ Never married ○ Divorced	

Example 28. Original Census Questionnaire Using a Matrix Format.

THE 1992 NATIONAL CENSUS TEST

Four very brief questions about each person living in your household on April 1, 1992.

Please complete both sides of this questionnaire and return in the postage-paid envelope.

OMB No. 0607-0728
Approval Expires 09/30/92

4916 Livingstone St.
Bremer, WA 90001

A. Please answer the four brief questions for each person living at this address → as of Wednesday, April 1, 1992.

DON'T INCLUDE PERSONS WHO:

- Usually live somewhere else
- Are away in an institution such as a prison, mental hospital, or nursing home
- Are college students living somewhere else while attending college
- Are in the Armed Forces and live somewhere else
- Stay somewhere else most of the week while working

BE SURE TO INCLUDE PERSONS WHO:

- Usually live here, such as family members, housemates and roommates, foster children, roomers, boarders, and live-in employees
- Are temporarily away on a business trip, on vacation, or in a general hospital
- Are college students staying here while attending college
- Are in the Armed Forces and live here
- Are newborn babies still in the hospital
- Are children in boarding schools below the college level
- Stay here most of the week while working, even if they have a home somewhere else
- Have no other home and are staying here on April 1

(Start with the household member, or one of the household members, in whose name this house or apartment is owned, being bought, or rented. If there is no such person, start with any adult household member.)

PERSON 1

NAME _____ (Print) Last _____ First _____ Initial

1. What is this person's sex?

☐ Female
☐ Male

2. What is this person's date of birth?

☐☐ Month ☐☐ Day ☐☐☐☐ Year

(It is important to answer both Questions 3 and 4)

3. Is this person of Spanish/Hispanic origin?

☐ No, not Spanish/Hispanic
☐ Yes, Mexican, Mexican-Am., Chicano
☐ Yes, Puerto Rican
☐ Yes, Cuban
☐ Yes, other Spanish/Hispanic *(Print group)* → _____

4. What is this person's race?

☐ White
☐ Black or Negro
☐ American Indian *(Print name of enrolled or principal tribe)* → _____
☐ Eskimo
☐ Aleut
☐ Chinese
☐ Japanese
☐ Filipino
☐ Asian Indian
☐ Hawaiian
☐ Samoan
☐ Korean
☐ Guamanian
☐ Vietnamese
☐ Other Asian or Pacific Islander *(Print race)* → _____
☐ Some other race *(Print race)* → _____

PERSON 2

NAME _____ (Print) Last _____ First _____ Initial

4. What is this person's race?

☐ White
☐ Black or Negro
☐ American Indian *(Print name of enrolled or principal tribe)* → _____

Example 29. Redesigned Census Questionnaire Using An Individual Space Format.

DUE BY JANUARY 1, 1991

OMB No. 0607-0672 Approval Expires 06 30 91

FORM 90-A17

U.S. DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS

UNITED STATES CENSUS OF AGRICULTURE

CENSUS

NOTICE — Response to this inquiry is required by law (title 13, U.S. Code). By the same law YOUR REPORT TO THE CENSUS BUREAU IS CONFIDENTIAL. It may be seen only by sworn Census employees and may be used only for statistical purposes. Your report CANNOT be used for purposes of taxation, investigation, or regulation. The law also provides that copies retained in your files are immune from legal process.

In correspondence pertaining to this report, please refer to your Census File Number (CFN)

Please complete this form and RETURN TO: BUREAU OF THE CENSUS, 1201 East Tenth Street, Jeffersonville, IN 47133

Note — If your records are not available, reasonable estimates may be used. If you cannot file by January 1, a time extension request may be sent to the above address. Include your 12-character Census File Number (CFN) as shown in your address label in all correspondence to us.

If you received more than one report form, enter extra Census File Number(s) here and return extra copies with your completed report. A ___ A ___

CENSUS USE ONLY: 035 036 037 038 039 040 041 042

90-A17

CFN 08603545545 EI 86-0354554
6 24 04 10 00 802100 F 8002
NAVAJO 08603545545 4H 00100

L. David Sinclair
4916 Livingstone St.
Bremer, WA 90001

Please correct errors in name, address, and ZIP Code. ENTER street and number if not shown.

SECTION 1

1. At **any** time **during 1990**, did you do any of the following:
   - operate a farm or ranch (including greenhouse and nursery)?
   - sell your farm or ranch?
   - retire from farming or ranching?

   1 ☐ Yes 2 ☐ No

2. At **any** time **during 1990**, did you grow or raise any:
   - crops (including **hay**, tobacco, fruits, vegetables, nuts, berries, etc.) for **sale**?
   - livestock or poultry (including **horses**, cattle, hogs, sheep, goats, fish, etc.) for **sale**?
   - crops, livestock, or poultry for **home use** or **as a hobby**?

   3 ☐ Yes 4 ☐ No

If you answered YES to EITHER of these questions, go to SECTION 2.
If you answered NO to BOTH of these questions, go to SECTION 10.

SECTION 2 — ACREAGE IN 1990

Report land owned, rented, or used by you, your spouse, or by the partnership, corporation, or organization for which you are reporting. *Include ALL LAND, REGARDLESS OF LOCATION OR USE — cropland, pastureland, rangeland, woodland, idle land, house lots, etc.*

	None	Number of acres
1. All land owned	☐	043
2. All land rented or leased FROM OTHERS, including land worked by you on shares, used rent free, in exchange for services, payment of taxes, etc. *Include leased Federal, State, and railroad land. (DO NOT include land used on a per-head basis under a grazing permit.)*	☐	044
3. All land rented or leased TO OTHERS, including land worked on shares by others and land subleased. *Also complete item 5 below.*	☐	045
4. **Acres in "THIS PLACE"** — *ADD acres owned (item 1) and acres rented (item 2), then SUBTRACT acres rented TO OTHERS (item 3), and enter the result in this space.* →		046

If the entry is zero, please refer to the Information Sheet, section 2.

5. Of the land you rented or leased to others, how many acres did you own? ☐ None 053 _____ Acres

6. In what county was the largest value of your agricultural products raised or produced? County name ____ State ____

SECTION 3 — LAND USE and IRRIGATION

• PART A — How were the ACRES in this place used in 1990?

	None	Number of acres
1. Cropland harvested — *Include all land from which crops were harvested or hay was cut, and all land in orchards, citrus groves, vineyards, and nursery and greenhouse crops.*	☐	161
2. Cropland on which all crops failed — *(Exception: Do not report here land in orchards and vineyards on which the crop failed.)*	☐	162
3. Cropland idle, cropland used for cover crops, or cropland in cultivated summer fallow	☐	163
4. Cropland used only for pasture, woodland pastured, and other pastureland and rangeland	☐	165
5. All other woodland, wasteland, houselots, etc. not reported in items 1 through 4 above	☐	167

SECTION 4

• PART A — CROPS HARVESTED from "THIS PLACE" in 1990
(Do not include crops grown on land rented to others.)

	None	Acres harvested	Quantity harvested	Gross value of crops sold: Dollars	Cents
1. Hay crops —					
a. Alfalfa and alfalfa mixtures	☐	105	104 Tons dry	143 $	00
b. Small grain hay	☐	106	107 Tons dry	147 $	00
c. Wild hay	☐	112	113 Tons dry	143 $	00
d. Other hay — Specify kind ____	☐	109	110 Tons dry	143 $	00
2. Corn for grain or seed	☐	087	088 Bu	115 $	00
3. Soybeans for beans	☐	086	085 Bu	115 $	00
4. Wheat for grain	☐	073	074 Bu	174 $	00
5. Tobacco — all types	☐	094 .10	095 Lbs	141 $	00
6. Potatoes, Irish — *(Do not include those grown for home use.)*	☐	087 .10	088 Cwt	186 $	00

7. All vegetables for sale *(Do not include those grown for home use.)*	None	Total acres	Dollars	Cents
	☐	175 .10	143 $	00
Specify kind(s)		.10		
		.10		

8. All fruit and nut orchards, vineyards, and berries	None	Total acres	Quantity harvested	Dollars	Cents
	☐	176 .10		144 $	00
Specify kind(s)		.10	Lbs		
		.10	Lbs		

9. Other crops — *For additional crops, enter the crop name and code from the list below. Report quantity harvested in the unit specified with crop name.*

Crop name	Code	Acres harvested	Quantity harvested	Gross value of crops sold: Dollars	Cents
				$	00
				$	00

If more space is needed use a separate sheet of paper.

Crop name	Code	Crop name	Code
Barley for grain (bushels)	079	Oats for grain (bushels)	078
Corn for silage or green chop (tons, green)	070	Sorghum for grain-milo (bushels)	082
Cotton (bales)	081	Other crops (pounds) — *Specify*	752

• PART B — NURSERY and GREENHOUSE CROPS GROWN FOR SALE on "THIS PLACE" in 1990

From the list below, enter the crop name and code for each crop grown.

Crop name	Code	Square feet under glass or other protection	Acres in the open in 1990: Whole acres	Tenths	Sales in 1990: Dollars	Cents

Example 30. Cover Page of the 1990 Census of Agriculture Questionnaire (Photo-reduced to 74 Percent of Original Size).

DISCUSSION

Jared B. Jobe
National Center for Health Statistics

These excellent papers approach the topic of self-administered questionnaires from different perspectives: Lessler and O'Reilly discuss the use of audio computer-assisted self-administered questionnaires for sensitive questions; Jenkins and Dillman describe the use of graphic design principles, as well as cognitive and motivational factors in the design of paper and pencil self-administered questionnaires. Its a privilege to have the opportunity to discuss their results, relate their findings to the literature, and offer some suggestions for future work.

Presentation by Lessler and O'Reilly

Lessler and O'Reilly describe the evidence that self-administered questionnaires result in more reports of sensitive behaviors. They describe the disadvantages of self-administered questionnaires; some of these problems were discovered by cognitive testing. They describe computer-assisted self-administered interviews (CASI), and audio CASI data collection, as well as some of audio CASI's advantages. Finally, they presented the results of an experiment conducted during the National Survey of Family Growth (NSFG) Cycle 5 Pretest.

The first issue that I'll address is respondents' ratings of their preferences of method of administration. Lessler and O'Reilly describe a small-scale study by O'Reilly, Hubbard, Lessler, Biemer, and Turner (forthcoming) showing that respondents preferred the audio CASI method. A greco-latin square design was used. Thus, all respondents' ratings were unbiased for the three self-administration methods tested. In the present study by Lessler and O'Reilly, respondents also recommended audio CASI for reporting abortions. However, such a finding should be interpreted with caution. Respondents could only compare audio CASI to the computer-assisted interview. Contrast that result with those from a collaborative study between the National Center for Health Statistics (NCHS) and the National Opinion Research Center (NORC) (Jobe, Pratt, Tourangeau, Baldwin, and Rasinski, forthcoming). In the NCHS-NORC study, focus group respondents recommended focus groups for collecting sensitive information, and respondents in a CASI pretest recommended CASI. Both the Lessler and O'Reilly study and the NCHS-NORC study lack appropriate comparison groups.

All these studies impose large demand characteristics on respondents (see e.g., Orne, 1969). Therefore, my conclusion is that respondents will usually prefer the administration method that they have just experienced. Thus, the audio CASI may not be as strongly preferred as Lessler and O'Reilly's results would suggest.

A second issue is whether method of administration effects occur when collecting data about sensitive topics. This is a

timely question. The O'Reilly *et al.* study, described by Lessler and O'Reilly, found that the two CASI methods, audio and video, produced more reports of marijuana and cocaine use than the paper and pencil self-administered questionnaire. Few differences in sexual behaviors were noted. In Lessler and O'Reilly's study, the 178 respondents reported 48 abortions when using audio CASI, compared to 42 abortions during the regular interview.

This result can be contrasted with results of the NCHS-NORC study mentioned above (Jobe *et al.*, forthcoming). We crossed computer-assisted and paper and pencil interview modes with interviewer and self administration. Computerization had no effects on a variety of reports of sensitive questions. Self administration, however, resulted in more reports of some sensitive behaviors: As shown in the top panel of Table 1, more sex partners were reported for the last year, last five years, and lifetime with self administration. As shown in the middle panel of Table 1, more condom use was reported in the last 30 days and past year (marginally significant) with self administration. As shown in the bottom panel of Table 1, more respondents reported a sexually transmitted disease with self administration. Neither computerization nor self administration had any main effects on reports of abortions or drug use.

Table 1

Mean sexual partners and rates of sexually transmitted diseases by method of administration

	Method of Administration		
	Self-Administered	Administered by Interviewer	Ratio
Number of Sexual Partners			
Past Year	1.71	1.44	1.19
Past Five Years	3.87	2.82	1.37
Lifetime	6.51	5.43	1.20
Condom Use			
Past 30 Days	46.7%	35.3%	1.32
Past Year	23.8%	17.9%	1.33
Sexually Transmitted Diseases	22.0%	17.0%	1.29

Source: National Center for Health Statistics

Recently, Boekeloo, Schiavo, Rabin, Conlon, Jordan, and Mundt (1994) reported that patients at a sexually transmitted disease clinic reported more high-risk sex behaviors to 2 of 16 questions for audio CASI compared to a written self-administered questionnaire. Both were superior to a face-to-face interview. They also found fewer missing responses with the audio CASI.

Thus, the studies I have discussed and several others, indicate that self-administered questionnaires may result in more reports of sensitive behaviors than interviewer-administered questionnaires. However, not all sensitive behaviors are reported more frequently in every study showing self-administration effects. Moreover, there is some evidence that computer or audio CASI is slightly superior to other forms of self-administered questionnaires. We clearly need more studies comparing different methods of administration, especially comparing written, video, and audio self administration. My hypotheses are that, across studies, the effects will be small and equivocal among those three, but that effects will be larger and more consistent for self administration over interviewer administration.

A third issue raised by Lessler and O'Reilly's presentation is the use of incentives in sensitive surveys. They manipulated incentives as a variable, in addition to audio CASI. Table 2 shows additional data from their study, reported by Mosher and Duffer (1994). Self administration and incentives worked in an additive fashion: The highest percentages of respondents reported an abortion with a $20 incentive and audio CASI (30%), and with a $40 incentive and no audio CASI (29%). Intermediate levels of abortion reporting were found with audio CASI and no incentive (25%), and with a $20 incentive and no audio CASI (22%). The lowest level of reporting was found with no audio CASI and no incentive (14%).

Table 2

National Survey of Family Growth Cycle 5 Pretest

Group	% Reporting Abortion	No. in Group
In-home, No $, No Audio CASI	14%	(n = 96)
In-home, $20, No Audio CASI	22%	(n = 72)
In-home, No $, Audio CASI	25%	(n = 98)
In-home, $20, Audio CASI	30%	(n = 80)
Off Site, $40, No Audio CASI	29%	(n = 147)

Source: National Center for Health Statistics

In the NCHS-NORC study (Jobe *et al.*, forthcoming), incentives were used only with neutral site interviews, and respondents interviewed in their own home were not paid. We found no main effects on the incentive/site variable for any of the tested sensitive behaviors. Thus, in one study incentives had an effect, and in another study incentives had no effect. Clearly, more research is needed on the effects that incentives have on responding to sensitive questions.

Presentation by Jenkins and Dillman

Jenkins and Dillman presented 20 principles for designing self-administered questionnaires. In her conclusions, Ms. Jenkins states, "Little information on the design of self-administered questionnaires existed until relatively recently. That which did was based primarily on common sense and individual experience." What is significant and interesting about this statement is that it is so close to statements researchers made about the design of interviewer-administered questionnaires before cognitive psychology began to make an impact a decade or so ago.

I am excited by their approach. Their principles have a high degree of face validity. Potentially, attention to graphic design features as well as cognitive and motivational factors could improve self-administered questionnaires as much as cognitive interviews have improved interviewer-administered questionnaires. However, in order for this to occur, two major differences must be overcome between how cognitive psychologists approached questionnaire design and how Jenkins and Dillman have approached self-administered questionnaires.

The first difference is that, from the beginning, cognitive psychologists involved in questionnaire design have utilized the theories and results from cognitive psychology (for reviews, see Jobe and Mingay, 1991; Jobe, Tourangeau, and Smith, 1993). These scientific citations helped convince people that questionnaire design could be more of a science and less of an art. Researchers integrated basic and applied cognitive research on language comprehension, memory encoding and retrieval, frequency and magnitude estimation, heuristics, and decision processes. These are described in articles and books dating back to the beginning of the survey research-cognitive science collaboration (e.g., Hippler, Schwarz, and Sudman, 1987; Jabine, Straf, Tanur, and Tourangeau, 1984; Moss and Goldstein, 1979).

A prime example is an excellent discussion by Fred Smith (Smith, 1991). He described how cognitive laboratory research on free recall, on frequency estimation, and on magnitude estimation applied to the respondents' tasks of recalling their previous day's intake, estimating the frequencies with which they eat foods, and estimating the sizes of their portions. Awareness of the literature on cognitive theory and research has resulted in the use of these theories and application of results in questionnaire design research. Researchers who are not cognitive psychologists

have been able to this knowledge to design excellent cognitive experiments on questionnaire design.

Jenkins and Dillman state that a major reason for their papers is to encourage experimental research on the issues raised by their 20 principles. Theirs is a laudable goal, and an attainable one. A paper with these well thought out principles will encourage more research, if it is well grounded scientifically. A necessary next step for Jenkins and Dillman is to integrate significant research, some cognitive, some social, which is applicable to their principles. Relevant research has been conducted on reading comprehension (e.g., Graesser and Bower, 1990), eye movement (e.g., Carpenter and Just, 1983), respondent effort (Krosnick, 1991), politeness (e.g., Grice, 1975), and impression management (e.g., Schwarz, 1993), to name a few areas.

The second difference between this presentation and the cognitive approach is that cognitive psychologists have well described the applicable methodology so that other people can use it. The most prominent example of this is the cognitive interview (e.g., Lessler, Tourangeau, and Salter, 1989; Willis, Royston, and Bercini, 1991). Although different types of cognitive interviews are used, the one most frequently used in questionnaire design is the concurrent think aloud with probes. This methodology has been described sufficiently so that the largest federal statistical agencies, university survey laboratories, and private survey organizations now use cognitive interviews and do it well. It is not very difficult to learn, although there are individual differences in skill at conducting cognitive interviews.

Jenkins and Dillman have NOT described appropriate techniques for all their principles so that others can use them. For example, they refer to graphic design principles in their paper. But, after reading this paper and a much longer version of the same paper--I am unable to describe these graphic design features. In Principle two they state about Example 3, "This cover page uses natural reading format and graphical design features." The same problem occurs on other principles such as numbers 11 and 12. Principle 11 uses the same two terms, and yet they are never defined. For these principles to be helpful, they must not merely give examples of how the Census Bureau successfully solved questionnaire design problems for a particular survey (I am impressed with their success), but they must educate people so that they can use them on their own questionnaires.

Several of the principles ARE self explanatory and easy to implement. For example, I developed a solution to the problem of multi-task formats identified in Principle 8 and illustrated in Examples 13 and 14. In the solution described by Jenkins and Dillman, the respondent must still perform two mental calculations at a time. In contrast, another solution would be to ask respondents to report the total number of employees in each category, such as teachers, guidance counselors, and teachers aides. Then the respondent can be asked to divide the employees in

each category into full time and part time workers. Not only does this solution require the respondent to perform only one mental task at a time, but it more closely matches how the information is likely to be organized in the respondent's long-term memory. The solution I just described also illustrates my earlier point that knowledge of relevant scientific literature can make these principles more effective.

A third issue, and one that illustrates the effectiveness of the principles, is the split-ballot experiment. The one described by Jenkins and Dillman used 5 experimental questionnaires, and is a dramatic example of how these principles can be tested experimentally and shown to be effective. This study demonstrated that large structural and organization resulted in large improvements in item and response rates. Note also that smaller changes produce smaller results.

I can conclude by stating that these 20 principles have the potential to revolutionize the design of self-administered questionnaires. However, the long-term effectiveness of these principles may be determined, at least in part, by how their scientific underpinnings are explicated, and by how their everyday use is described.

ACKNOWLEDGEMENT: The author is greatly indebted to Douglas Herrmann for his many valuable suggestions on an earlier version of this paper.

REFERENCES

Boekeloo, B. O., Schiavo, L., Rabin, D. L., Conlon, R. T., Jordan, C. S. and Mundt, D. J. (1994). Self-reports of HIV risk factors by patients at a sexually transmitted disease clinic: Audio vs. written questionnaires. *American Journal of Public Health, 84*, 754-760.

Carpenter, P. A. and Just, M. A. (1983). What your eyes do while your mind is reading. In K. Rayner (Ed.), *Eye movements in reading: Perceptual and language processes*. New York: Academic Press.

Graesser, A. C. and Bower, G. H. (1990). (Eds.). *The psychology of learning and motivation: Inferences and text comprehension*. New York: Academic Press.

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.), *Syntax and semantics: Speech acts 3* (pp. 41-58). New York: Academic Press.

Hippler, H. J., Schwarz, N. and Sudman, S. (Eds.). (1987). *Social information processing and survey methodology*. New York: Springer-Verlag.

Jabine, T. B., Straf, M. L., Tanur, J. M. and Tourangeau, R. (Eds). (1984). *Cognitive aspects of survey methodology: Building a*

bridge between disciplines. Washington, DC: National Academy Press.

Jobe, J. B. and Mingay, D. J. (1991). Cognition and survey measurement: History and overview. *Applied Cognitive Psychology, 5*, 175-192.

Jobe, J B., Pratt, W. F., Tourangeau, R., Baldwin, A. K. and Rasinski, K. A. (forthcoming). Effects of interview mode on sensitive questions in a fertility survey. In L. Lyberg *et al.* (Eds.), *Survey measurement and process quality*.

Jobe, J. B., Tourangeau, R. and Smith, A. F. (1993). Contributions of survey research to the understanding of memory. *Applied Cognitive Psychology, 7*, 567-584.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*, 213-236.

Lessler, J., Tourangeau, R. and Salter, W., (1989). Questionnaire design in the cognitive research laboratory. *Vital and Health Statistics*, Series 6, No. 1 (DHHS Publication No. PHS 89-1076). Washington, DC: U.S. Government Printing Office.

Mosher, W. D. and Duffer, A. P., Jr. (1994, May). *Experiments in survey data collection: The National Survey of Family Growth Pretest*. Paper presented at the meeting of the Population Association of America, Miami.

Moss, L. and Goldstein, H. (Eds.). (1979). *The recall method in social surveys*. London: NFER Publishing Co.

O'Reilly, J. M., Hubbard, M. L., Lessler, J. T., Biemer, P. P. and Turner, C. F. (forthcoming). Audio and video computer-assisted self interviewing: Preliminary test of new technology for data collection. *Journal of Official Statistics*.

Orne, M. T. (1969). Demand characteristics and the concept of quasi-controls (pp. 143-179). In R. Rosenthal and R. L. Rosnow (Eds.), *Artifact in behavioral research*. New York: Academic Press.

Schwarz, N. (1993, November). *What respondents learn from questionnaires: The survey interview and the logic of conversation*. Third Annual Morris Hansen lecture presented to the Washington Statistical Society, Washington, DC.

Smith, A. F. (1991). Cognitive processes in long-term dietary recall. *Vital and Health Statistics*, Series 6, No. 4 (DHHS Publication No. PHS 92-1079). Washington, DC: U.S. Government Printing Office.

Willis, G. B., Royston, P. N. and Bercini, D. (1991). The use of verbal report methods in the applied cognitive laboratory. *Applied Cognitive Psychology, 5*, 251-267.

DISCUSSION

Roger Tourangeau
National Opinion Research Center

1. Introduction

Both of the papers in this session concern the impact of the mode of data collection--in particular, the effects of self-administration--on data quality. The papers share a more specific concern with the difficulties respondents may have in reading survey questions and following skip patterns and other instructions for completing the questions. The two papers explore very different approaches to addressing these problems. The paper by Jenkins and Dillman describes some procedures for making it easier to read questionnaires; the paper by Lessler and O'Reilly discusses a method that eliminates the need for the respondent to read the questions entirely.

Although a good deal has been written about the impact of mode of interviewing on the results obtained (see, for example, Bradburn et al., 1991; deLeeuw and van der Zouwen, 1988; Groves and Kahn, 1979; Hochstim, 1967), there is no general model of the effects of the different methods of collecting survey data. The various popular methods of collecting survey data--in telephone or face-to-face interviews or in self-administered questionnaires--differ on several important dimensions; further, each of these basic procedures can be carried out on paper or using a computer. I suggest that the different modes of data collection vary on at least three key psychological dimensions:

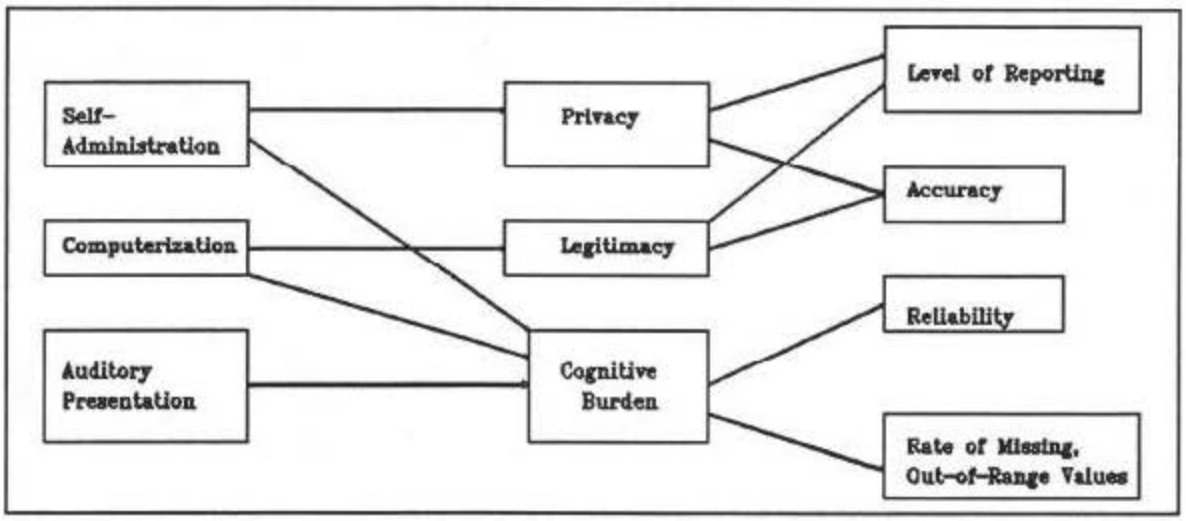
1) **Cognitive demands.** Conventional paper-and-pencil questionnaires require either the respondent or the interviewer to read the questions and to follow the instructions; this requirement may sometimes exceed the reading abilities of respondents or interviewers.

2) **Level of privacy.** By eliminating the need for respondents to tell the interviewer their answers, self-administered questionnaires may reduce respondent concerns about the interviewer's reaction or about other family members overhearing sensitive information.

3) **Perceived importance of the study.** Laptop computers are still a novelty for most of the population, and the use of laptops in face-to-face data collection may enhance the perceived importance or objectivity of the study.

A general model incorporating all three dimensions is depicted in Figure 1. According to the model, features of the method of data collection (such as use of self-administered questions) affect the three psychological dimensions (e.g., level of privacy), which in turn affect data quality (willingness to report accurately about sensitive behaviors). Much of what is currently known about the different modes of data collection is captured in the model.

Figure 1. Path Model of Mode Effects



2. Paper by Lessler and O'Reilly

The paper by Lessler and O'Reilly reports results from a study that compared computer-assisted personal interviews (CAPI) with interviews in which the computer administered the questions directly to the respondent via earphones (audio computer-assisted self-interviewing, or audio-CASI). This study was done as a pretest for Cycle V of the National Survey of Family Growth (NSFG), and the pretest comparison naturally focusses on the two modes of data collection most likely to be used in that survey. Unfortunately, the pretest was not designed to separate out the effects of the several key advantages that audio-CASI offers relative to other modes of data collection (computerization, auditory presentation, self-administration). By comparing audio-CASI and CAPI, the pretest mainly examines the impact of self-administration rather than the other variables distinguished in Figure 1.

A fair number of studies have already shown that self-administration increases the level of reporting of sensitive

behaviors. For example, Turner, Lessler, and DeVore (1992) demonstrated increases in reported drug use with self-administered questions, a finding replicated by Schober and colleagues (Schober et al., 1992). London and Williams (1990) find that more abortions are reported on a self-administered questionnaire than in a face-to-face interview (see also Mott, 1985). A study recently conducted by NCHS and NORC demonstrates increased reporting of sexual behaviors in self-administered questionnaires (see Table 1). That study compared face-to-face interviews and self-administered questionnaires in both a computer-assisted and conventional paper-and-pencil formats; as is apparent in Table 1, the effects of self-administration were larger and more consistent that those of computerization.

Table 1. Mode Effects in the Women's Health Study

Experimental Group	**Mean Reported Sexual Partners**		
	Past Year	Past 5 Years	Lifetime
Self-Administered Questions	1.72	3.88	6.54
Conventional (SAQ)	1.56	3.37	6.88
Computer-assisted (CASI)	1.89	4.40	6.25
Interviewer-Administered Q's	1.44	2.82	5.43
Conventional (PAPI)	1.56	2.86	4.58
Computer-assisted (CAPI)	1.36	2.79	6.27

Note: Each mean based on approximately 240 interviews; total rows for self- and interviewer-administration are based on approximately 500 completed interviews.

The results of the NSFG pretest on the differences in abortion reporting by mode are not very dramatic--audio-CASI increased the proportion of the sample cases who reported an abortion to 27.1% as compared to 23.8% when those same cases were interviewed via CAPI. Moreover, the audio-CASI abortions questions were different from those in the CAPI questionnaire, and they were administered after the respondents had already completed the CAPI interview. So the results from the study are perhaps better characterized as suggestive than definitive. In any case, they are certain to be useful to those charged with making practical decisions about the NSFG design.

Despite any weaknesses in the evidence regarding the advantages of audio-CASI, I predict that this technology will quickly be widely adopted by survey organizations. Having the capability, we will embrace the objective. The advantages of the new technology are, in some sense, too clear cut to require a lot of experimental confirmation--computerization virtually eliminates skip errors, self-administration minimizes privacy concerns, and auditory presentation eliminates the need for respondents to be literate. All in all, audio-CASI is a package that should prove irresistible.

3. Paper by Jenkins and Dillman

The paper by Jenkins and Dillman proposes 20 principles for improving the readability of self-administered questionnaires. The sensible tone of these recommendations reminded me of the admonitions in Tufte's *The Visual Display of Quantitative Information* and of the advice offered to writers in Strunk and White's *Elements of Style*. I did, however, wish that the authors had followed one additional principle; here is my proposed addition to the list:

Principle 21. Follow the Lord's example; never present more than ten commandments at any one time.

I am not advocating that Jenkins and Dillman abandon any of their principles! But I do think that they might try to formulate some larger principles from which their more specific guidelines follow. As I read their paper, it seemed to me that their recommendations reflected four key underlying axioms. First, the flow of a questionnaire should follow the natural reading order of the respondents. In English, this means questions should flow from left to right and from top to bottom. Second, questionnaires should use familiar, readily-understood graphical conventions. For instance, the same design element should always cue the same respondent action. Third, the questionnaire should call attention to the key information (via boldfacing and other methods). Finally, there should be a clear path for respondents to follow. Graphical features should emphasize this path. Table 2 groups 19 of the 20 principles discussed by Jenkins and Dillman under these four general themes.

I found almost all of their recommendations quite compelling. The one major exception involved matrix items--for example, questions that are asked for each family member or for each event of a given type. Jenkins and Dillman argue against giving respondents the choice on how to proceed through the matrix, and this may be the best way to ensure that they answer every question.

Table 2. Four Underlying Axioms

1) Where possible, take into account the natural reading sequence (left-to-right, top-to-bottom).

   --Include key information in the question, not after it (5)

   --Align questions and answers vertically (6)

   --Make top headings more prominent than those in the middle (10)

2) Use easily understood graphical conventions.

   --Use familiar formats (1)

   --Use same design feature to request the same action (13)

   --Avoid variability (14)

   --Use different layouts to distinguish different types of questions (16)

3) Call attention to the **key information** (15).

   --Present only the most relevant information (2)

   --Feature questions rather than explanations (4)

   --Put instructions where they are needed (7)

   --Put captions, units for answers where they will be seen (17)

4) Establish a clear path through the questionnaire (12).

   --Avoid multi-task questions (8)

   --Avoid matrix questions (9)

   --Use graphical instructions (such as arrows) to make the path salient (11)

   --Use graphical features to emphasize the path (18)

   --Avoid separating questions with lines and boxes (19)

   --Provide structures that make sense leave the R no choice! (20)

Note: Numbers correspond to those used by Jenkins and Dillman.

The counterargument is that there are sometimes good reasons for letting respondents follow their natural chain of associations in recalling specific incidents. For some respondents, it may be easiest to recall events that involve one person before recalling those involving the next person; for others, however, an event involving one family member may trigger the recall of similar events involving a different family member. As a result, there could be advantages to letting respondents follow whatever order the flow of memories seems to impose. Only further work can determine whether the advantages of imposing an order on the questions outweigh those of letting the respondents select the order they find most congenial.

The work that Jenkins and Dillman are doing is, in my view, quite important. Mail questionnaires are likely to remain a major method for collecting survey data for the foreseeable future. As this paper demonstrates, our current practice in developing these questionnaires often falls far short of the ideal. At NORC, instructions on self-administered questionnaires are sometimes put in boxes to distinguish them from the questions. During cognitive pretests, I have found that respondents often use the box as a cue for identifying material they need not bother to read. So, I agree with Jenkins and Dillman in thinking that bad graphical design can lead serious errors.

4. A Final Point

The method developed by Lessler and O'Reilly and the principles articulated by Jenkins and Dillman share an underlying goal--that of improving data quality, primarily by reducing missing data. One of the main advantages of audio-CASI over other methods of self-administration is that the software automatically computes which item the respondent is to answer next; this eliminates data that are missing due to incorrectly skipped items. Similarly, many of Jenkins and Dillman's principles stress methods to make sure that respondents answer all the applicable questions by making it easier for the respondents to figure which questions they are supposed to answer. Although audio-CASI has other noteworthy features and the principles proposed by Jenkins and Dillman will help address other response problems (such as questions that are misunderstood rather than missed entirely), a major objective of both approaches is the reduction of missing data. Valuable though this endeavor is, I cannot help but wonder whether this is the most pressing data quality issue that we face. The development of computer-assisted data collection methods has greatly increased the capital requirements of survey organizations; I sometimes wonder whether the gains in terms of data quality have provided a return commensurate with the investment. Perhaps it would make more sense to worry about whether respondents answer the questions accurately than to worry so much about whether they answer at all.

REFERENCES

Bradburn, N. M., Frankel, M., Hunt, E., Ingels, J., Schoua-Glusberg, A., Wojcik, M., and Pergamit, M. (1991), A comparison of computer-assisted personal interviews with personal interviews in the National Longitudinal Survey of Labor Market Behavior—Youth Cohort. In *Proceedings of the U.S. Bureau of the Census Annual Research Conference*, Washington, D.C.: U.S. Bureau of the Census.

deLeeuw, E.D. and van der Zouwen, J. (1988), Data quality in telephone and face to face surveys: A comparative meta-analysis. In R. Groves et al. (Eds.), *Telephone Survey Methodology*, New York: Wiley.

Groves, R.M. and Kahn, R.L. (1979), *Surveys by Telephone*, New York: Academic Press.

Hochstim, J. (1967), A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association, 62*, 976-989.

London, K., & Williams, L. (1990, May), A comparison of abortion underreporting in an in-person interview and self-administered questionnaire. Paper presented at the Annual Meeting of the Population Association of America, Toronto, Canada.

Mott, F. (1985), *Evaluation of fertility data and preliminary analytic results from the 1983 survey of the National Longitudinal Surveys of Work Experience of Youth*. A report to the National Institute of Child Health and Development by the Center for Human Resources Research, January, 1985.

Schober, S., Caces, M., Pergamit, M., & Branden, L. (1992), Effects of mode of administration on reporting of drug use in the National Longitudinal Survey. In C. Turner, C., J. Lessler, & J. Gfroerer (Eds.), *Survey Measurement of Drug Use: Methodological Studies* (pp. 267-276). Rockville, Md.: National Institute on Drug Abuse.

Tufte, E. (1983), *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Turner, C., Lessler, J., and Devore, J. (1992). Effects of mode of administration and wording on reporting of drug use. In C. Turner, J. Lessler, & J. Gfroerer, (Eds.), *Survey Measurement of Drug Use: Methodological Studies* (Pp. 177-220). Rockville, Md.: National Institute on Drug Abuse.

Session 10

STATISTICAL USES OF ADMINISTRATIVE RECORDS

Improving Data Quality Through Increased Data Sharing

Edward A. Trott[1]
Bureau of Economic Analysis

In this time of major cutbacks throughout the federal government the Bureau of Economic Analysis (BEA), along with other statistical agencies, is being asked to improve the quality of statistical information and its availability. This is a big challenge. It will require a lot of thinking "outside the lines," followed by action.

To emphasize this point, I offer the following quote;

> "In a rapidly changing world, the best solution is not to keep redesigning the organizational chart; it is to melt the rigid boundaries between organizations. The federal government should organize work according to customers' needs and anticipated outcomes, not bureaucratic turf. It should learn from America's best-run companies, in which employees no longer work in separate, isolated divisions, but in project- or product- oriented teams.
>
> To do so, the government must make three changes. It must give federal workers greater decision making authority, allowing them to operate effectively in cross-cutting ventures. It must strip federal laws of prohibitions against such cooperation. And it must order agencies to reconsider their own regulations and tradition-bound thinking."[2]

One of the recommendations of the National Performance Review is the elimination of legislative barriers to the exchange of business data among federal statistical agencies. This recommendation is referenced as DOC11 and cites the Census Bureau, Bureau of Labor Statistics, and the Bureau of Economic Analysis as participating agencies. It further recommends the reduction of reporting burden on American business.

Today I will mention several areas in BEA's international, national, and regional accounts that utilize data from the Internal Revenue Service (IRS), Bureau of the Census (BOC), and Bureau of Labor Statistics (BLS)--where we are today--and then take a look at where we could be if more barriers to information exchange are removed.

---

[1]The author is the Assistant to the Chief, Regional Economic Analysis Division, Bureau of Economic Analysis, United States Department of Commerce. He is grateful for the contributions of Ralph Kozlow, International Investment Division; Brooks Robinson and Kenneth Petrick, National Income and Wealth Division; and Wallace Bailey, Regional Economic Measurement Division to the various parts of the paper.

[2]*"Creating a Government That Works Better and Costs Less; The Report of the National Performance Review."* Penguin Books, USA, 1993.

I.Current Uses

International Accounts

BEA and the Census are authorized to exchange certain confidential statistical data with one another for the purposes of augmenting and improving the quality of BEA's Title 22 (direct investment and international services) data and Census' Title 13 (economic census and related programs) data, under the Foreign Direct Investment and International Financial Data Improvements Act of 1990. Under the same act, the Bureau of Labor Statistics is authorized, for the purposes of augmenting and improving its data, to have access to selected information collected by BEA. Only data collected directly by BEA can be shared with Census and BLS; thus, for example, data BEA obtains from Census cannot, in turn, be shared with BLS. Similarly, only data collected directly by Census can be shared with BEA. Thus, BEA cannot have access to administrative records data obtained by Census from the IRS.

BEA and Census have made extensive use of their limited authorization to exchange data with one another and, as a result, have significantly augmented and improved the data collected in their business statistics programs. Several data exchanges have been for the purpose of obtaining detailed, establishment-level data on foreign direct investment in the United States. In 1992, BEA information on the identity of enterprises, or companies, that were foreign owned was linked to the Census' establishment-level data on the Standard Statistical Establishment List covering all U.S. establishments, to obtain detailed BOC data (State by 4-digit SIC) for 1987 on the number, employment, payroll, and shipments or sales of foreign-owned and all U.S. establishments. In 1993, BEA and Census released detailed data (mostly at the 3-digit SIC level by State) for 1989 and 1990, covering nearly all the data items on Census' Annual Survey of Manufactures (ASM) for foreign-owned and all U.S. manufacturing establishments. In 1994, BEA and Census will release detailed 1988 and 1991 ASM data for foreign-owned and all U.S. establishments. These exchanges have enabled statistical analyses that offer important new insights into the operations of foreign-owned companies and their impact on particular sectors of the U.S. economy.[3]

Other data exchanges between BEA and BOC have been for the purpose of supplementing the mailing lists for BEA's mandatory international surveys. BEA obtained confidential Census data on the identity of companies that are foreign owned, of companies that have foreign affiliates, and of companies that export services, in order to supplement mailing lists for BEA surveys of inward and outward direct investment and of international services transactions, respectively. The use of Census data to supplement the mailing lists for BEA surveys of international services transactions were particularly useful, because a significant number of potential new respondents were added

---

[3] See, for example, "Foreign Direct Investment in the United States: Establishment Data for 1987," in the October 1992 issue of the *Survey of Current Business*, and "Characteristics of Foreign-Owned U.S. Manufacturing Establishments," in the January 1994 issue of the *Survey*.

to BEA's mailing list as a result. (To further this purpose, questions on whether a given company or establishment sold, or exported, any services to foreign persons were included on a number of additional 1992 economic census surveys at BEA's request.)

In addition, BEA and the BOC are currently studying the feasibility of linking BEA's data on foreign-owned enterprises to Census' establishment-level product and product class data from the economic censuses, and product-by-country merchandise export data from its Exporter Data Base. The feasibility of other data link projects between the two Bureaus is also being discussed.

As mentioned above, in a parallel data link project, BEA's data on foreign direct investment in the United States have been linked to BLS data on all U.S. businesses. The initial results of that link, released in 1992 by BLS[4], provided data for 1989 and 1990 on the number, employment, and payroll of foreign-owned establishments. In October 1993, BLS released information on the occupational structure of foreign-owned manufacturing establishments for 1989[5].

BEA and the Department of the Treasury are also implementing limited data exchanges. BEA collects its data on direct investment and international services transactions, and Treasury collects its data on portfolio investment, under the same act (P.L. 94-473, as amended) and, in many cases, from the same U.S. companies. In order to ensure that companies are reporting correctly to both agencies, the Secretaries of Commerce and Treasury obtained OMB authorization to exchange and compare certain data. OMB approval for the data exchanges was obtained in January 1993, but BEA and Treasury are still working out the details necessary to implement the data exchanges. Note that any additional data exchanges between BEA and Treasury will require this same slow implementation process (beginning with letters from the Secretaries of Commerce and Treasury requesting OMB authorization, followed by negotiation between the agencies on the details for sharing the data.) It is hoped the time and resources required to implement data exchanges would be greatly reduced or eliminated if broad data sharing legislation were enacted.

National Accounts

BEA currently is accessing both noncorporate and corporate Statistics of Income (SOI) data via letters of agreement between IRS and BEA and the Secretary of Commerce and the IRS Commissioner as permitted by statutes. The

---

[4]United States Department of Labor. Bureau of Labor Statistics. "Employment and Wages in Foreign-Owned Businesses in the United States, 1989." Press Release, USDL 92-473 (July, 1992). and (*ibid*) "Employment and Wages in Foreign-Owned Businesses in the United States, Fourth Quarter 1990." Press Release, USDL 92-663 (October 1992).

[5]United States Department of Labor. Bureau of Labor Statistics. "New Research on Occupations in Foreign-Owned Manufacturing Establishments in the United States." Press Release, USDL 93-455 (October 1993).

Bureau's relationship with IRS has been in place for many years--dating back to 1944. At that time the Secretary of the Treasury, Henry Morganthau Jr., signed an executive order granting Commerce staff access to both noncorporate and corporate returns for statistical purposes. This executive order was superseded by statute in the early 1980's and access to SOI data from then until now has been limited in scope.

Following is a list of some of the specific topics that were recently researched via corporate return access--unless otherwise noted, adjustments to data resulting from this analysis improved the current year NIPA estimates of profits, interest, and dividends by adjusting extrapolating series to better conform to SOI definitions:

> Identified discrepancies between SOI tabulations and trade association data in profits reported by life insurance companies for 1986-1987. Once identified, the SOI tabulations could be adjusted for changes in the tax law.
>
> Identified differing reporting practices for commercial bank bad debt losses in SOI and FDIC tabulations and adjusted the current year FDIC series.
>
> Reconciled pension and profit-sharing reporting practices between SOI and IRS Form 5500 preparatory to revising estimation methodology and data sources.
>
> Separated business and interest receipts of credit agencies to more accurately measure interest flows.
>
> Identified public utility joint partnership formations to ensure unduplicated reporting of new plants.
>
> Analyzed captive finance companies' consolidation practices by parent corporations to ensure unduplicated reporting of income items.
>
> Made adjustments to shift data for individual corporations to different industry classifications, which are more consistent with NIPA definitions and other data sources.

A secondary use of the individual company reports has been in the analysis of peculiar movements in SOI data. Because BEA has direct access to the IRS' corporate tax return information (not for the Foreign Direct Investment project), company by company panel comparisons have been compiled to review several questions regarding reporting practices, consolidation methods, or compilation errors.

Currently, actual use of noncorporate tax information has been limited to a study to determine the reporting patterns of partnership "pass-through" income reported on schedule K and by the individual partners on their Form 1040's. Results of the study led to a revision of the NIPA estimates of noncorporate partnership income and a redesign of the Form 1065 to explicitly determine the amounts of Schedule K income distributed to partners.

An additional area of data sharing between BEA and another federal agency--the Census Bureau--involves the use of the Quarterly Financial Reports (QFR) to adjust and analyze the quarterly NIPA estimates. Through a memorandum of agreement with the Census, BEA's sworn Census employees have continued to access individual company reports since 1982. Prior to 1982 BEA had access to the reports through an agreement with the Federal Trade Commission.

Primary use of this access is for the quarterly NIPA estimates. Each quarter BEA staff provide a list to Census of large non-recurring income or expense items which have been noted in a company's individual QFR. These items--mainly capital related and not associated with current production--are not treated as current income or expense items in the NIPA's. Census reviews the company report to determine if the questioned item was reported as a component of the indicator used to estimate pretax profits, or if it was excluded via several other reporting options. If the amount was included, Census identifies the exact pretax impact, which often differs from the amount in published after-tax profit reports, and the industry where the company is classified in their tabulations. The QFR also is used to augments the BEA company list with additional companies Census discovers during the review process of the quarterly reports.

In 1989 BEA entered into another cooperative arrangement with the Bureau of the Census; to explore the feasibility of using BOC data to prepare construction price indexes. In 1989, BEA staff--after being sworn in as Census agents--and the Census staff of the Construction Statistics Division (CSD) began to work on producing improved construction price indexes. The CSD provided micro-level construction data that were either collected through the Survey of Construction or the Value of New Construction Put in Place Survey. BEA staff used the data to test prospects for producing hedonic price indexes for residential and nonresidential buildings.

BEA staff produced a multifamily structures hedonic price index in 1991 for 1978-1989, and BEA began using it to prepare constant dollar estimates of the multifamily structures component of GDP at the time of the 1991 comprehensive revision of the NIPA's. Prior to this time, BEA used the BOC's Single-family Hedonic Price Index for Houses Under Construction. An article that describes the development of the index was published.[6]

Regional Accounts

For the preparation of BEA's estimates of personal income by State and county, the Regional Economic Measurement Division (REMD) receives several statistics tabulated by county from IRS Form 1040. The data are coded geographically and tabulated by the Census Bureau; the IRS, SOI Division provides these unpublished tabulations to BEA. The sample used by SOI for the Form 1040 statistics that they publish is too small to yield data for regional use. The

---

[6]See U.S. Department of Commerce, Bureau of Economic Analysis. "A Price Index for New Multifamily Housing." *Survey of Current Business*. Washington, DC: U.S. Government Printing Office, February 1993.

tabulations of dividends, taxable interest, and gross rents are the principal bases for REMD's estimates of the monetary components of property income. The tabulations of wages and salaries are vital for adjusting the components of personal income that are initially estimated on a place-of-work basis--wages and salaries, other labor income, and most of personal contributions for social insurance--to a place-of-residence basis. The tabulations of number of returns, number of exemptions, and adjusted gross income are used for REMD's quality control of the tabulations that are used directly in the estimates.

Proprietors' income and employment are important parts of the State and county estimates of personal income and employment prepared by REMD. Annual filings to the IRS by sole proprietors and partnerships are the only direct source of data for regional nonfarm proprietors' income and employment. The sample-based statistics from these filings that are prepared and published by the SOI Division are used for the calculation of national nonfarm proprietors' income. For the State and county estimates, REMD has acquired, under a contractual arrangement, individual records for sole proprietors from IRS Form 1040 Schedule C and for partnerships from IRS Form 1065. These records were then geographically and industrially coded, tabulated and combined to form the basis for REMD's State and county nonfarm proprietors' estimates. REMD-SOI data contracts have been in place for tax years 1981-83 and 1987-91.

II.Future Uses: How BEA Programs Would Be Improved Upon Implementation of DOC11

For illustrative purposes, several examples of how BEA's economic accounts can be improved with increased data sharing are given below. These are only a few of the examples that could be cited.

International Accounts

First, increased data sharing would improve the data obtained in the data link projects on foreign direct investment in the United States that BEA conducts with the BOC and BLS. As discussed earlier, BEA has shared its information on the identity of foreign-owned companies with both Census and BLS, to obtain those agencies' detailed data, by industry and State, for companies that are foreign owned. However, at present, Census and BLS cannot share their data with one another either directly or through BEA. For the data items released that were defined most similarly by both Census and BLS (number of operating establishments and employment), there are very large unexplained differences between the two agencies' data for foreign-owned establishments. For example, within manufacturing, the differences between Census and BLS data for 1990 are often at least 20 percent at the 2-digit SIC classification level; at the 3-digit and 4-digit SIC levels, the differences are often even larger. (Other industries may show even larger differences than manufacturing.) With such large unexplained differences, the results of the data link projects, and the conclusions based upon them, may be questioned. Reasons for differences may include different definitions, differences in industry coding, errors in the data, or that one agency was better able than the other agency to link its establishments to a given foreign-owned company. Because Census and BLS cannot share these data with one another currently, the exact reasons for the differences cannot now be determined. With increased data sharing, the

reasons for the differences would be learned, and appropriate corrections to the data sets, if necessary, could be made.

Second, BEA's ability to analyze the data from the BEA-Census data link project would be improved if BEA had access to IRS data on Census' file of foreign-owned establishments. BEA employees are allowed to obtain limited administrative records data directly from the IRS, but they are not allowed to access the same data at the Census Bureau, because redisclosure of administrative record data to BEA is prohibited. (For 1987, IRS data account for about 15% of the estimate of employment for all foreign-owned establishments combined.) The Commerce Department must analyze and report to Congress on the employment, market share, value added, productivity, profitability, etc. of foreign-owned business enterprises compared to all U.S. business enterprises (P.L. 101-533 section 3(c)(1)). In order that BEA may efficiently perform these analyses, they need access to the full foreign direct investment (FDI) data file at Census. (The IRS has said it will support a regulatory change that will allow BEA access to the full FDI file, but this change may take considerable time to implement and will only facilitate this one project.)

Third, full data sharing would improve BEA's direct investment data if it permitted BEA access to complete IRS data on foreign-owned U.S. businesses and on U.S. businesses that have foreign affiliates. Both BEA and the IRS have released data on the net income, assets, sales, and number of foreign-owned U.S. companies and of U.S. companies that have foreign affiliates, and the data of the two agencies frequently differ substantially. Some of these differences might be eliminated if BEA were able to obtain information from the IRS for comparing with its own information on which companies are foreign owned and which companies own foreign affiliates. Also, appropriate corrections to BEA's data sets could be made, and BEA's sample frames could be improved.

National and Regional Accounts

First, for companies that own more than one establishment, Census and BLS both annually "map" all the individual establishments on their registers to the owning enterprises. Census obtains data needed for this purpose from the mandatory annual Report of Organization survey; BLS obtains the needed data from the Multiple Worksite Report (executed by each of the States and funded by BLS) and from commercial sources, such as Dun and Bradstreet. Under data sharing, only one agency would need to gather the information and share it with the other. This change would result in one set of company/establishment relationships that BEA could use to allocate data available only for companies to an establishment industry level.

Second, REMD currently makes intensive use of tabulations of ES-202 wage and employment data in estimating the wage-related components of personal income and employment for States and counties. Any improvement in the geographic or industrial classification of the ES-202 data that might result from ES-202/Standard Statistical Establishment List integration would increase the accuracy of BEA's regional estimates.

Third, full data sharing between Census and BLS would improve the derivation of the sampling frame for BLS' Producer Price Index. BLS uses the ES-202 file for its sample frame. The price indexes would be improved if the probability of inclusion in the sample were based on its shipments or sales, because it would allow for more precise weighting of the data that are collected. Producer Price Indices are used extensively in preparing constant dollar estimates for several expenditure components of GDP and for the industry distribution of GDP.

Fourth, with full data sharing, the industry coding used for calculating both the BLS producer price indexes and the Census shipments data could be fully consistent. At present, BEA must apply BLS' price indexes, which are based upon BLS' industry codes, to Census' shipments data, which are based on Census' industry codes, in calculating real GDP by industry.

Fifth, increased data sharing will improve BEA's I-O tables and estimates of gross State product and GDP by industry. In developing its detailed industry and State estimates, BEA utilizes Census data on establishment receipts and value added, and BLS data on employment and payroll. As discussed in connection with BEA's estimates of real GDP, the accuracy and usefulness of the I-O tables and of BEA's estimates of gross State product and GDP by industry would be considerably improved through consistent industry coding of establishments. These estimates would also be improved with full BEA access to unsuppressed Census Bureau estimates of value added, cost of materials, etc.; BEA now must indirectly estimate these items for suppressed Census Bureau data cells.

Lastly, REMD has identified additional statistics from the Individual Income Tax Returns (see under "Current uses") that would strengthen its current estimates and would help to extend its current estimates to prepare for the adoption of measures more consistent with the System of National Accounts (SNA).

The transfer of these data series to BEA do not require additional legislative action, only that the data be requested from the IRS. A relatively new and growing income entry on the individual tax form--interest from municipal bonds--will yield data to replace proxy information currently used to estimate the nontaxable interest component of personal income at both the State and county level. The tabulation of pensions and annuities received would be essential for the estimation of the corresponding components of the SNA aggregates.

III Conclusions

From the current uses section of this paper, it is apparent that even the limited amount of data sharing permitted under existing legislation has already produced a remarkable increase in the amount and quality of data and analyses, with no increase in respondent burden and minimal increased costs to the statistical agencies. The examples of future uses suggest improvements in data sharing that will benefit both the customers and staff of BEA as the recommendations of DOC11 are met.

It is apparent that significant duplication and inefficiency exist in business statistics data collection programs currently. It is particularly important that we continue our efforts to promote fuller data sharing in light of the dual challenges of limited agency resources and having to monitor developments in a rapidly changing world.

The opportunity to improve the accuracy of the income and product accounts produced by BEA is here. I believe that with the implementation of DOC11, estimates prepared not only at BEA but throughout the government will be improved as to their accuracy, timeliness, and will be less costly to produce.

Health Reform Information Systems: Great Expectations, Uncertain Prospects

Edward L. Hunter
National Center for Health Statistics

Introduction

In the opening keynote address, Graham Kalton described Roger Herriot (who organized this session and died suddenly this spring) as someone that had a vision beyond what we did every day, a person that encouraged us to look up from our work occasionally and check for shifting paradigms. Roger recognized that the types of information systems proposed for health reform may well represent such a paradigm shift. The purpose of this paper is to convey a sense of the great expectations that have accompanied discussion of health reform, the importance of these proposals for research and statistics, and to discuss some of the uncertainty over whether these expectations will be realized.

Health reform

The national debate over health care reform is reaching a crescendo this summer in Washington. More than a dozen congressional committees are weighing decisions that might - or might not - lead to wholesale changes in the U.S. health care system and, as a result, in our Nation's health information systems. The work of these committees is focused on piecing together legislation that can pass both Houses before members leave the Capitol to face voters in November.

While each health reform proposal may get there by a different route, most share the same common goals: improve health insurance security for the 38 million persons without health insurance, and for those that have insurance but risk losing it due to illness or unemployment; improve access to health care; control and even reduce health care costs, and even improve the health of the American people.

Among the questions yet to be resolved:

1) Will health reform result in universal coverage for health insurance, or will more incremental approaches be adopted?

2) If there are standard health insurance packages, how generous will the benefits be?

3) How will coverage for the uninsured be financed? Options include a mandate that employers provide coverage, a mandate that individuals buy insurance, or broad-based tax increases.

4) To what extent will the Government be involved in regulating activity in the health care system?

Major proposals still under consideration include modifications to the President's proposed Health Security Act and more modest or market-oriented alternatives proposed by moderate Democrats and Republicans. At this point in the debate, it is not yet clear what type of reform - if any - will be enacted.

Role of information in the current health reform debate

It does seem clear, however, that issues of health reform will continue to dominate our agenda for the coming years regardless of what happens this session of the Congress. And it is clear that information - or the lack thereof - has become central to the emerging debate over health reform. Questions, such as who lacks health insurance and why; how and why do businesses cover employees; how is insurance related to use of services, and to health; what drives health costs up; and what will expanded benefits cost, have become central to crafting a legislative package. Yet, these are questions for which our current information systems - both survey and administrative - can provide only partial answers.

Importance of information in a reformed health system

Nearly all advocates of health reform visualize a new health care system in which information plays an increased role. Since reform will almost certainly be implemented at the State level, State governments will need better information to tailor their approaches. Health plans will need better data to select member institutions and facilities, and manage the quality and cost of patient care. Consumers will need far greater information on quality of services, satisfaction, and outcomes in order to make informed choices between competing plans. Further down the road, States will need to assure that health plans are meeting standards - and the Federal Government will need to assess the expected and unexpected impacts of reform. Finally, there will almost certainly be a debate over what went right - or wrong - that can only be conclusive if there is adequate information.

There is a growing, and encouraging, consensus among the sponsors of competing health reform proposals regarding several key information provisions. Most important among these are the need for administrative simplification - digging health providers and payers out from under the ever-growing pile of forms, and the need for greater uniformity in the way health transactions are recorded and reported. Nearly all parties to the debate recognize that we need to hold providers and health plans more accountable, that we need better information on health outcomes to guide practice, and that consumers will play a greater role as information users.

The Health Security Act: a national framework for health information

President Clinton's proposed "Health Security Act"[1] includes the most ambitious - and controversial - information proposals of any plan currently being considered. The balance of this paper will outline these proposals, and address some of the implications if current legislative deliberations result in enactment of provisions resembling these proposals.

The architects of the Health Security Act regard the availability of health information as essential to the success of managed competition, the basis for the Act. Toward that end, the Act calls for a national framework under which standardized, nationwide information would be collected for all patients, providers, and encounters with the health care system. This system would be built on a consensus around core, minimum data sets that can be used by multiple parties; standardization and simplification of currently burdensome systems of administrative and payment records; unique identification of individuals and providers to facilitate linkages; national legislation to protect the privacy of records in the system; and regional data centers to process the records and provide for access for research and statistical purposes.

Information on individuals enrolled in health insurance plans

There are three major elements in this design: enrollment, encounter, and administrative data. In the first, information would be obtained on each individual eligible for health insurance when they are enrolled in a health plan. At a minimum, this information would include identifying information (including name and address); a unique identification number; and additional demographic information such as age, sex, and (hopefully) race/ethnicity. Information on the individual's source of coverage would be included, as would any other information deemed necessary to adjust premium payments to health plans based on the individual's risk. It is possible that this list could eventually expand to include information on the individual's health status and medical history.

Data on encounters with the health care system

In the second element of the framework, a standard, minimum data set will be collected at the point of each encounter with the health care system. This data would begin with that now provided on claims forms, such as reason for visit, diagnoses, procedures, prescriptions, followup, and disposition. But since care will increasingly be delivered by prepaid systems that do not require individual claims, information would also be obtained for all such encounters. Demographic data will already have been obtained for individuals through the enrollment system, so it would not be necessary to collect such information at each encounter.

---

[1] HR 3600, "The Health Security Act of 1994", Title V.

Debate over implementing this encounter data system will focus on the length and content of the minimum data set, the extent to which an encounter might be used to obtain risk factor or other information not immediately relevant to the clinical encounter, and the types of coding and classification systems that will be used to translate recorded information into usable statistics. To make this information of maximum use for research purposes, encounter data will need to go beyond what we currently collect through the payment system, and be of better quality.

Data on plans and providers

Third, in order to facilitate the operation of the health care system, including payment of claims, certification of plans, etc., administrative data on the characteristics and operations of health plans and providers will be included in the system. This may range from the multiple affiliations that plans and providers will have with each other, to the types of services provided and patients served.

Privacy and data access

To summarize, the information framework envisioned in the Health Security Act provides for demographic information on all individuals eligible for health insurance, information on the characteristics of all health providers, and a small set of information on each encounter an individual has with the health care system.

Such a system has enormous potential - the potential to meet a wide range of research needs, and the potential - if badly designed or implemented - to jeopardize privacy protection that Americans have come to expect.

As a result, privacy protection is being given a great deal of attention by those currently debating the Health Security Act. The Administration is committed to the enactment of strong confidentiality protection,[2] and there are serious efforts being made in the Congress to craft comprehensive privacy legislation for medical records.[3]

The privacy debate over the health reform information network is critical to the eventual ability of researchers to use data that may be part of the network, for a variety of reasons. First, without addressing the public's concerns through enactment of strong protection, the Congress may not be willing to create such a network in the first place. Second, if the

---

[2] Testimony of Nan D. Hunter, Deputy General Counsel, DHHS, before the Subcommittee on Information, Justice, Transportation, and Agriculture, Committee on Government Operations, U.S. House of Representatives, April 20, 1994

[3] HR 4077, "Fair Health Information Practices Act of 1994," introduced by Representative Gary Condit (D-CA), is one example of pending legislation.

public doesn't have confidence in the privacy of their medical records, it will be impossible to assure the quality and accuracy of the information they provide to the health care system. Third, in an effort to reassure the public on privacy, it is possible that the Congress will over-react and place restrictions on access to data that will unnecessarily restrict research and statistical uses. And fourth, even with the best of intentions, it is a tough job to craft legislation that balances privacy with access to data, and we must make efforts to ensure that there are no inadvertent limits are placed on our ability to use these sources.

Statistical uses of *ADMINISTRATIVE* records?

The Federal statistical community argues that information produced through any system, including purely administrative data systems, should serve multiple purposes, and be put into a form that is conducive to research and compilation of statistics. It is an encouraging development that the Administration and the Congress are giving serious consideration to ways in which the health reform information framework can be put to maximum public use.

In fact, it is not at all clear that this system should be considered an *ADMINISTRATIVE* system - since research and statistics are among the most important uses the system is being designed to meet.

Federal statistical agencies have a strong history of making data available to the public, while protecting privacy, by producing anonymous, person level public use files. It is encouraging that this history is guiding policy development for access to data in the health reform network. As envisioned by the authors of the Health Security Act, regional data centers will be responsible for creating general-purpose files, as well as responding to individual researchers with special requirements, while being given special privacy authorities to assure the protection of individual records.

Potential uses of data from the health reform network

Public health provides a number of useful examples of how this network might serve research needs. The network will be broad and comprehensive, but will lack detailed diagnostic and clinical information typically included in patient charts. In effect, it will be a "mile wide and an inch deep." But this "mile" of data can fulfill many important public health needs, including ones we have great difficulty meeting today.

It can provide a broad picture of the population, and reasonably complete coverage of the delivery of health services. The existence of a population base for these records also provides greater opportunity to link medical events to denominators, and develop better incidence/prevalence data on certain diseases. It allows us to address subgroups of the population with greater confidence.

Examples of health applications

One specific example is the tracking of progress in improving the immunization of preschool children, a national public health priority on which we will spend hundreds of millions of dollars in the coming years. From enrollment files, we will know the ages of children that should be receiving vaccines - and will be able to use encounter data to determine immunization rates by health plans and local areas. Likewise, we will be able to monitor our success in promoting other preventive services, such as encouraging women of certain age groups to have a mammogram, and men of appropriate ages to have cholesterol screening. With this information, we will be able to better target intervention programs to prevent disease.

If a major health reform is implemented, we will also experience significant change in nearly every aspect of health care and, potentially, the health of the public. The data in this network will help us to monitor health status and outcomes, produce routine measures of quality of care, and assess changes in the organization, financing, and delivery of health services.

Potential uses in other areas

Since the network will include information on a broad population basis through the enrollment records, it may prove to be a valuable tool for demographic research. These files will, at a minimum, include a variety of demographic and geographic items, may provide us a picture of the family status of individuals, and may even record changes in individual circumstances over time.

Since enrollment for health insurance will be primarily employment-based, the network will include considerable information about the employment status and occupation of covered individuals and their family members. This may prove to have applications for labor market and employment-related research.

And, the network will include detailed information on health care providers, including facilities, professionals, and other organizations. This will provide the basis for research on the establishments and workforce of a large segment of the economy.

Potential uses for Census

The broad population coverage anticipated for the enrollment file has obvious potential uses in conducting censuses of populations. Congressional testimony by the Administration reflects current thinking that <u>some</u> access to identifiable information in the network by statistical

agencies may be appropriate.[4] As the Bureau of the Census looks toward the use of other administrative sources of data for population censuses, the health insurance enrollment file may prove to be among the most complete source of demographic information of the population. This is said, of course, with full recognition of the limitations of this system, which will be discussed briefly below.

It is possible that some provisions allowing for access to identifiable records for such purposes may be built into the health security act as it is considered by the Congress. However, it is also likely that these sorts of uses will be carefully specified to ensure that the public is not left with the perception that there will be widespread sharing of identifiable records for non-health related purposes. It is also important to remember that the Act was written with privacy protection in mind, and limits use of identifiable files to health applications except in limited circumstances.

Applications for survey research

To this point, this review has not focused on the limitations of the proposed health reform information system, but clearly, there are many. For example, it is likely that even a universal entitlement to insurance coverage will not entice all eligible individuals to participate - leaving enrollment files short of complete population coverage. We know that certain population groups are likely to be underrepresented, or excluded. We know from other administrative data sets that we have reason to suspect the quality of reporting of certain types of data.

Finally, since the framework relies principally on the reporting of encounters with the health care system, it will lack information on events, conditions, and other health issues of interest that occur outside the health care delivery system.

For the purposes of this paper, however, as we speculate on what this system might do for research and statistics, it is assumed that we can devise methods to compensate or adjust for these limitations. For example, it will be critical to augment the network with surveys and other data collection mechanisms to fill critical data gaps and provide quality checks on administrative reporting.

The information framework, then, will both rely on data collected in surveys, and can at the same time facilitate the conduct of surveys.

---

[4] Testimony of Nan D. Hunter, Deputy General Counsel, DHHS, before the Subcommittee on Census, Statistics, and Postal Personnel, Committee on Post Office and Civil Service, U.S. House of Representatives, March 16, 1994

Surveys required for health reform

At the National Center for Health Statistics, and elsewhere in the Federal statistical establishment, we also face an enormous task in developing surveys to provide information for implementing and evaluating health reform, and we have reason to hope that new systems will facilitate this work. These surveys will be essential to augment information available through the network, since no single data source will be able to provide the depth of data needed for the types of analyses we will need to be performing. We are already beginning to address some of these needs with development of new surveys, such as the National Employer Health Insurance Survey.

Sampling frames

The first, and most significant benefit of the new system will be in the construction of sampling frames. Many current surveys rely on sampling approaches that are time consuming and inefficient. The three components of the network will facilitate sampling as follows:

1) **Population sampling from the enrollment data base**, allowing for selection of sample individuals by demographic characteristics such as age, race, sex, geographic location, occupation, and possibly other proxies. Sampling and screening for rare and non-clustered population groups, currently difficult and expensive, would be greatly facilitated. Similarly, it would be possible to quickly identify and survey selected population groups, allowing for quicker turnaround for topics of current policy or research interest.

2) **Provider sampling from administrative records.** Many current health provider surveys (e.g., hospitals, physicians, nursing homes) construct sampling frames from lists provided by professional associations, phone directories, licensing agencies, or private marketing firms. These lists are subject to considerable error and possible bias, and are costly and time consuming to create. Sampling from complete and accurate lists of providers, with detailed characteristics already included in administrative files, would make provider surveys faster, less costly, and of greater quality.

3) **Sampling based on conditions, diagnoses, or procedures** that can be identified through encounter records. For example, cohorts of individuals that received specific medical treatments can be identified for followup or interviews to assess outcomes. Similarly, persons with certain diagnoses or conditions could be identified to conduct studies of access to appropriate medical care; persons with only limited use of the health care system could be identified and included in studies of barriers to access to care.

In each of these examples, it is clear that extensive research will be required to assess the extent to which sampling frames created entirely from administrative records systems will adequately represent the universe, and approaches will need to be developed to augment these frames. It is also clear that there will continue to be an important role for traditional sampling approaches where administrative frames are not adequate.

Other potential uses - record linking

Access to enrollment and encounter data in the network can also facilitate efficient linkage of detailed health status, risk, and behavior information that we will obtain through population-based surveys to the utilization and outcomes data included in encounter files.

These files can facilitate longitudinal studies, in which individuals could be enrolled in cohorts according to characteristics in the enrollment files (e.g., occupation), or based on encounters. Followup of theses cohorts could then be conducted through surveys and through analysis of encounter data.

Conclusion: prospects

To conclude, the information system described in the health security act, and touched upon in other health reform proposals, holds enormous potential for those of us interested in research and statistics. Data from enrollment, encounter, and provider data systems will augment or replace existing approaches, and provide survey researchers with new tools for sampling, conducting longitudinal followup studies, and linking survey data with outcomes. Along with a variety of new survey approaches, this information system provides us with the opportunity to fill many long-standing gaps in our understanding of the health care system and public health, and there are a variety of potential applications beyond health.

Getting there from here requires that at least four difficult questions be answered in the affirmative: 1) Can Congress actually enact a broad reform that will change the health care system? 2) In enacting any information requirements, can an appropriate balance between privacy and access to data be found that will allow the system to be useful? 3) Can all the affected interests come together to agree on standards and minimum data sets that include items of use for statistical purposes? and finally, 4) Can we develop and manage the technology for handling and protecting the volume of records that will result, and turn these records into usable statistics?

Improvements to Economic and Health Statistics: Discussion

Miron L. Straf[1]
Committee on National Statistics
National Academy of Sciences--National Research Council

It is a privilege and pleasure to discuss the two papers by Edward A. Trott, Jr., and Edward L. Hunter. Both papers are excellent and exceptionally well prepared. As such they tributes to the memory of Roger Herriot, to which we have dedicated this session. On behalf of all of us at this conference and the many others who knew or were influenced by Roger, I thank our speakers for their quality presentations.

Discussion of "Improving Data Quality Through Increased Data Sharing: The National Performance Review (NPR) Initiative," by Edward A. Trott, Jr.

Ted Trott is being modest. The data sought are even more important than one might glean from his paper. As our economy becomes an increasingly a global one, our current system of economic statistics becomes woefully out of date. Let me give you one example from the report of one of our Committee on National Statistics panels, *Behind the Numbers: U.S. Trade in the World Economy* (National Academy Press, 1992). The traditional balance of payments framework classifies transactions by geographical boundaries, so that when foreign affiliates of U.S. firms sell to other foreign firms, no effect is recorded in our trade balance (See Figure 1-1). By supplementing reports on exports and imports across geographic boundaries with information on other international business activities, the panel estimated that the difference in what the U.S. bought from and sold to foreigners in 1987 was $64 billion, less than half the reported $148 billion trade deficit for that year.

Trott shows how sharing data can improve efficiency and accuracy, but the benefits are more. In most of the cases that he mentions, more data are or would be created than the sum of what we have in separate agencies.

Sharing data also promotes new research and serves to test new theories and methods. It helps us improve our models and better understand sources of error. It furthers the use of empirical studies and other analyses in public policy formulation and evaluation. And it respects the respondents who provided the data by seeing that their information benefits society in the most effective way.

---

[1]The remarks herein are those of the author and not necessarily those of the Committee on National Statistics nor of the National Academy of Sciences--National Research Council.

Trott's paper is an excellent documentation of costly problems from not sharing and the benefits from sharing data. It also provides some innovative approaches to the problems of missing data. One example raises a question asked in the session with George Duncan and Nancy Kirkendall on confidentiality and disclosure limitation. The Census Bureau can't provide detailed micro data to the Bureau of Economic Analysis (BEA) and so suppresses some cells in the tables provided. BEA indirectly estimates these missing data. Suppose it does so through a model based on some assumptions. The Census Bureau has the original data and can determine whether the model holds or the assumptions are wrong. Should it inform the BEA?

My friends who are into Transcendental Meditation provide some important advice for missing data. Once, when I was delivering a lecture on multiple imputation methods for nonresponse, one of them intonated to the audience that the methods were unnecessary, because, if you search really hard, you can find the missing data within yourselves.

And, to me, that is Trott's message to the federal statistical system. Only, as he points out, you really don't have to search very hard.

Thus, we are led to ask if what we need is really going to be achieved by a data sharing agreement in fulfillment of recommendation 11 to the Department of Commerce from the National Performance Review. Perhaps, but not likely in our lifetime.

What is needed is a new framework to provide the incentives for finding the ways for data sharing to occur while fulfilling the intent of confidentiality law. the Committee on National Statistics explored one approach recently with a number of statistical agency heads. The basic issue is that there needs to be a quid pro quo for one party to share information with another. Truly cooperative agreements are rarely built on one-way streets. But what can be offered in return for data provided? One approach is to return analyses that depend on the data. But more can be offered.

A statistical agency can engage in designing, developing, and even managing a data collection program of another agency, so that it can obtain important data for statistical purposes. What is more the data program of the other agency could be for administrative purposes. Thus, BEA or the Census Bureau could design for the Treasury a tax data base for tax policy or even for enforcement purposes that could also provide statistical data. It's not far fetched. As David Binder has reminded us here, Statistics Canada already does it.

Discussion of "Health Reform Information Systems: Great Expectations, Uncertain Prospects," by Edward L. Hunter.

Hunter's paper is a consummate exposition of one of the greatest challenges to the federal statistical system: the information system that may arise in a reformed health care

system. Ed lays out, in as much detail as has been developed, the data to be collected and the potential uses of them. I want to complement his description with another perspective that raises some issues for the federal statistical system.

The structure of the data is in three parts: enrollment, encounter, and claims. In addition, data would be compiled on health care providers. Consumer information would be issued as some sort of report card on plans and providers. Here measures of the quality of health care services become important.

The regional data centers may be entirely new public-private ventures. They may be the places where public-use files are compiled and where linkages to other data, such as surveys, are performed. Here privacy and confidentiality are crucial.

The health care information system envisioned can provide many benefits. The information is needed for physicians to make accurate diagnoses and recommend appropriate treatment. But the information is also needed to monitor trends that affect the costs of health care, to plan for the changes in the needs for health care, and to achieve a better understanding of how behavioral and social factors are related to health and health care coverage.

It is important to note however, that, although health care information alone may provide information on health outcomes, it is not possible to understand what may have caused the outcomes unless data can be combined with other data, such as from surveys. The information serves other purposes too, including being a valuable source of social, economic, and demographic characteristics of our population. However important these purposes are, however, the health care information system would be driven by the needs to implement the health care system and not by research and statistical needs.

So that the public might benefit from access to the information for research and statistical purposes, the Committee on National Statistics expressed concerns to the Congress about legislative provisions that were proposed. As Hunter points out, some of these provisions could preclude data for important research and statistical purposes.

The Committee's first and foremost concern is that privacy and confidentiality of health care information be adequately protected. It is not necessary to sacrifice either confidentiality or the benefits of information: both are possible if legislation provides for responsible access and demonstrated, effective means to protect confidentiality. The Committee also argued that health care legislation can protect confidentiality of information and yet permit important research and statistical uses of that information by

- Prohibiting data about an individual that are collected or maintained for research and other statistical uses from being used in any administrative or enforcement action affecting that individual. This principle is referred to as *functional separation*.

- Extending confidentiality protection to identifiable data about individuals, wherever the data are maintained.

- Providing sanctions against unauthorized disclosures by any user.

- Authorizing access to health care data about individuals for research and statistical purposes whenever confidentiality can be assured.

- Creating an independent federal advisory body charged with fostering a climate of enhanced protection for all federal data about persons and responsible data dissemination for research and statistical purposes.

The Committee conveyed these points in letters to members of Congress working on health care legislation. The result is that bills were modified to permit access to health care information for public health research and for research on behavioral and social factors affecting health. Without these changes, the viability of several of our major national surveys would be threatened.

Many difficult problems remain, however. Who, for example, rules on access? One proposal is that Institutional Review Boards grant permission for access. Some parties, such as the Institute of Medicine, would not even permit access with consent. The concern is that, if access were allowed with consent, then employers might require it as a condition of employment.

Hunter talks about the system enhancing the legally mandated reporting to local and state health departments. Not everyone wants to facilitate providing this information, however, because of confidentiality concerns.

Hunter also shows the many benefits of the data in and of themselves. We must take care however, that people do not get the impression that a single, large anonymous file can serve most needs. Such a public-use tape might do so for public health purposes, but many policy purposes require different data to be combined or data to be combined in other ways. And public use tapes cannot be combined with further data that may be needed, such as information from Social Security earnings records.

The examples Hunter gives of uses for demographic research, for labor market and employment-related research, and for providing information on health care providers shows the utility of the data to other agencies: the Census Bureau, the Bureau of Labor Statistics, and the National Center for Health Statistics. What are these agencies doing to assure access to the data? I am not aware of any major initiative.

Many serious confidentiality problems remain. For example, is it appropriate to screen the system to develop a sampling frame of people with a disease like AIDS? Are

we going to allow them to be contacted? How? Should we permit contact by a call or letter that might go to a member of their household?

A specific confidentiality practice that can be harmful is when an agency specifies that information will not be used for purposes other than for which it was collected . That is a bad principle, and some bills before Congress have adopted it. Such a blanket prohibition denies many legitimate research and statistical uses. No one can foresee all potential uses of data that would benefit society. If such prohibitions were enacted and enforced, society would lack important information that it could obtain only at greater cost through new data collection that might further intrude on individual privacy.

Despite Hunter's expectations, the administration may have missed the boat on privacy and confidentiality provisions. Bills are moving now quickly through the Congress, and Representative Condit's bill can stand on its own with or without a health reform bill.

Nevertheless, the challenges for implementing such an ambitious data system are before us, and the statistical agencies can offer a special expertise here. We can look toward a health care information system to allow through research and statistics the means of providing the information required by stakeholders represented by the six P's: policy makers, public health officials, payers, providers, patients, and the public.

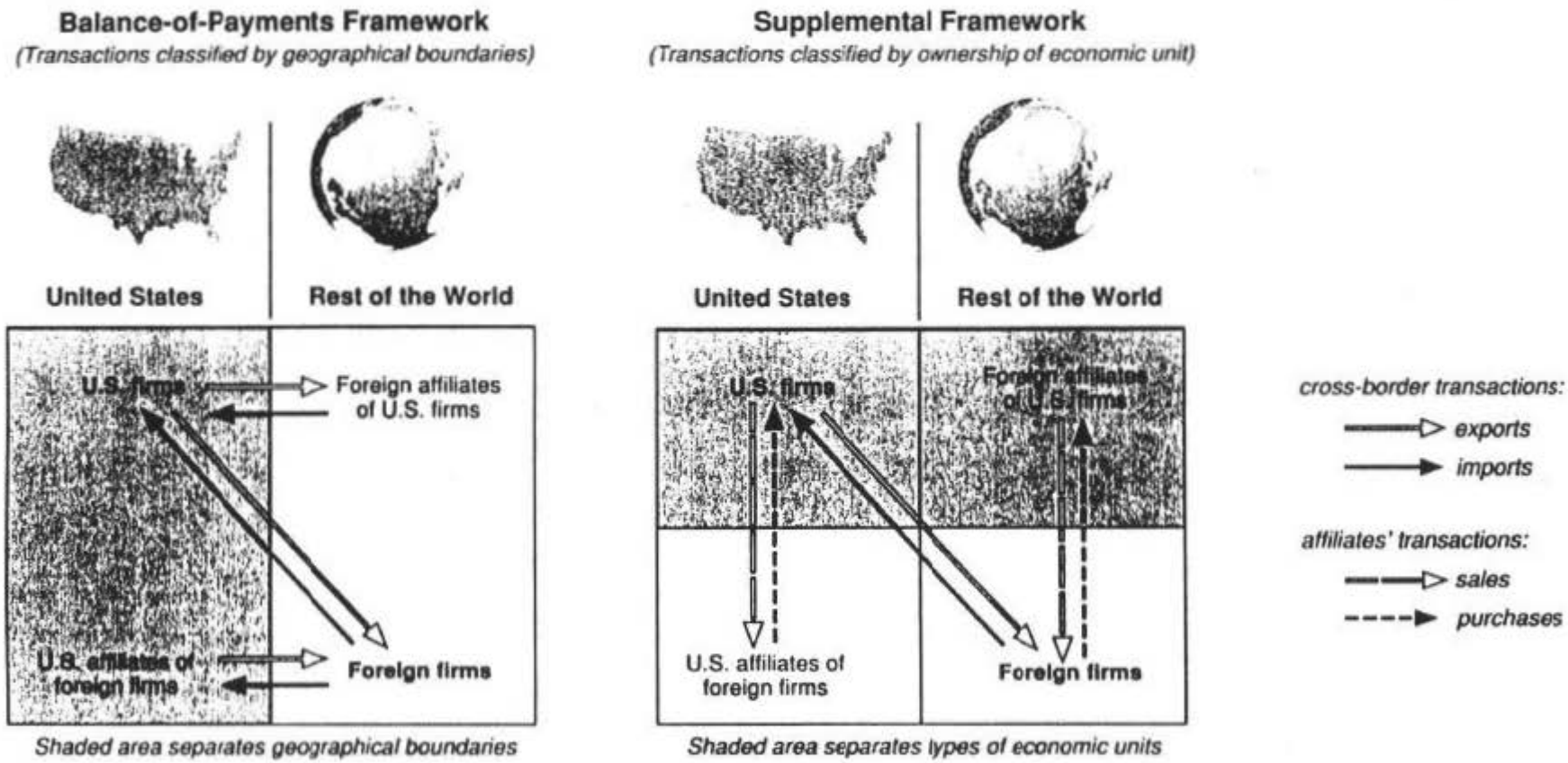


FIGURE 1-1 The balance-of-payments framework and a proposed supplemental framework.

Source: Kester, Anne Y. (1992). *Behind the Numbers: U.S. Trade in the World Economy*, Report of the Panel on Foreign Trade Statistics, Committee on National Statistics, National Academy of Sciences--National Research Council. Washington, D.C.: National Academy Press.

Session 11

SMALL AREA ESTIMATION

Small Area Estimation for the National Health Interview Survey Using Hierarchical Models

DONALD MALEC
National Center for Health Statistics; Hyattsville, MD

J. SEDRANSK
State University of New York; Albany, NY

1. Introduction

There is a continuing need to assess health status, practices and resources at both the national level and subnational levels. Estimates of these health items help determine the demand for quality health care and the access individuals have to it. Although NCHS survey data systems can provide much of this information at the national level, little can be provided directly at the subnational level, except for a few large states and metropolitan areas. The need for State and substate health statistics exists, however, because health and health care characteristics are known to vary geographically. Also, health care planning often takes place at the state and county level.

Using a hierarchical model, our focus is on the development of state estimators using data from the National Health Interview Survey (NHIS). Information on health status, practices and resources is collected annually in the NHIS and direct national estimates of these items are also produced annually. The NHIS is a multistage, personal interview sample survey. The current sample design uses 1,983 primary sampling units (PSU's), each PSU consisting of a single county or a group of contiguous counties (minor civil divisions are used instead of counties in New England and Hawaii). The population of 1,983 PSU's is stratified and approximately 200 are sampled with probability roughly proportional to their population sizes. Within each sampled PSU clusters of households are formed and sampled. Areas within a PSU with a high concentration of blacks are oversampled. The NHIS is a cross-sectional survey: each year, a new sample containing approximately 50,000 households and 120,000 individuals is selected. (For additional details about the design of the NHIS see Massey et al. 1989.) Although the total sample size is large, the sample size in most states is too small to produce direct estimates that are sufficiently precise.

Malec and Sedransk (1985) have described Bayesian methodology appropriate for the analysis of some multi-stage sample surveys when the variables are normally distributed. We have extended this methodology to accommodate binary random variables, the predominant variables in the NHIS. Our model is similar to that of Wong and Mason (1985). However, the objective in Wong and Mason (1985) is inference about parameters in the model rather than finite population quantities. While Dempster and Tomberlin (1980) investigate small area estimation

methods for binary random variables they, like Wong and Mason, provide an empirical Bayes rather than a fully Bayes solution. Since empirical Bayes procedures often account for only a fraction of the error correctly represented in a fully Bayes approach, we prefer the latter. Recent advances in numerical methods (e.g., the Gibbs sampler) permit the employment of a full Bayesian analysis; see, e.g., Gatsonis, et al. (1993), Malec and Sedransk (1993a), and Malec, Sedransk, and Tompkins (1993).

The notation and model are described in Section 2 while the estimation methodology is presented in Section 3. Section 4 describes the techniques for fitting the proposed models, and displays the final model using data from the 1987 NHIS on utilization of physician care. There is a comparison of alternative estimators in Section 5, and evaluation of the proposed methodology is described in Section 6.

2. Model Specification

The model in (2.1), (2.2) and (2.3) below includes the most important features of the sample design. Our objective is to produce accurate point estimates and appropriate measures of variability by accounting for geographic variability of the response and using available covariate information.

Let Y_{ij} denote a binary random variable for individual j in county i where $i=1,\ldots, L$ and $j=1,\ldots,N_i$. Within county i and conditional on the p_{ij}, the Y_{ij} are assumed to be independent Bernoulli random variables; i.e.,

$$Pr(Y_{ij}=y_{ij}) = p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}}, y_{ij}\in\{0,1\}. \tag{2.1}$$

Given the vector of M covariates corresponding to individual j, $\underline{X}_{ij}^t = (X_{ij1},\ldots,X_{ijM})$, and $\underline{\beta}_i$, it is assumed that

$$\ln\{p_{ij}/(1-p_{ij})\} = \underline{X}_{ij}^t\underline{\beta}_i. \tag{2.2}$$

To allow for the possibility of a linear regression between each element of $\underline{\beta}_i$ and a set of covariates, $\underline{Z}_i^t=(Z_{i1},\ldots, Z_{ic})$, available at the county level, assume

$$\underline{\beta}_i \sim N(G_i\underline{\eta},\Gamma) \tag{2.3}$$

where, conditional on $\underline{\eta}$ and Γ, the $\underline{\beta}_i$'s are independent and $G_i=\mathrm{Diag}(\underline{g}_{i1}^t, \underline{g}_{i2}^t, \ldots, \underline{g}_{iM}^t)$ and $\underline{g}_{ij}^t$ is a row vector of dimension c_j containing a subset of covariates from $\underline{Z}_i^t$. Additionally, $\eta^t=(\eta_{11},\ldots,\eta_{1c_1},\eta_{21},\ldots,\eta_{2c_2},\ldots,\eta_{M1},\ldots,\eta_{Mc_M})$, conforming to the dimensions of G_i, and Γ is an MxM positive definite matrix. Finally, reference prior distributions are assigned to $\underline{\eta}$ and Γ; i.e.,

$$p(\underline{\eta},\Gamma) \propto \text{constant}. \tag{2.4}$$

Taking $\Gamma=0$ provides a specification that is consistent with the basic assumption of synthetic estimation. In the following "synthetic estimation" refers to the use of (2.1), (2.2) with $\underline{\beta}_i=G_i\underline{\eta}$, and $p(\underline{\eta}) \propto$ constant.

We include as variables in (2.2) those individual level characteristics that provide the best prediction for p_{ij}, and are reliably estimated at the county level in non-census years. Candidate variables at the county level (i.e., G_i in (2.3)) include the variables used to define the NHIS strata (defined at the county level). By predicting for demographic groups within counties and then weighting by postcensal population estimates, estimates are automatically weighted to fixed population totals.

3. Estimation Methodology

3.1 Bayesian Predictive Inference

In this paper, our objective is to make inference about finite population means. By first summing the Y_{ij}'s within a county and then within a state, the population mean within a state can be expressed as

$$\theta = \sum_i \sum_{j=1}^{N_i} Y_{ij}/N. \tag{3.1}$$

Formula (3.1) can represent either a mean for the entire state or for a subpopulation. The first sum is over the collection of counties within the state, while N_i is the size of the population or subpopulation in county i. Here, $\Sigma_i N_i=N$.

In (2.2) we use the variables age, sex and race because these are the only variables for which reliable estimates are available at the county level for non-Census years. In this case, (3.1) can be simplified. Suppose that in the population there are K different values of the vector $\underline{X}_{ij}$. Then write $\underline{X}_{ij} = \underline{X}(k)$ for all ij having pattern k (k=1,...K).

From (3.1),

$$\theta=\Sigma_{(i,j)\in s}\frac{y_{ij}}{N}+\sum_i\sum_{k=1}^{K}\left(\frac{N_i(k)-n_i(k)}{N}\right)\overline{Y}_{ik}^{(ns)} \tag{3.2}$$

where $N_i(k)$ and $n_i(k)$ are, respectively, the population and sample sizes in county i with $\underline{X}_{ij} = \underline{X}(k)$ and $\overline{Y}_{ik}^{(ns)}$ is the mean of the nonsampled individuals with demographic characteristic k in county i.

Letting $\underline{y}_s$ denote the vector of sample observations, we emphasize the first two moments of θ, $E(\theta|\underline{y}_s)$ and $Var(\theta|\underline{y}_s)$.
From (3.2),

$$E(\theta|\underline{y}_s) = \Sigma_{(i,j)\in s}\frac{y_{ij}}{N} + \sum_{i}\sum_{k=1}^{K}\left(\frac{N_i(k)-n_i(k)}{N}\right)E(p_{ik}|\underline{y}_s) \tag{3.3}$$

where $p_{ik} = \exp\{\underline{X}^t(k)\underline{\beta}_i\} / [1+\exp\{\underline{X}^t(k)\underline{\beta}_i\}]$

and

$$\begin{aligned} Var(\theta|\underline{y}_s) = &\sum_{i}\sum_{k=1}^{K}\{N_i(k)-n_i(k)\}E\{p_{ik}(1-p_{ik})|\underline{y}_s\}/N^2 \\ &+Var\left(\sum_{i}\sum_{k=1}^{K}\{N_i(k)-n_i(k)\}p_{ik}|\underline{y}_s\right)/N^2. \end{aligned} \tag{3.4}$$

3.2 Numerical Evaluation

Since the posterior moments of θ are nonlinear functions of $\{\underline{\beta}_i: i=1,\ldots,L\}$, and the posterior distribution,

$$f(\{\underline{\beta}_i: i=1,\ldots, L\},\underline{\eta},\Gamma|\underline{y}_s), \tag{3.5}$$

cannot be expressed in a simple form, numerical evaluation is needed. We generate from (3.5) R sets of parameters, $\Omega = \{\Omega^{(r)}: r=1,\ldots,R\}$, where $\Omega^{(r)} = \{\{\underline{\beta}_i^{(r)}: i=1,\ldots, L\},\underline{\eta}^{(r)},\Gamma^{(r)}\}$. Then we evaluate the $p_{ij}^{(r)}$ using (2.2), and obtain estimates of $E(\theta|\underline{y}_s)$ and $Var(\theta|\underline{y}_s)$,

$$\hat{E}_\theta = \Sigma_{(i,j)\in s}y_{ij}/N + R^{-1}\sum_{r=1}^{R}\left[\Sigma_{(i,j)\in s}p_{ij}^{(r)}/N\right] \tag{3.6}$$

and

$$\begin{aligned} \hat{V}_\theta = &R^{-1}\sum_{r=1}^{R}\left[\Sigma_{(i,j)\in s}p_{ij}^{(r)}(1-p_{ij}^{(r)}) + \{\Sigma_{(i,j)\in s}p_{ij}^{(r)}\}^2\right]/N^2 \\ &-\{R^{-1}\sum_{r=1}^{R}\left[\Sigma_{(i,j)\in s}p_{ij}^{(r)}/N\right]\}^2 . \end{aligned} \tag{3.7}$$

This numerical evaluation is accomplished using a Gibbs sampler; see Malec and Sedransk (1993b) for details.

4. Variable Selection

Using data from the 1987 NHIS we select the variables to be included in (2.2) and (2.3) where the binary variable Y has Y=1 if there has been at least one visit to a physician during the past twelve months. We proceed in two steps by first fitting an individual-level model using (2.1) and (2.2), and then considering the county-level model in (2.3).

Our initial objective is to ascertain the general form of (2.2). We do this by ignoring county variation and estimating $\underline{\beta}$ in the "national" model, (4.1). If $\underline{X}_{ij} = \underline{X}(k)$, (2.1) and (2.2) are replaced by

$$Pr(Y_{ij}=y_{ij}) = p_k^{y_{ij}}(1-p_k)^{1-y_{ij}}, y_{ij} \in \{0,1\},$$

and

$$\ln\{p_k/(1-p_k)\} = \underline{X}'(k)\underline{\beta}. \tag{4.1}$$

First, we obtain estimates based on the saturated model where the sample proportion of individuals in class k, $\hat{p}_k$, is used to estimate p_k. Figure 1 shows the effect of age, race and sex on $\ln\{\hat{p}_k/(1-\hat{p}_k)\}$.

The variation in log odds in Figure 1 corresponds to an expected pattern. First, for a given sex and age, the probability of a physician visit is generally larger for whites than for nonwhites. Second, the general patterns are similar for both races for a given sex. For males, the probability of a physician visit decreases steadily until about age 22.5, and then increases steadily. (Recall that we are using five year age groups.) For females, physician visits decrease steadily until age 12.5 and then increase to about age 27.5. Physician visits remain roughly constant from 27.5 until 62.5 and then increase steadily.

Due to the complex form seen from Figure 1, various spline models, linear in age, were used. A fixed knot spline can be defined as a linear model (Smith 1979) and, hence, used in (2.2). We include the possibility of a knot at each five year age group. The general model investigated included all possible splines that are linear in age, a race effect, a sex effect, a race by sex interaction and, finally, all interactions between these categorical variables and the linear age splines. The set of possible variables is

1) Categorical variables: Intercept, race (r), sex (s) and race by sex (rs)
2) Linear age splines: $X_a(k)=\max(0, \text{age}(k)-a)$, $a=0,5,10,\ldots,85$ and age(k) is the age for individuals in class k.
3) Categorical by age-spline interactions: r by $X_a(k)$, s by $X_a(k)$, rs by $X_a(k)$.

To determine a subset of terms to include in (4.1) the SAS forward stepwise logistic regression procedure, PROC LOGISTIC, was used. This procedure selects variables for inclusion and exclusion using a residual chi-squared test. Since the sample size is approximately 120,000 persons, variables possibly having only a small effect may be included in the model. To determine the total number of variables to use in the model a quantity like R^2 was used. Define the deviance D^2 for the model M_1 as $(\text{Dev}(M_1)-\text{Dev}(M_0))/(\text{Dev}(M_s)-\text{Dev}(M_0))$, where M_0 is the null model (with only an intercept term) and M_s is the saturated model (a parameter is fitted for each age by race by sex group). Note that
$0 \leq D^2 \leq 1$, and equals R^2 for the linear model. The variables, intercept, $sX_{15},\ldots,r$, listed in the table below were included in the model. Adding other variables does not increase the value of D^2 appreciably (note the small contributions of the next best variables, rX_0 and sX_{45}, to D^2).

Variable	Intercept	sX_{15}	X_0	X_{15}	sX_{25}	X_{35}	X_{25}	r	rX_0	sX_{45}
Cumulative D^2	0.00	17.17	22.09	58.88	75.02	87.41	91.55	94.41	95.37	95.70

To check the fit of the model, partial residuals were plotted. Corresponding to each observation there is the residual, $r_{ij} = (y_{ij} - \hat{p}_{ij})/\hat{p}_{ij}\ (1-\hat{p}_{ij})$, which is then averaged over subsets of interest. A typical residual plot has, for a given domain (e.g., sex by race), r_{ij} averaged over all individuals of a given age plotted against age. The particular form of the residual is used because it will estimate a missing term in the logistic model (see Fienberg and Gong's comment to Landwehr, Pregibon and Shoemaker 1984). The residual plot in Figure 2 indicates that the eight variable model provides a good fit to the data. The one large remaining residual (for Black males, aged 85+) corresponds to an estimate based on a very small sample size.

The second step in the data analysis is to identify county-level covariates that affect an individual's probability of visiting a doctor, after having removed the effects due to the individual level covariates. To do this, we combined the individual level and county level models in (2.2) and (2.3) but set $\Gamma=0$. Then

$$\ln\{p_{ij}/(1-p_{ij})\} = X_{ij}^{t} G_i \mathbf{n} \ . \tag{4.2}$$

To reduce the scale of this investigation we consider only the eight individual level variables identified earlier. As indicated by (4.2), we allow main effects of county-level variables and interactions of these county covariates with the individual-level variables. The collection of county covariates that we considered are ones included in the Area Resource File or county mortality file, and thought by subject-matter specialists to be relevant. We have also included county variables related to the formation of the NHIS strata. The procedure we used was to force the eight individual-level variables into the model, and let the SAS stepwise logistic regression procedure add variables. (We have also used graphical methods as described in Malec and Sedransk 1993a and Malec, Sedransk, and Tompkins 1993.) We found no county-level covariates that increased D^2 appreciably. However, there is still considerable county-to-county variation to be captured by (2.3) with $G_i = I$. For other dependent variables (e.g., health status), county-level covariates play a more significant role.

5. Comparison of Alternative Estimators

In this section we use data from the 1987 NHIS to compare the Bayes estimates with the standard alternatives, synthetic and design-based estimates.

For the largest states, the conventional design-based estimates should have relatively small variances, and there should be good agreement between them and estimates based on (3.2). In Figure 3 we plot, for each state and type of estimator (design-based, Bayes, synthetic), the estimated percent of the state population who visited a physician against state sample size. The Bayes estimates (based on a normal approximation to the posterior distribution) are close to the design-based estimates for the largest states, as one would hope. For the same states, the synthetic estimates are always further from the design-based estimates than are the Bayes estimates. As the state sample sizes become smaller the design-based estimates become increasingly unreliable, and the Bayes estimates look less like the design-based estimates, and more like the synthetic estimates. We have also used this same model to produce state estimates of the percent visiting a physician for subpopulations such as persons 65+, non-whites and females. These estimates exhibit the versatility of Bayes estimates; the between county variability, based on Γ, is different for these three cases, leading to different amounts of "gaining of strength". See Malec and Sedransk (1993b) for details and estimates.

Corresponding to Figure 3, Figure 4 is a plot for the 51 areas of posterior standard deviations vs. state sample sizes where we consider both the hierarchical Bayes (formulas (2.1) - (2.4)) and "synthetic"

estimates. For the states with smaller populations, the standard deviations based on the hierarchical Bayes model more properly account for the uncertainty associated with inference about θ.

6. Evaluation

We have investigated whether the conventional sample weights are informative. Figure 5 is a partial residual plot similar to Figure 2. (For this analysis, $r_{ij} = \{y_{ij} - E(p_{ij}|\underline{y}_s)\}[E(p_{ij}|\underline{y}_s)\{1-E(p_{ij}|\underline{y}_s)\}]^{-1}$.) The ordinate of each point is the average residual for all individuals having a sample weight within the range centered at the corresponding abscissa. There is no evidence that the model should include the sample weight as a covariate.

Since the frequency of persons who visit a physician is not available for the entire NHIS population, it is not possible to compare the small area estimates with the true state values. However, by removing a portion of the sample, cross-validation methods can be used to assess how well the model and estimation procedure predict the part of the sample that has been deleted.

The cross-validation procedure that we plan to use is described below. Define the set of sampled elements that are set aside as "A". Let $\underline{y}_A$ denote the vector of observations that correspond to the elements in A and $\underline{y}_{(A)}$ the remaining sampled elements. Also, $\underline{Y}_A$ is the random variable (with observed value $\underline{y}_A$) that represents the removed elements. The predictive distribution, $f(\underline{Y}_A|\underline{y}_{(A)})$, can be used to make comparisons between the observed data, $\underline{y}_A$, and the values of $\underline{Y}_A$ predicted from the model. Specific functions comparing $\underline{Y}_A$ and $\underline{y}_A$, denoted $g(\underline{Y}_A|\underline{y}_A)$, can be defined to evaluate features of the predictions. (See Gelfand, Dey and Chang 1991 for a general review of Bayesian model assessment.)

We shall remove sets of sample elements in ways that permit us to see if our model captures the most important features of the NHIS data. Our evaluation will be based on how well the model predicts the deleted sample,

$$\theta_{AU} = \frac{\sum_{i \in U} \sum_{k=1}^{K} \sum_{j \in A_{ik}} y_{ikj}}{\sum_{i \in U} \sum_{k=1}^{K} n_{Ai}(k)}$$

where the first sum is over all counties in state "U", A_{ik} denotes the set of deleted individuals in demographic group k and county i, and $n_{Ai}(k)$ is the size of A_{ik}. Two choices for the error in prediction are

$$g_{1U}(\underline{Y}_A, \underline{y}_A) = (\theta_{AU} - E(\theta_{AU}|\underline{y}_{(A)}))^2$$

and

$$g_{2U}(\underline{Y}_A, \underline{y}_A) = \left(\frac{\theta_{AU} - E(\theta_{AU}|\underline{y}_{(A)})}{E(\theta_{AU}|\underline{y}_{(A)})} \right)^2.$$

To evaluate how well the model can predict the error of the estimate one may use

$$g_{3U}(X_A, X_A) = \frac{g_{1U}(X_A, X_A) - E(g_{1U}(X_A, X_A | X_{(A)})}{E(g_{1U}(X_A, X_A) | X_{(A)})}.$$

Numerical results from this cross-validation will appear in a forthcoming report.

References

Dempster, A.P., and Tomberlin, T.J. (1980), "The Analysis of Census Undercount From a Postenumeration Survey," Proceedings of the Conference on Census Undercount, Arlington, VA, 88-94.

Gatsonis, C., Normand, S-L., Liu, C. and Morris, C. (1993), "Geographic Variation of Procedure Utilization: A Hierarchical Model Approach," Medical Care, 31, YS54-YS59.

Gelfand, A.E., Dey, D.K. and Chang, H. (1991), "Model Determination Using Predictive Distributions with implementations via Sampling-Based Methods". In Bayesian Statistics 4, eds., J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Clarendon Press.

Landwehr, J.M., Pregibon, D., and Shoemaker, A.C. (1984), "Graphical Methods for Assessing Logistic Regression Models," Journal of the American Statistical Association, 79, 61-83.

Malec, D., and Sedransk, J. (1985), "Bayesian Methodology for Predictive Inference for Finite Population Parameters in Multistage Cluster Sampling," Journal of the American Statistical Association, 80, 897-902.

Malec, D., and Sedransk, J. (1993a), "Bayesian Predictive Inference for Units with Small Sample Sizes: The Case of Binary Random Variables," Medical Care, 31, YS66-YS70.

Malec, D., and Sedransk, J. (1993b), "Small Area Inference for Binary Variables in the National Health Interview Survey," Technical Report.

Malec, D., Sedransk, J., and Tompkins, L. (1993), "Bayesian Predictive Inference for Small Areas for Binary Variables in the National Health Interview Survey," In Case Studies in Bayesian Statistics, eds., C. Gatsonis, J.S. Hodges, R.E. Kass and N.D. Singpurwalla. Springer Verlag.

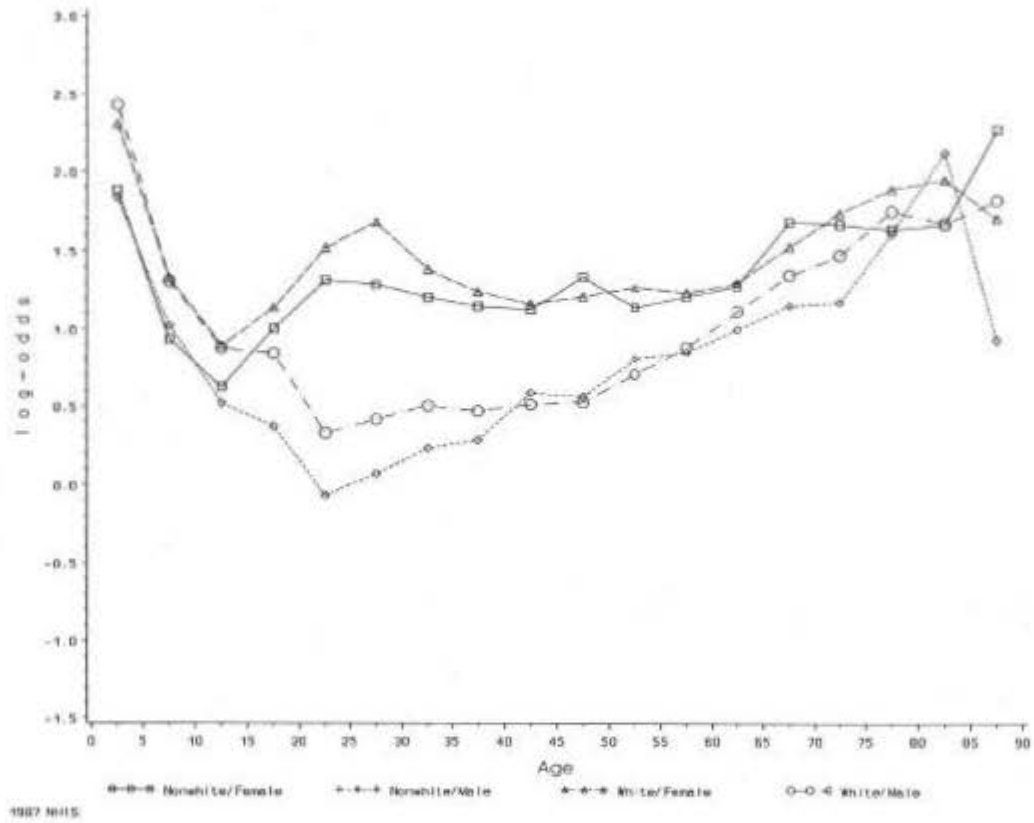
Massey, J.T., Moore, T.F., Parsons V.L., and Tadros, W. (1989), "Design and Estimation for the National Health Interview Survey, 1985-94," National Center for Health Statistics. Vital and Health Statistics, 2:110.

Smith, P.L. (1979), "Splines As a Useful and Convenient Statistical Tool," The American Statistician, 33, 57-62.

U.S. Department of Health and Human Services (1989), "The Area Resource File (ARF) System". ODAM Report No. 7-89.

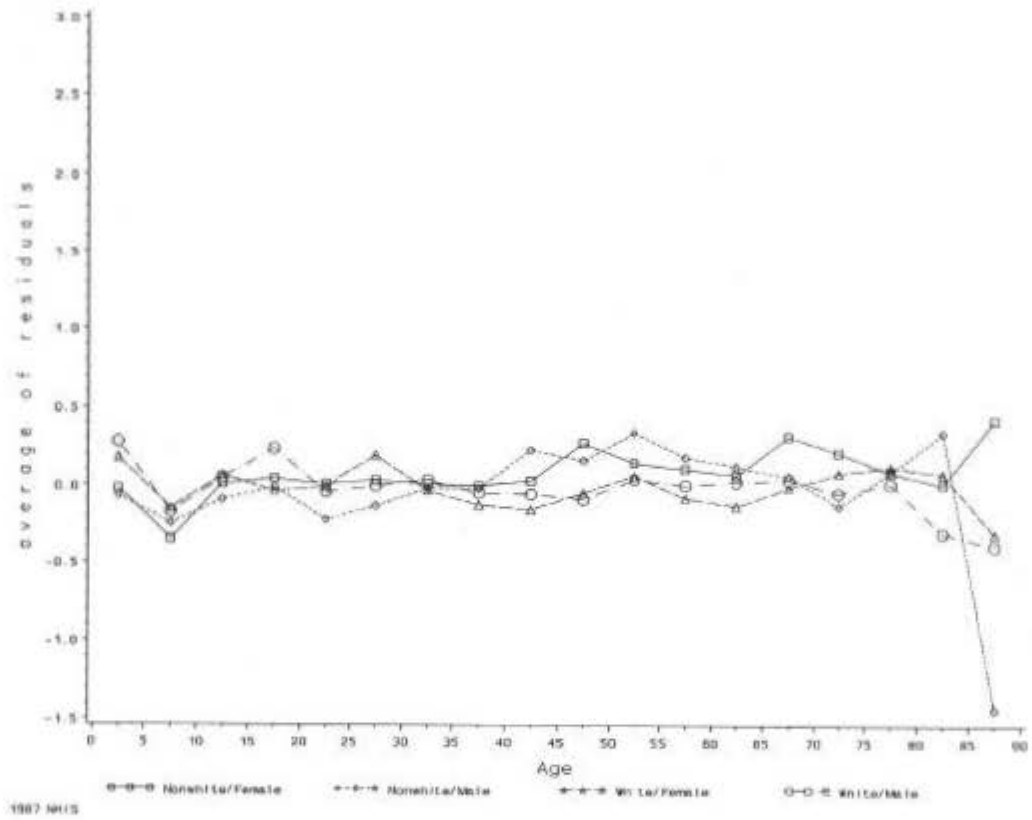
Wong, G.Y., and Mason, W.M. (1985), "The Hierarchical Logistic Regression Model for Multilevel Analysis," Journal of the American Statistical Association, 80, 513-524.

Figure 1: Relationship of log odds of a sample proportion and age for each race by sex group.



NOTE: The binary variable is the presence or absence of at least one visit to a physician within the past year.

Figure 2: Relationship for each sex x race group of average residual and age using the model in section 4.



NOTE: The ordinate of each point is the average residual from the model for all members of a given sex by race by age group.

Figure 3: Estimated percent of population in a state who visited a physician in the past year plotted against state sample size: Hierarchical Bayes, synthetic and design-based estimates.

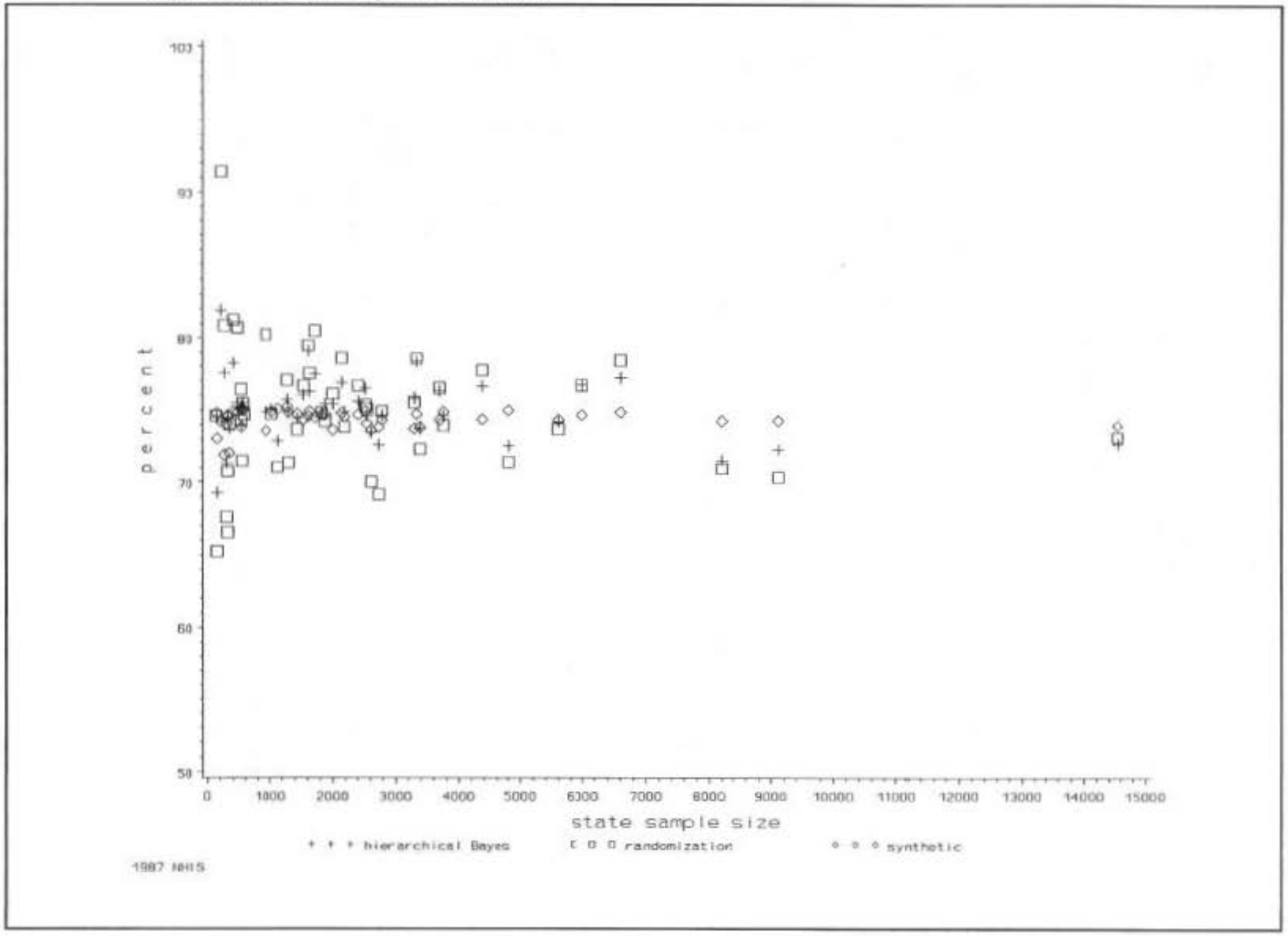


Figure 4: Posterior standard deviations plotted against state sample sizes: Hierarchical Bayes and synthetic models.

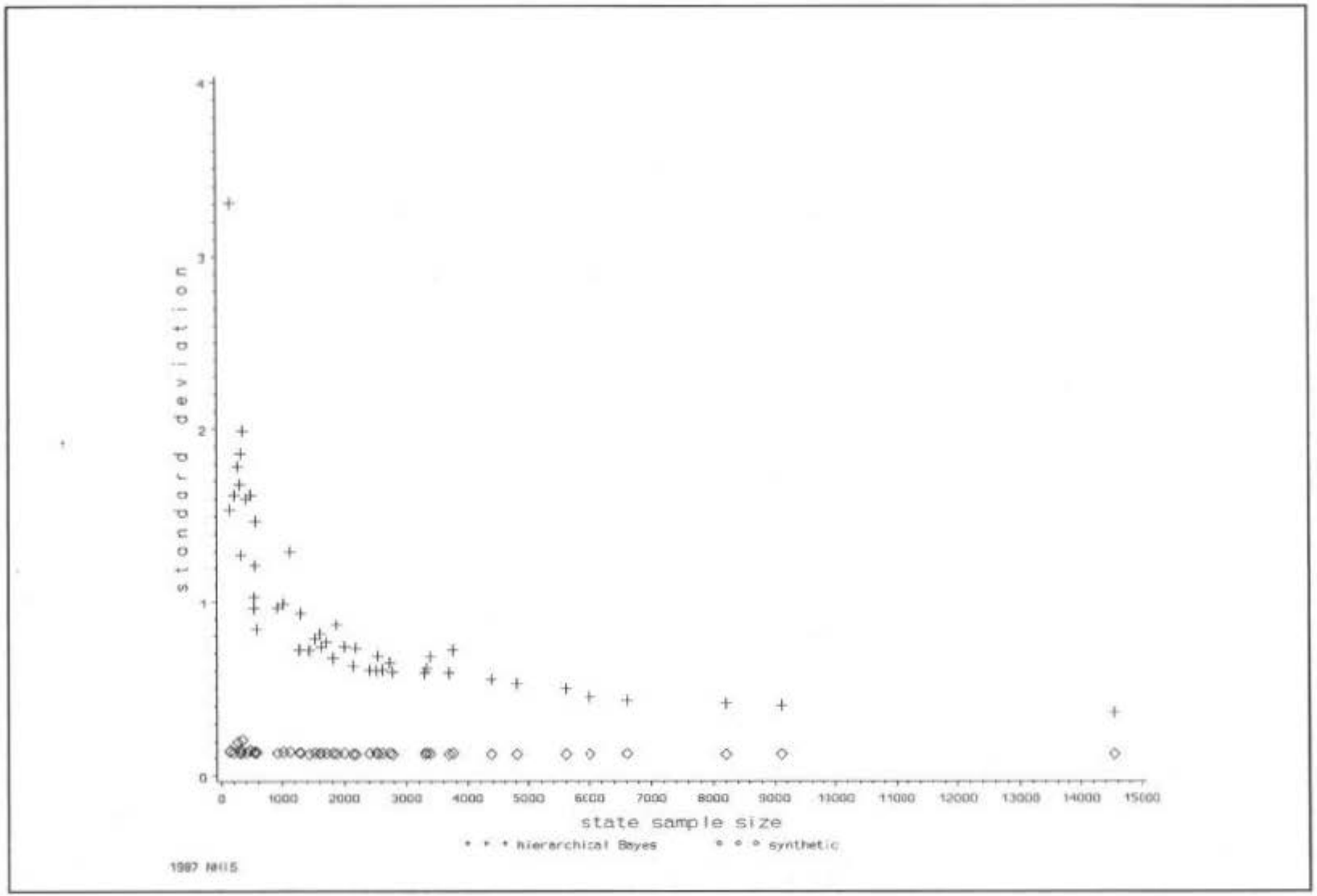
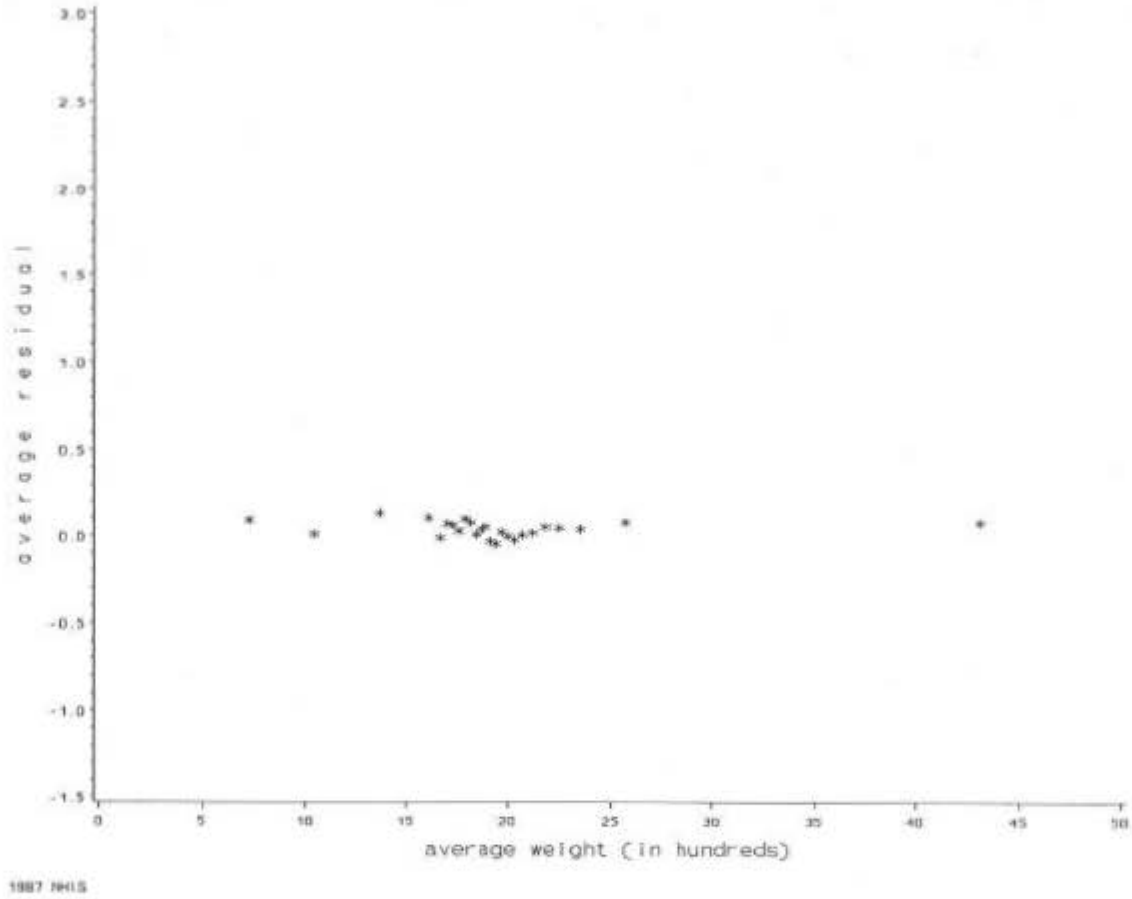


Figure 5: Relationship of average residual and sample weight using the model in section 4.



NOTE: The ordinate of each point is the average residual from the model for all individuals having a sample weight within a narrow range.

THE ROLE OF DESIGN BASED VARIANCES AND COVARIANCES IN SMALL DOMAIN ESTIMATION

Robert E. Fay[1]
U.S. Bureau of the Census

1. Introduction

Two recent reviews provide the context for this paper. The Subcommittee on Small Area Estimation, Federal Committee on Statistical Methodology (Schaible and Gonzalez 1993) surveyed applications of indirect estimators in official U.S. government use. The first chapter of their report noted the predominance of direct estimators in federal statistics. In other words, official estimates are almost always "direct," through exclusive or almost exclusive use of data from both the time and domain of interest. Indirect estimators, relying heavily on data from either other domains or times (or both), are the exception in federal statistics. The report enumerates and discusses indirect estimators in current federal use. (That is, the report considered only those applications published as official estimates, not including methodological tests and discontinued series. Generalizations of survey variance estimates, for example, those often included in source and reliability statements at the end of Census Bureau reports, were also not included.)

Although infrequently employed in federal practice, indirect estimation generally reflects an attempt to address a need for estimates that cannot be reliably produced directly given constraints on resources. The concluding chapter of the report urged caution in the use of indirect methods and eschewed advocacy of them as a general purpose and easily developed solution.

Ghosh and Rao (1994) reviewed the statistical methodology underlying several types of indirect estimators. Their review included demographic and other methods specific to postcensal population estimation; synthetic, composite and related estimators for domain characteristics; and empirical best linear unbiased predictors (EBLUP), empirical Bayes (EB), and hierarchical Bayes (HB). This paper employs their review as a point of departure for comparisons of existing theory to practice.

Several small domain applications that have appeared in the literature share enough common features to be studied as a group. One class of applications, which represents the scope of this paper, combines information from survey estimates at the domain

---

[1] This article represents results of research undertaken by a staff member of the Census Bureau. The views expressed are attributable to the author and do not necessarily reflect those of the U.S. Bureau of the Census.

level with domain-level characteristics available from independent sources. Examples include estimates of 1970 census income for small places (Fay and Herriot 1979), estimates of 1980 census undercount (Ericksen and Kadane 1985, Cressie 1992), estimates of 1990 census undercount produced by the U.S. Census Bureau in 1991 (subsequently revised under a different methodology), and estimates of median family income by state (Fay, Nelson, and Litow 1993).

In some of these applications, independent data provide a basis for evaluating the methodology. For example, in estimating median family income by state, the decennial census figures serve as a gold standard by which to judge the performance of the resulting small domain estimates. Although this comparison is available only every 10 years, the empirical results support the application. As a second example, the relatively small number of available special censuses taken after the 1970 census also corroborated the application to 1970 census income for small places. In other cases, however, including the analysis of census undercount, there is no gold standard by which to evaluate the resulting estimates. Consequently, the validity of the application of the underlying theory for both the properties of the resulting estimates and the measurement of their reliability is of considerable importance.

Comparison of these and other applications to the available theory generally shows that the explicit theoretical conditions are not completely satisfied, although to varying degrees. Consequently, each application implicitly requires that the departures from the theory do not pose serious consequences. As the title of this paper suggests, the theoretical results typically assume that the sampling errors of the small domain estimates are known, whereas in practice they are frequently estimated from the data, either directly or through a model to generalize the variances.

Section 2 reviews much of the existing theory for the class of estimators under discussion. Section 3 then compares the applications just mentioned to the requirements of the theoretical formulations to note implicit extensions of the theory that, for the most part, still lack a theoretical foundation. Section 4 reports the results of simple Monte Carlo studies to assess evidence in some of these areas. Although mathematical proof is preferable to computer demonstration, the empirical results present useful evidence on the significance of various issues arising from the practical application of these procedures and suggest directions for new research.

2. Theoretical Results for a Class of Small Area Estimators

As noted in the previous section, Ghosh and Rao (1994) reviewed several general small area approaches. The class of

models of interest to this paper employs auxiliary data $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$, which are assumed measured without sampling error. In their notation, the parameters of interest, θ_i, are assumed to be related to the $\boldsymbol{x}_i$ by

$$\theta_i = \boldsymbol{x}_i\beta + \nu_i z_i, \qquad i = 1, \ldots, m, \tag{2.1}$$

Frequently, the model takes the simpler form:

$$\theta_i = \boldsymbol{x}_i\beta + \nu_i, \qquad i = 1, \ldots, m, \tag{2.2}$$

where β is a vector of regression parameters, and the ν_i are independent, identically distributed (iid) random variables with:

$$E(\nu_i) = 0, \quad V(\nu_i) = \sigma_\nu^2. \tag{2.3}$$

In (2.1), the z_i's are known positive constants. Ghosh and Rao (1994) develop the theory in the general form (2.1). Results specific to the simpler model (2.2) are offered here because the formulas are more accessible.

The θ_i in (2.1) and (2.2) represent the parameters of interest for the small areas, such as local area per capita income, the ratio of correction population to census population, the number of employed, etc. The model reflects a possible lack of fit between the regression $\boldsymbol{x}_i\beta$ and the actual value through random effect terms, ν_i.

In this class of models, direct estimates, $\hat{\theta}_i$ are available at the domain level with

$$\hat{\theta}_i = \theta_i + e_i, \tag{2.4}$$

where the e_i represent sampling errors with

$$E(e_i|\theta) = 0, \quad V(e_i|\theta) = \psi_i. \tag{2.5}$$

(In this section, the sampling errors are also assumed independent, but extensions have reflected correlated sampling errors.) In other words, the $\hat{\theta}_i$ are design-unbiased estimators. Ghosh and Rao comment that these conditions may be quite restrictive. For example, the estimators may not be unbiased, as in the case of undercount adjustment. In addition, the sampling variances ψ_i may not be known.

The combined model, using (2.2) and (2.4), is

$$\hat{\theta}_i = \boldsymbol{x}_i\beta + \nu_i + e_i . \tag{2.6}$$

As Ghosh and Rao note, (2.6), which is a linear combination of fixed and random effects, is a special case of the general mixed linear model.

Ghosh and Rao (1994) discuss the estimation of (2.6) from the perspectives of EBLUP, EB, and HB; this paper will primarily focus on the EBLUP formulation. They cite Henderson (1950) as the originator of best linear unbiased predictors (BLUP) for models such as (2.6), when the variance components are known. Ghosh and Rao express the BLUP of θ_i as

$$\hat{\theta}_i^H = \gamma_i\hat{\theta}_i + (1-\gamma_i)\boldsymbol{x}_i\hat{\beta} , \tag{2.7}$$

where

$$\hat{\beta} = (\boldsymbol{x}^T\boldsymbol{V}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{V}^{-1}\hat{\theta} \tag{2.8}$$

is the BLUE of β , $\boldsymbol{V}$ is the diagonal matrix with elements $\sigma_\nu^2+\psi_i$, and

$$\gamma_i = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \psi_i} . \tag{2.9}$$

When the variance components are known, the mean square error of (2.7) under model (2.6) is

$$E(\hat{\theta}_i^H - \theta_i)^2 = g_{1i}(\sigma_\nu^2) + g_{2i}(\sigma_\nu^2) , \tag{2.10}$$

where

$$g_{1i} = \gamma_i \psi_i , \tag{2.11}$$

and

$$g_{2i}(\sigma_v^2) = (1-\gamma_i)^2 \, \boldsymbol{x}_i (\boldsymbol{x}^T \boldsymbol{V}^{-1} \boldsymbol{x})^{-1} \boldsymbol{x}_i^T . \tag{2.12}$$

Because σ_v^2 is typically unknown, a two-stage estimator, $\hat{\theta}_i^H$, arises by first estimating σ_v^2 from the data and then using it to obtain (2.7). Ghosh and Rao reference several options for estimating σ_v^2. A simple moment estimator $\sigma_{v(1)}^2 = \max(\sigma_{v(1)}^2, 0)$, where

$$\sigma_{v(1)}^2 = (m-p)^{-1} \left[(\hat{\theta} - \boldsymbol{x}\beta^*)^T (\hat{\theta} - \boldsymbol{x}\beta^*) - \sum_i \psi_i \{1 - \boldsymbol{x}_i (\boldsymbol{x}^T \boldsymbol{x})^{-1} \boldsymbol{x}_i^T\} \right], \tag{2.13}$$

and

$$\beta^* = (\boldsymbol{x}^T \boldsymbol{x})^{-1} \boldsymbol{x}^T \hat{\theta}$$

is the ordinary least squares estimator of β, has the advantage of not requiring iteration.

The remaining methods to be considered here each require iteration, unless the sampling errors ψ_i are equal. For a given trial estimate of σ_v^2, β is estimated through (2.8) at each cycle for each of the methods.

Fay and Herriot (1979) used an estimator $\sigma_{v(2)}$, based on the method of moments as the solution to the equation

$$(\hat{\theta}_i - \boldsymbol{x}\beta)^T \boldsymbol{V}^{-1} (\hat{\theta}_i - \boldsymbol{x}\beta) = m - p \tag{2.14}$$

or 0 in the case that no solution exists. If the sampling errors ψ_i are equal, then (2.13) and (2.14) have the same solution.

A third alternative, maximum likelihood, $\sigma_{v(3)}$, maximizes

$$L(\beta, \sigma_v^2) = -1/2 \log(|\boldsymbol{V}|) - 1/2 (\theta - \boldsymbol{X}\beta)' \boldsymbol{V}^{-1} (\theta - \boldsymbol{X}\beta), \qquad (2.15)$$

which is the log-likelihood up to a constant.

Cressie (1992) suggested the application of another estimator, $\sigma_{v(4)}$, based on restricted maximum-likelihood (REML), which maximizes the adjusted likelihood,

$$\begin{aligned} L^*(\beta, \sigma_v^2) = & -1/2 \log(|\boldsymbol{V}|) - 1/2 (\log(|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}|) \\ & - 1/2 (\theta - \boldsymbol{X}\beta)' \boldsymbol{V}^{-1} (\theta - \boldsymbol{X}\beta), \end{aligned} \qquad (2.16)$$

where terms not involving the parameters have been dropped. Cressie (1992) further describes this procedure, which was originally developed by Patterson and Thompson (1971, 1974). In short, however, the procedure examines the likelihood of the residuals from the regression. When all the sampling errors ψ_i are equal, then (2.13), (2.14), and (2.16) have the same solution, while (2.15) yields a generally smaller estimate of σ_v^2.

Under normality of the error terms, Kackar and Harville (1984) showed that

$$E(\hat{\theta}_i^H - \theta_i)^2 = E(\hat{\theta}_i^H - \theta_i)^2 + E(\hat{\theta}_i^H - \hat{\theta}_i^H)^2,$$

where, for large m, the second term may be approximated by

$$g_{3i}(\sigma_v) = \psi_i^2 (\sigma_v^2 + \psi_i)^{-3} \overline{V}(\hat{\sigma}_v^2),$$

where $\overline{V}(\hat{\sigma}_v^2)$ is the asymptotic variance of $\hat{\sigma}_v^2$.

Prasad and Rao showed that an approximately unbiased estimator of the mean square error of the EBLUP estimator is

$$mse(\hat{\theta}_i^H) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2 g_{3i}(\hat{\sigma}_v^2) \qquad (2.17)$$

with bias of order lower than m^{-1}.

The MSE estimators studied in Section 4 share (2.17) but differ in the approach to estimate $\bar{V}(\hat{\sigma}_v^2)$. Section 4 describes these differences.

3. Assumptions Made by Some Previous Applications of EBLUP

Section 2, and the more complete review in Ghosh and Rao (1994), detail the assumptions of the available theory for EBLUP. This section briefly reviews potential discrepancies between the theory and some previously published applications.

One feature is common to all of the applications discussed here and can be assumed to occur almost universally for sample surveys, namely, that the sampling variances, ψ_i, are estimates rather than known values. The following discussion notes the consequent adaptations, which range from direct use of estimated variances to variance generalization.

Fay and Herriot (1979) described a large-scale implementation of EBLUP/empirical Bayes estimation to estimate per capita income in 1969 for small places and minor civil divisions with population below 1000 persons. The sample estimates, $\hat{\theta}_i$, were based on the long form sample of the 1980 census. Predictors included the county average PCI, 100% data from the census on housing value, and reported income from IRS returns. Because of computing constraints at the time, the authors refrained from any recalculation of the census sampling variances but instead employed the available variance generalization. The generalization was a simple national model without any allowance for geographic variation. Since the generalization yielded a linear relationship between ψ_i and θ_i^2, a logarithmic transformation of $\hat{\theta}_i$ gave a closer fit of the application to the theory. They employed (2.14) to estimate σ_v^2. Generally, the compositing, (2.7), drew on both the sample estimates and the regression in approximately equal amounts, rather than relying almost exclusively on one of the two. The authors employed (2.2) but observed some evidence of variation in σ_v^2 by size of place. The evidence suggested (2.1) with z_i decreasing with increasing size, n_i, although at a rate closer to $n_i^{-1/4}$ than $n_i^{-1/2}$. The authors did not attempt MSE estimation, but presented some limited empirical evidence from special censuses favoring the EBLUP approach.

Application of EBLUP to sample estimates of decennial census undercounts has been controversial, and the review here will simply focus on assumptions incorporated in the implementations rather than systematically evaluating the merits of the work on this subject. As Ghosh and Rao (1994) comment in passing, survey estimates of undercount in both 1980 and 1990 have been subject to substantial sources of bias, and the existing theory does not provide a clear measure of how EBLUP behaves under such conditions. Furthermore, gains from EBLUP and estimators of MSE have figured

prominently in the undercount debate, since the 1980 PES estimates at the state level and the 1990 PEP estimates based on the original 1392 strata have such high sampling variability as to preclude adjustment without EBLUP or other smoothing. The estimators placed high weight on the regression and little on the direct estimates.

In both 1980 and 1990, estimates of ψ_i have appeared to depend on $\hat{\theta}_i$. The published 1980 analysis used the estimated variances in spite of this departure from the model. The 1991 analysis of the 1990 PES applied a variance generalization. Although opinions have been offered on the subject, a systematic analysis of the effect of the generalization on the 1991 estimates remains to be done. Furthermore, the 1991 smoothing was multivariate and employed large covariance matrices, formed from the generalized variances and directly estimated correlations. Fay (1992) showed through stratified bootstrap samples that this approach induced substantial additional variability not reflected in the MSE's computed by the Census Bureau.

The 1980 PES estimates were subject to substantial amounts of missing data, yet no estimates of missing data variance are available, and the author is unaware of systematic analysis showing what possible effect this factor might have had on the 1980 analysis.

Ericksen and Kadane (1985) and the 1991 EBLUP for the 1990 PES both employed (2.2), whereas Cressie (1992) reanalyzed 1980 estimates with $z_i = n^{-1/2}$. Although Cressie argued for this choice on intuitive grounds, empirical evidence on this question is limited and virtually impossible to obtain from the undercount estimates themselves. The 1990 application employed (2.2); yet the sample estimates suggested that it failed to hold because σ_v^2 appears much larger in minority poststrata than elsewhere.

The U.S. Census Bureau has employed an EBLUP procedure to estimate median family income for 4-person families by state from the Current Population Survey (Fay, Nelson, and Litow 1993). The model can be calibrated against census values every 10 years. These calibrations have favored continued use of (2.2) at the state level, distinctly rejecting proposals such as $z_i = n^{-1/2}$ (Cressie 1992) in this application. The authors account for different approaches to estimating ψ_i and σ_v^2 over the evolution of the model. Over time, more emphasis has been placed on direct estimates.

In short, 1) each of these applications has rested on implicit extensions of the existing theory, 2) some empirical evidence suggests that these procedures can be useful under some conditions, but 3) a more systematic approach to assessing effects of uncertainty for EBLUP is still needed. The next section does not fully meet this need, but it does suggest the value of large scale Monte Carlo simulation as a productive approach to some of these questions.

4. Monte Carlo Evaluation

4.1 Basic Design of the Study

As noted earlier, the derivation of the estimators of mean square error rest on expectations taken both over repetitions of the sample and over the random effects. The more usual perspective of finite population sampling considers the population as fixed but unknown. In order to bridge the consequences of these two points of view, this study generated several finite populations, θ, and compared the properties of the mean square error estimates for each resulting population. Although the expressions in Section 2 focused on errors for individual components, (2.17), this section studies the accuracy of the estimated sum over domains of squared errors, much as the literature on the James-Stein estimator. The actual MSE's are compared to (2.17) summed over i.

Two values of m, 20 and 50, offer some indication of the effect of number of domains on the estimators. The first offers an approximate lower boundary on the range of usual application, while the second illustrates the effect of somewhat larger m. The primary emphasis will be on 50. Obviously, results for larger numbers of domains, such as 200, would also be desirable.

The section reports results that share the following common elements:

1) A set of population values for the domains, θ, is drawn from some distribution. For normal applications, for example, the domain means are selected.

2) Samples, $\hat{\theta}$, are drawn from the domain population and sampling variances estimated.

3) One or more EBLUP's, $\hat{\theta}^{H}$, are constructed.

4) Steps 2) and 3) are repeated for a total of 2000 samples from the population defined at step 1).

5) The MSE's of the EBLUP's, estimated as the average value of $(\hat{\theta}_i^{H} - \theta_i)^2$, calculated from the 2000 samples and summed over the domains, i, become the standard for comparison to the corresponding estimated MSE sums of (2.17) over i.

6) The bias and mean square error of the MSE estimators are then derived, and also key frequencies, such as the percent of samples in which the estimated MSE understates the actual MSE by 25 percent.

Steps 1)-6) generate one point in the Monte Carlo study. In other words, each point represents a specific population realized from

the superpopulation, where the performance, over repeated sampling, of each of the EBLUP estimators and MSE estimators is evaluated.

Obviously, the MSE's at step 5) are not entirely free from sampling error themselves, but the relatively large number of samples provides practical justification for this procedure. The results presented in Figures 1-13 show that this procedure produces stable values.

Thus, the perspective is similar to design-based finite population sampling, since the criteria assess the performance for individual over repetitions of the sampling design.

4.2 Results for Normal Populations

For simplicity, four X variables were constructed:

1) The grand mean.

2) An indicator variable dividing the domains into halves according to domain number, $i = 1, \ldots m$. For example, for $m = 20$, the variable distinguishes the first 10 from the second.

3) A linear term, increasing with the domain number.

4) A similar quadratic term.

The sample size, n, for each domain was fixed at either 10 or 20, and the observations were scaled by $n^{1/2}$ in order to give the sample means unit variance. Because the analysis is invariant to the true β, these coefficients were set to 0 in generating the Monte Carlo samples.

Eight combinations were studied:

1) Use of known sampling variances, $\psi_i = 1$, in combination with (2.14) and $\overline{V}(\sigma_v^2)$ from Prasad and Rao (1990, p. 167, (5.19)):

$$\overline{V}(\sigma_v^2) = 2m^{-1}\left[\sigma_v^4 + 2\sigma_v^2 \sum \psi_i/m + \sum \psi_i^2/m\right] \qquad (4.1)$$

2) Use of generalized sampling variances assuming that the ψ_i are equal to some unknown constant, which is then estimated as the average of the sample estimates of ψ_i. The remaining estimation is done as in 1).

3) MLE using known sampling variances, $\psi_i = 1$, and:

$$\overline{V}(\sigma_v^2) = 2\left[\sum (\sigma_v^2 + \psi_i)^{-2}\right]^{-1} \qquad (4.2)$$

4) MLE using the estimated variances and the estimation approach of 3).

5) REML using the estimated variances and $\overline{V}(\sigma_v^2)$ from Cressie (1992, p. 82, (3.22) and p. 85, (4.11)).

6) REML using the estimated variances and the more approximate expression (4.2).

7) The method of moments estimator, (2.14), and:

$$\overline{V}(\sigma_v^2) = 2m\left[\sum(\sigma_v^2 \cdot \psi_i)^{-1}\right]^{-2} \qquad (4.3)$$

8) The simple moment estimator, (2.13), and (4.1).

Cressie's (1992) estimator studied included in 5) is the only one of the group to explicitly incorporate the effect of the regression in estimating $\overline{V}(\sigma_v^2)$. All others depend on m being large compared to p. In fact, however, differences between 5) and the simpler 6) were extremely modest. Potentially, similar refinements could be incorporated into the other estimators of $\overline{V}(\sigma_v^2)$, but their impact is again likely to be small unless p is a substantial proportion of m.

Figures 1-13 results for $m = 50$, $n = 10$, that is a comparatively large number of domains with comparatively few degrees of freedom in each domain to estimate the variance in each. Of course, no one choice of these values is appropriate to represent the usual situation in most small domain estimation. Comments will follow about the results obtained for $m = 20$ and for $n = 20$.

A series of 28 populations are represented: 4 drawn from $N(0, \sigma_v^2)$ with $\sigma_v^2 = .125$, and 8 each from $\sigma_v^2 = .25$, .5, and 1.0. Figures 2, 4, 6, 8, 10, and 12 each omit the points for $\sigma_v^2 = .125$, which are generally far off the scale; further comments on this point follow.

Figure 1 shows the actual MSE for 2) as a function of $\sum\theta_i^2$, which is called the "SS of true deviations" in the figures. Over the entire range studied, the EBLUP improves on the direct sample estimates, but the improvement is most dramatic at the leftmost portion of the range, where $\sigma_v^2 = .125$, and the true values θ almost fit the regression line. The pluses and x's distinguish between different super population values of σ_v^2 used to generate the θ_i,

but quite clearly this distinction is unimportant once the results are conditioned on $\sum \theta_i^2$.

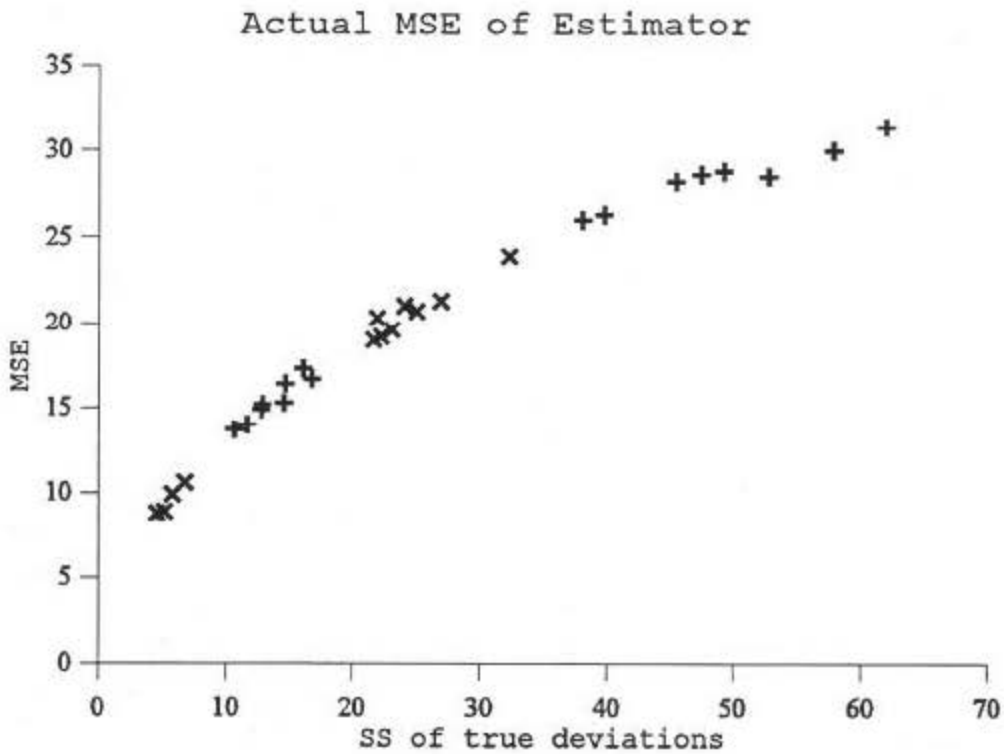


Figure 1 Actual MSE for "unbiased methods," generalized equal variances, m = 50, p = 4, 10 obs. per domain.

If the actual MSE's of any of the alternatives were superimposed on Figure 1, there would be substantial overlap. The MSE's and other performance characteristics of 1), with known variances, are virtually identical to 2). The actual MSE's of 3), MLE with known variances, are also almost identical to those in Figure 1. When sampling errors are instead estimated, the actual MSE's are a bit larger: by about 15-30% for REML and 8-20% for the other alternatives when the actual MSE is below 20, and by lesser amounts over the upper end of the range.

Since the MSE of the sample means is 50, Figure 1 includes a broad range of outcomes. At MSE=30, EBLUP yields distinct gains that, nonetheless, many practitioners might choose to forego in favor of the greater simplicity and interpretability of the direct

sample estimates. At MSE=15, the gains from EBLUP may have a substantial impact on the utility of the estimates.

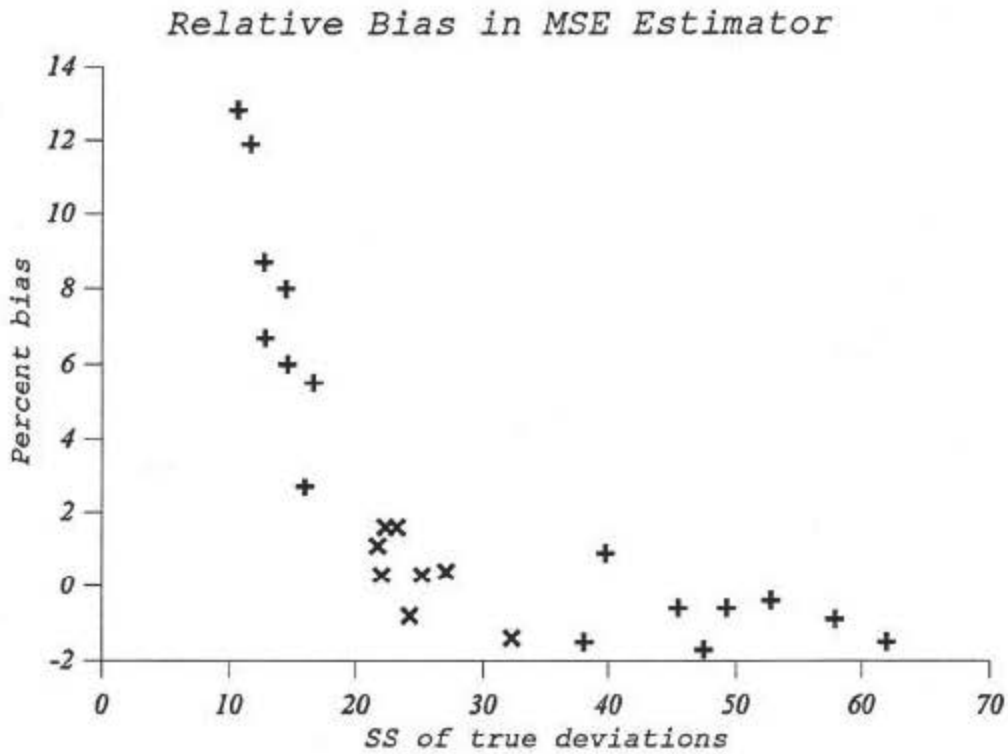


Figure 2 Percent bias in Prasad-Rao MSE estimator, generalized equal variances, "unbiased methods," m = 50, p = 4, 10 obs. per domain. Note: The first 4 points have been omitted.

Figure 2 reports the relative bias of the MSE estimator for 2) over the range of $\sum \theta_i^2$. The leftmost 4 points have been omitted from the graph because the bias increases dramatically, to around 30-40%, in that region. As noted previously, the performance for 1), with known sampling variances, is virtually identical to Figure 2.

From the perspective of bias, the performance of the MSE estimator is quite satisfactory over a large part of the range, but it becomes upwardly biased under conditions where the EBLUP estimator has the most pronounced effect, that is, in the leftmost portion of the range, below MSE = 15 or so.

Figure 3 evaluates the performance of the MSE estimator in a different manner, by showing the proportion of times that the estimated MSE falls below the actual MSE by 25% or more. For example, when the actual MSE=20, the figure reports the percentage of samples in which the estimated MSE is below 15.

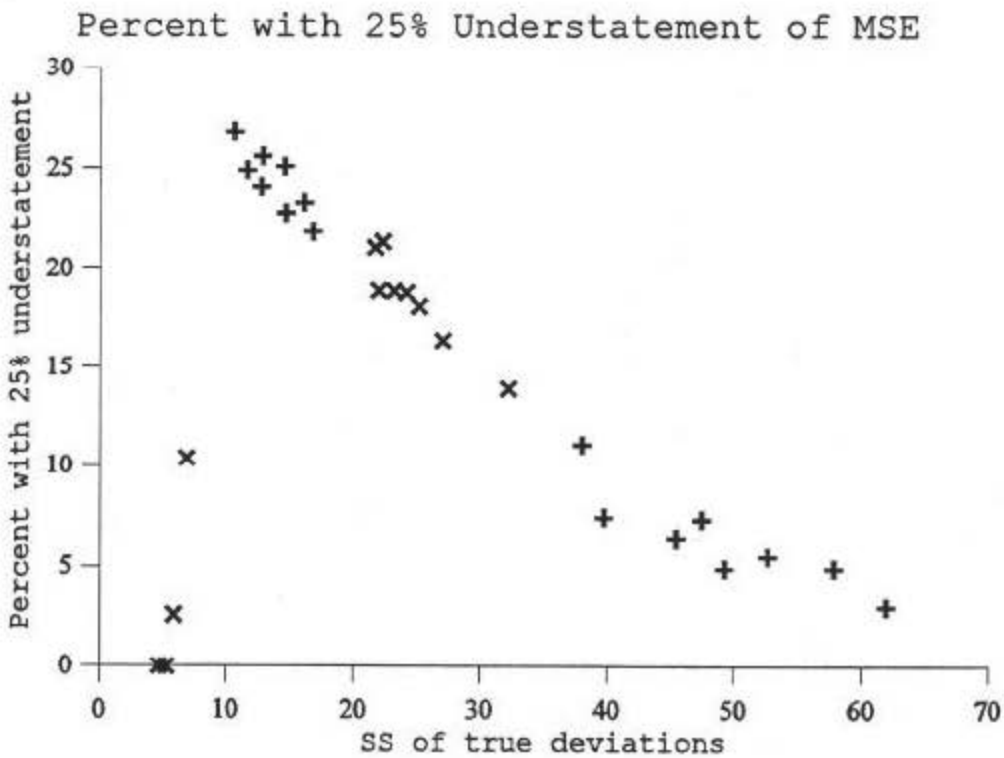


Figure 3 Percent understatement of true MSE by 25 percent or more, generalized equal variances, "unbiased methods," m = 50, p = 4, 10 obs. per domain.

The findings of Figure 3 are not easily predicted from Figure 2. In spite of the low level of bias in the MSE estimator over the range of MSE=20 and above, the probability that the estimator will substantially understate the actual MSE rises steadily as MSE decreases. Even more striking, however, is the dramatic fall towards 0% at the right of the figure. In fact, in this lower range, the contributions of the more stable components of (2.17), namely its second and third terms, are able to prevent a large understatement regardless of the contribution of the far more erratic first term.

Figure 4 presents comparable results for 3), MLE with known variances. Figure 4 reports a consistent downward bias in the estimated MSE for MLE. This finding agrees with a comparison of REML and MLE by Cressie (1992). Presumably, this downward bias could be even more severe when the ratio of p to m, which is 4 to 50 in this case, is larger.

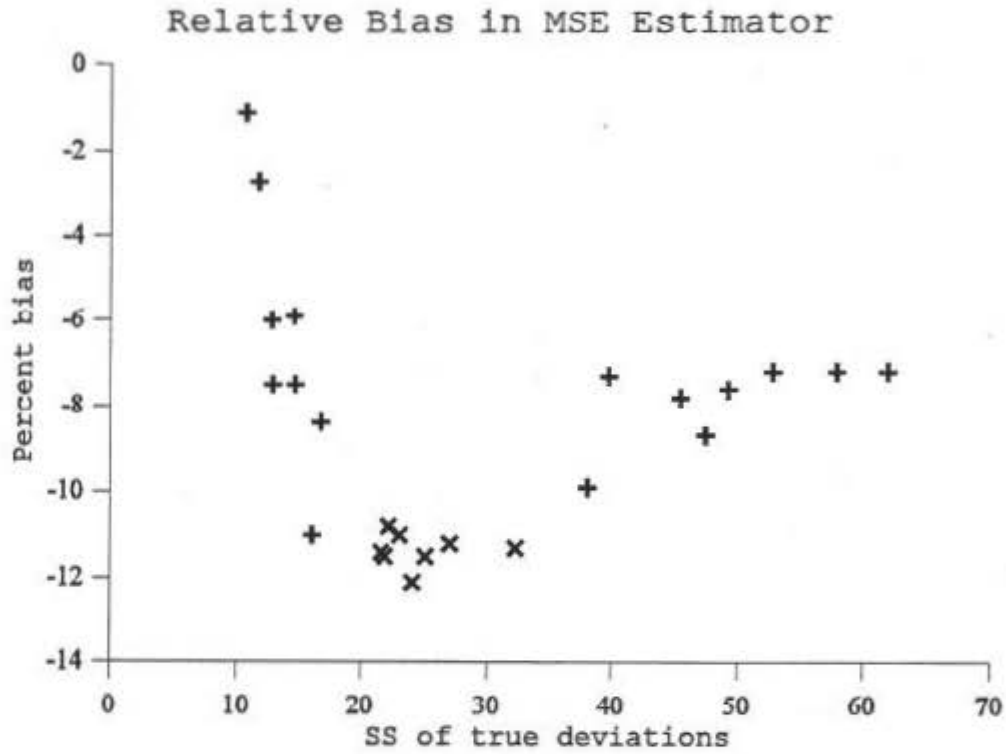


Figure 4 Percent bias in Prasad-Rao MSE estimator, generalized equal variances, MLE, $m = 50$, $p = 4$, 10 obs. per domain. Note: The first 4 points have been omitted.

In spite of the general downward bias in the MSE estimate, the bias changes sign and increases up to about 15-30% for the 4 lowest points included in the study.

Figure 5 presents results for MLE analogous to those in Figure 3. Noting the change in scale between the two figures, Figure 5 shows even higher proportions of significant understatement of the MSE over a large proportion of the range. This finding is consistent with the general downward bias exhibited in Figure 4.

As in Figure 3, however, the probability of significant understatement falls off dramatically near 0.

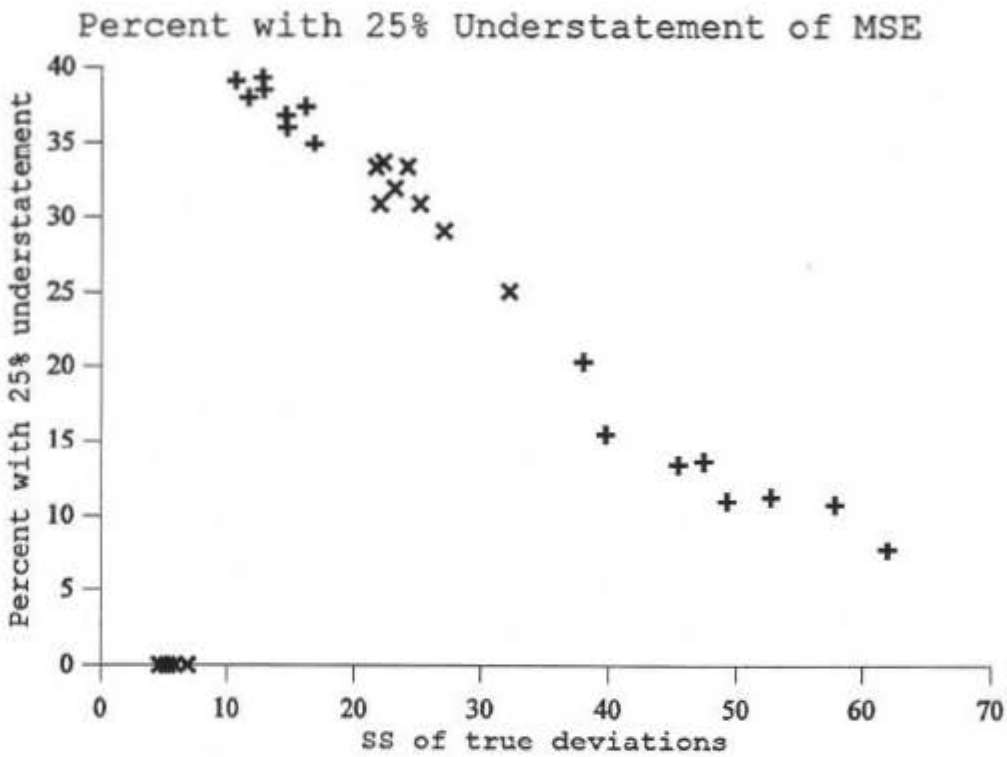


Figure 5 Percent understatement of true MSE by 25 percent or more, generalized equal variances, MLE, $m = 50$, $p = 4$, 10 obs. per domain.

As noted earlier, shifting from known variances to estimated variances for each of the domains increases the actual MSE of the MLE by about 8-20% for actual MSE's below 20, and somewhat less for larger actual MSE's. Figure 6 reports the performance of the MSE estimator in this instance, as an estimator of the actual, and now larger, MSE. Comparison of Figures 4 and 6 indicates some common features but considerable differences as well. On the right of Figure 6, the downward bias is even more pronounced than in Figure 4. For decreasing MSE, however, the bias crosses 0 earlier than in Figure 4. The bias for the omitted points rises to approximately the same range, that is, about 15-30%.

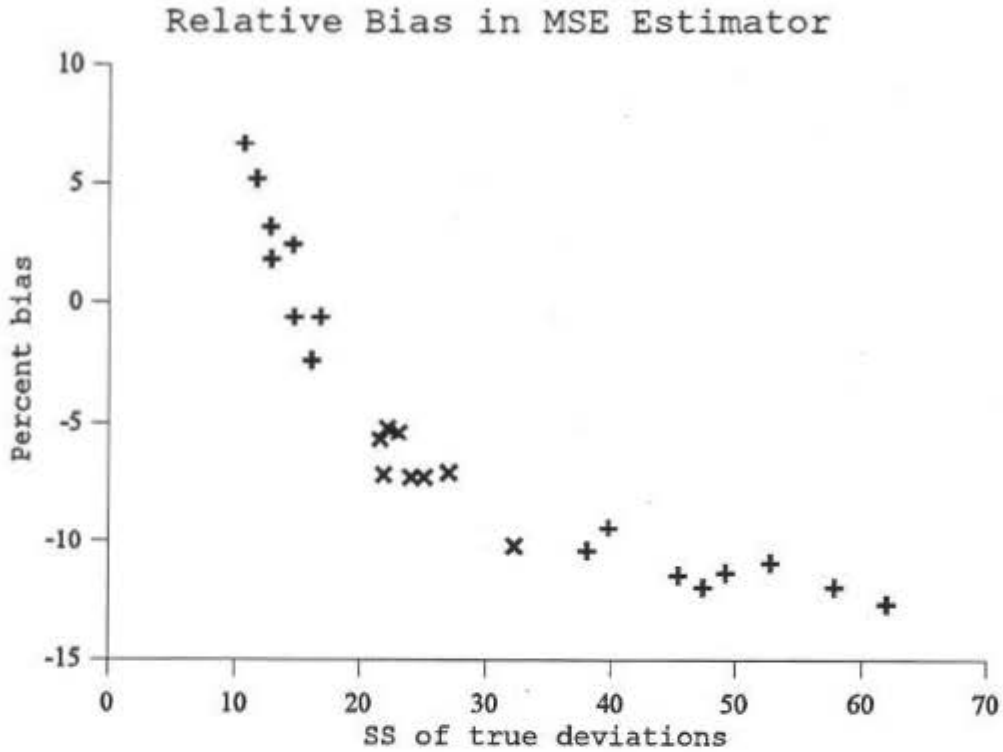


Figure 6 Percent bias in Prasad-Rao MSE estimator, estimated variances, MLE, $m = 50$, $p = 4$, 10 obs. per domain. Note: The first 4 points have been omitted.

Figure 7 shows the effect on 25% understatement of the actual MSE when the sampling errors are estimated. Compared to Figure 5, the results are much flatter, in the range of 15-20%, compared to the much more dramatic swings in Figure 5. Unlike Figures 3 and 5, the combination of the extra variability from estimating the sampling variances and the somewhat larger actual MSE eliminates the phenomenon of the dramatic drop towards 0% at the right end of the scale.

It was previously noted that REML applied to the sample data and estimated sampling variances yielded estimates with the largest actual MSE. Specifically, choice 5), with the estimator from Cressie (1992), appears here, although it was previously noted that the alternative 6) produces essentially identical results. Figure 8 shows the bias in the estimated MSE for REML. Figure 8 closely resembles Figure 6 in shape but has estimated biases moved up by

roughly 5-10 percentage points. Again, results of this comparison to MLE are consistent with a greater downward bias in the MSE for the latter.

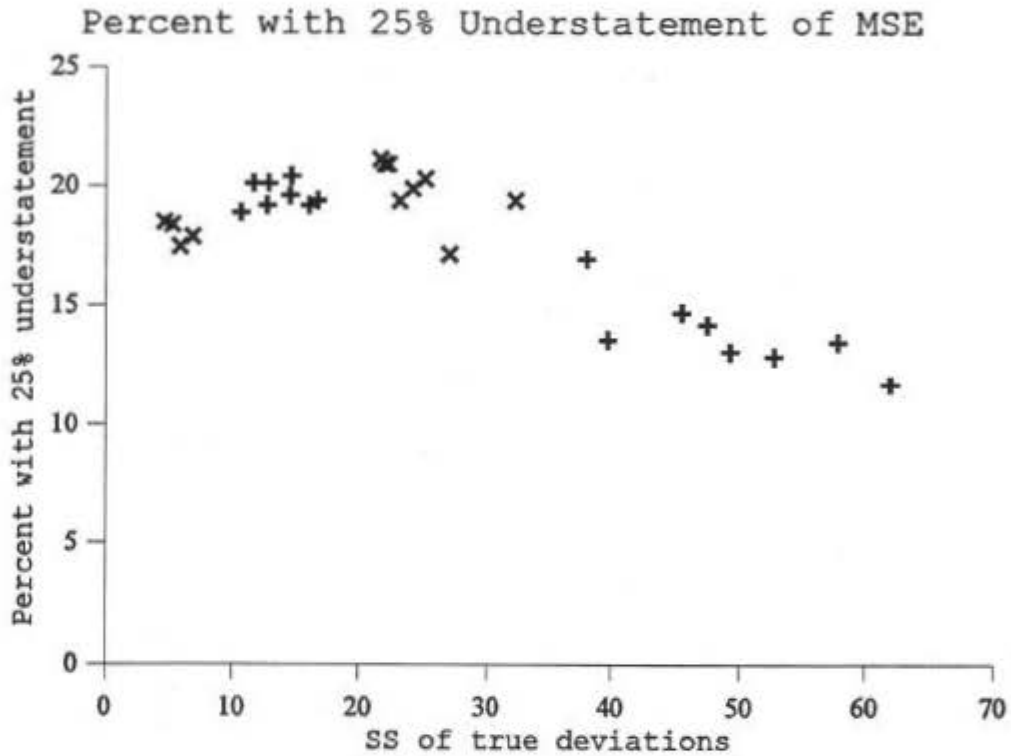


Figure 7 Percent understatement of true MSE by 25 percent or more, estimated variances, MLE, $m = 50$, $p = 4$, 10 obs. per domain.

Figure 9 resembles Figure 7, in showing a flatter performance over the range than Figures 3 and 5. Overall, however, the comparison of Figure 9 to Figure 7 awards a significant advantage to REML compared to MLE in preventing marked understatement of the true MSE. This finding is consistent with the relative shift in bias of the MSE estimators compared in Figures 6 and 8.

As noted earlier, use of sample variances in the method of moments estimator produces an increase in actual MSE comparable to the increase for MLE. Figure 10 shows performance comparable or slightly better than that of REML in Figure 8 under the same

circumstances. Again, the MSE estimates exhibit less downward bias than for MLE in Figure 6.

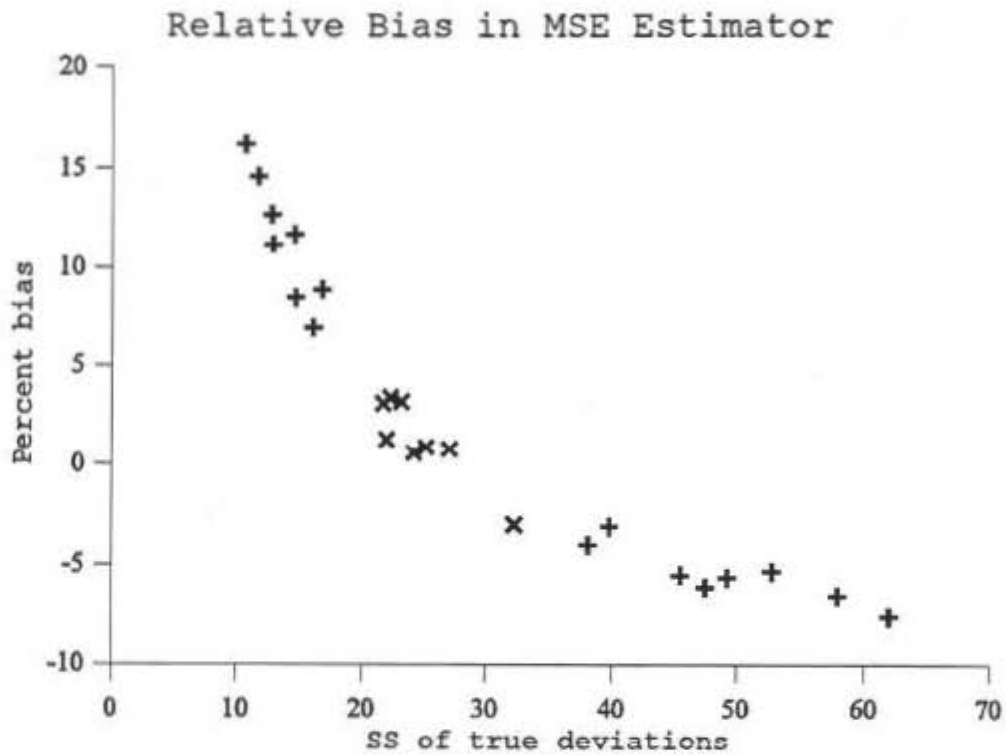


Figure 8 Percent bias in Prasad-Rao MSE estimator, estimated variances, REML, $m = 50$, $p = 4$, 10 obs. per domain. Note: The first 4 points have been omitted.

Comparison of Figures 11 and 9 reveals that the slight bias advantage of the method of moments approach compared to REML, shown previously by Figures 10 and 8, is traded against more frequent understatement of the actual MSE by 25% or more. Consequently, there is not a single winner in the contest of these alternatives.

Generally, the method of moments approach does appear to outperform MLE in Figures 6 and 7. The method of moments is subject to less downward bias than MLE at the upper end of the range studied and exhibits less frequent understatement.

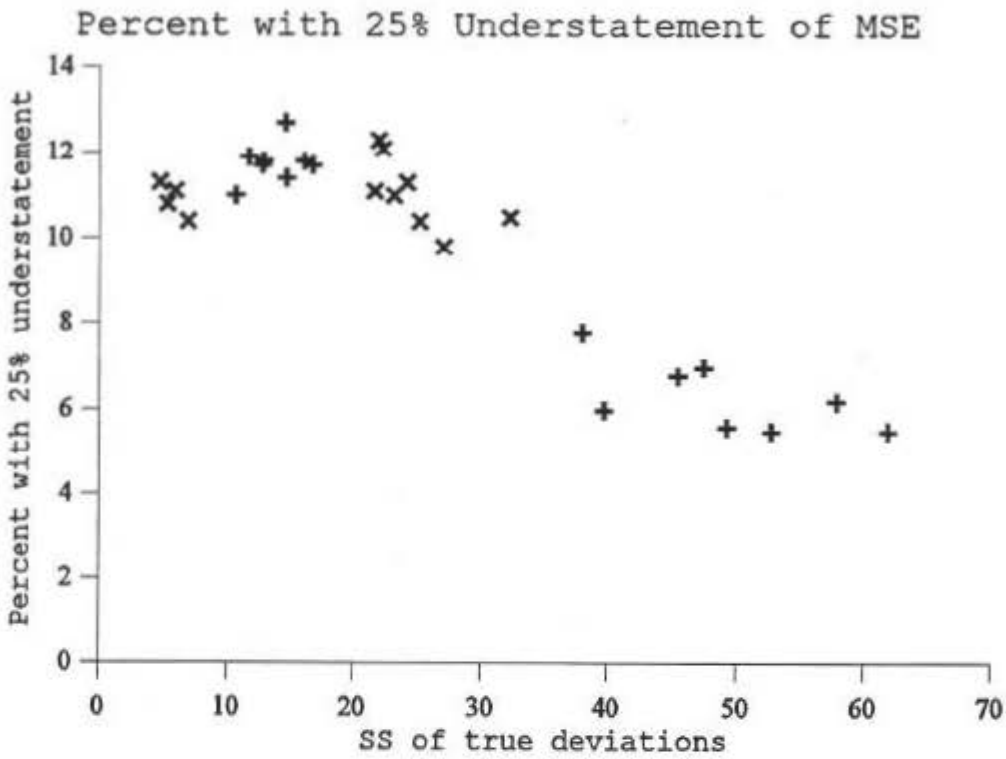


Figure 9 Percent understatement of true MSE by 25 percent or more, estimated variances, REML, $m = 50$, $p = 4$, 10 obs. per domain.

Figures 12 and 13 present the results for the last alternative, 8), which weights observations equally in estimating σ_v^2 and which does not require iteration. The findings show a considerable downward bias in MSE estimation under these conditions. For example, comparison of Figure 12 to Figure 10 shows a more consistent downward bias over much of the range studied. In turn, the probability of 25% understatement is higher in Figure 13 than Figure 11.

Generally, the findings show that the properties of the MSE estimators are affected to a significant degree as a result of estimating sampling variances when there are relatively few observations or degrees of freedom in each of the domains. These empirical findings do not appear to be a straightforward consequence of the available theoretical results.

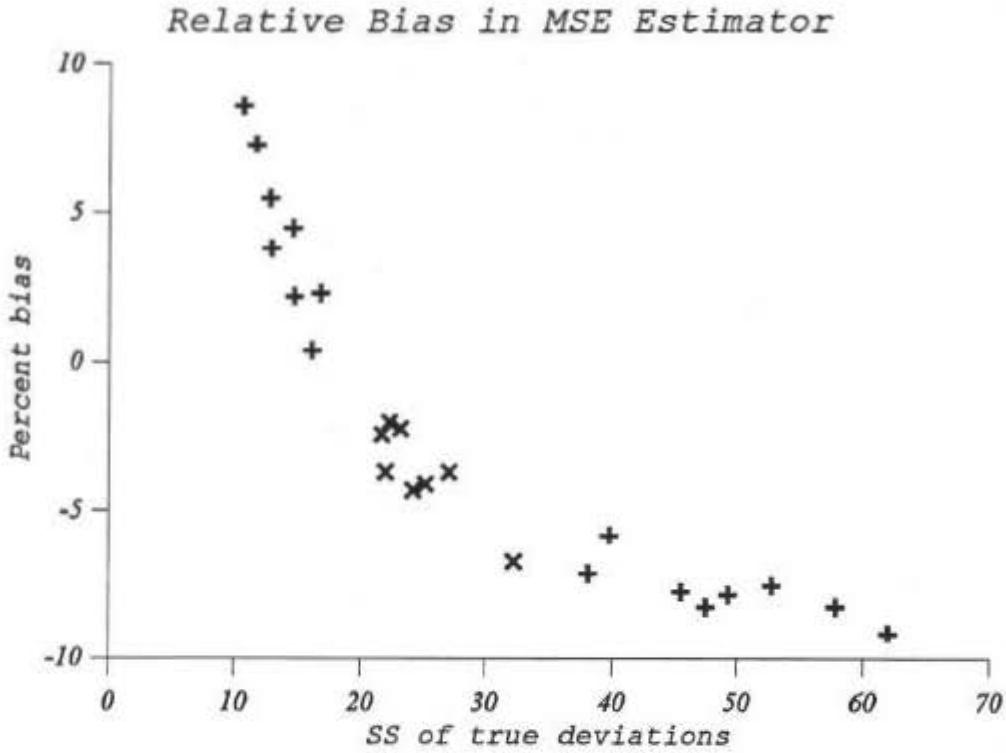


Figure 10 Percent bias in Prasad-Rao MSE estimator, estimated variances, method of moments, $m = 50$, $p = 4$, 10 obs. per domain. Note: The first 4 points have been omitted.

When $n = 20$ observations are instead available for variance estimation within each cluster, the effects of estimating the variances becomes less pronounced. In other words, the corresponding Figures 6 and 7 for $m = 20$ become more like Figures 4 and 5, and the pairs of Figures 8 and 9, 10 and 11, and 12 and 13 each resemble Figures 2 and 3 more closely. Consequently, and not surprisingly, the effect on MSE estimation depends on the degree of precision of the sampling variances in the domain, and not simply on the fact that the sampling variances have been estimated.

Translation of the implication of these results to application will, in the author's opinion, not be simple. Compared to the estimation of variance for standard estimates, such as the sample mean, the issue of the variance of the variance, that is, the design-based variance of a variance estimator, is a fairly arcane subject that has consequently received relatively little attention.

A simple count of the algebraic degrees of freedom will not typically provide an adequate indication of the expected performance of the variance estimator, except in the sense that a variance estimator based on a small number of observations or clusters is certain to be highly variable. Generally, non-normality of the individual or clustered observations may increase the variance of the variance substantially compared to its behavior under normality.

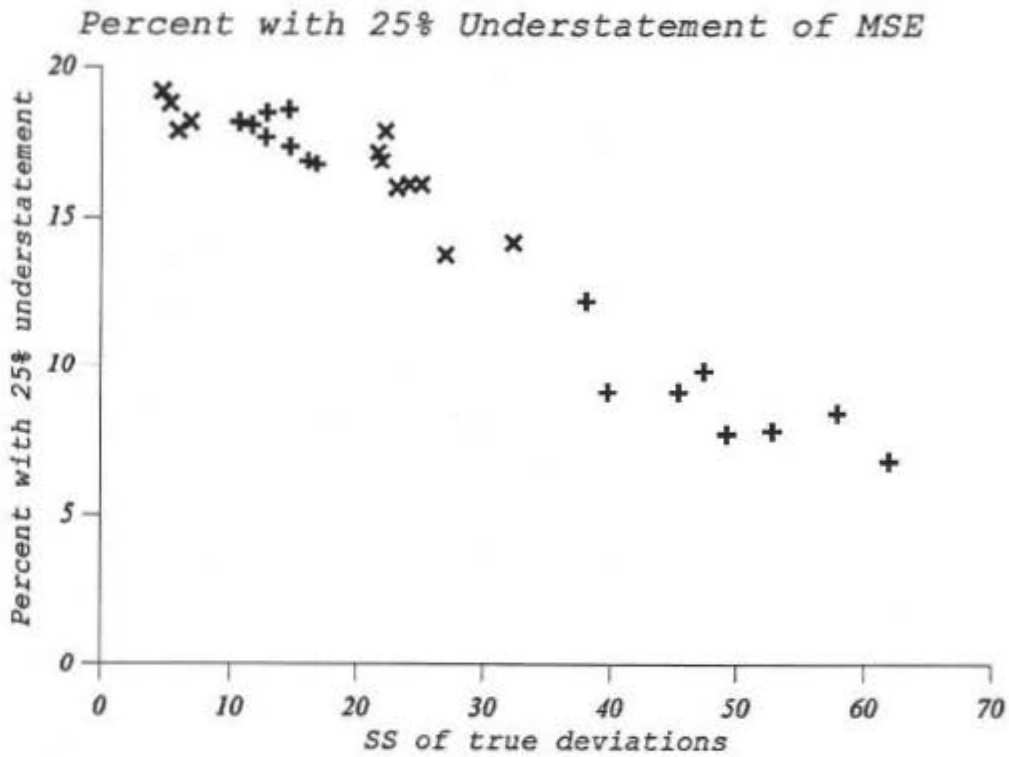


Figure 11 Percent understatement of true MSE by 25 percent or more, estimated variances, method of moments, $m = 50$, $p = 4$, 10 obs. per domain.

Results for $m = 20$ domains follow many of the same patterns as $m = 50$. Overall, however, there is substantially less evidence to evaluate whether the EBLUP has yielded substantial declines in MSE. When $m = 50$, the MSE estimators begin to exhibit relatively extreme behavior, including their upward bias, when the actual reductions are 75% or more. For $m = 20$, the same patterns appear much earlier, at around 50% actual reduction. Similarly, the phenomenon

in Figure 3 and others where the MSE estimator suddenly stops overestimating the true MSE by 25% or more shows up much earlier for $m = 20$. Thus, effective MSE estimation in situations where the gains from EBLUP are substantial requires numbers of domains on the order of $m = 50$. Specific findings are available from the author.

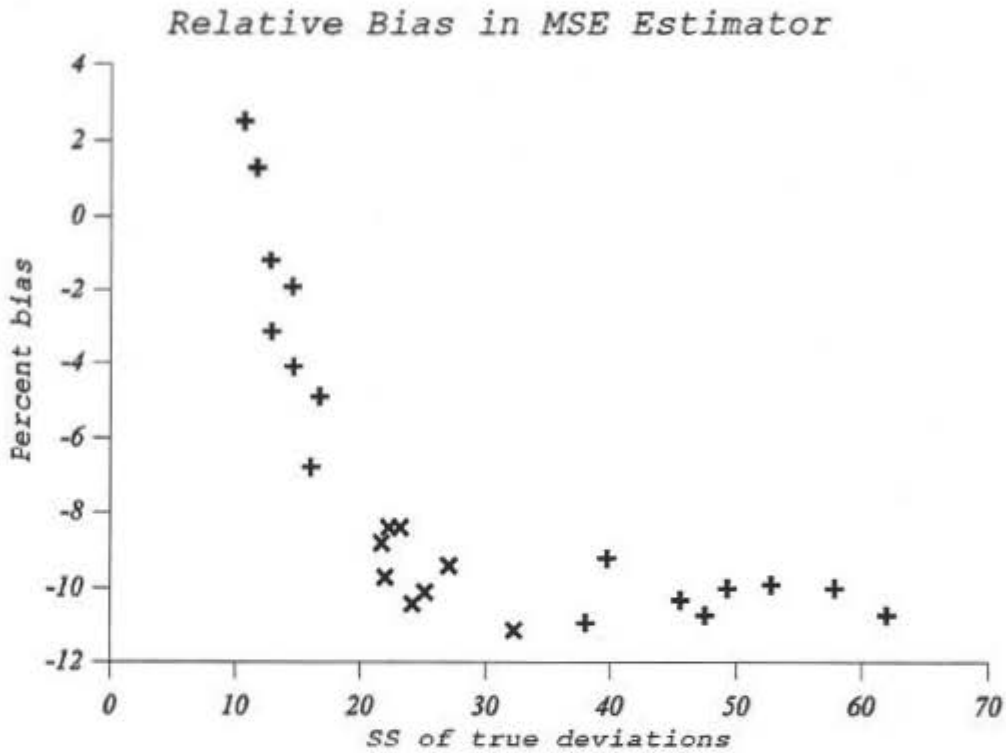


Figure 12 Percent bias in Prasad-Rao MSE estimator, estimated variances, equally weighted method, $m = 50$, $p = 4$, 10 obs. per domain. Note: The first 4 points have been omitted.

Except for separate FORTRAN programs to generate the sample data used in the Monte Carlo study, the variance program VPLX calculated the EBLUP estimators and summarized the results. PC's with 486-class processors performed the calculations for $m = 20$, and a Sun SPARC 10 for $m = 50$, although selective problems were checked against each other to verify independence of results on the choice of platform.

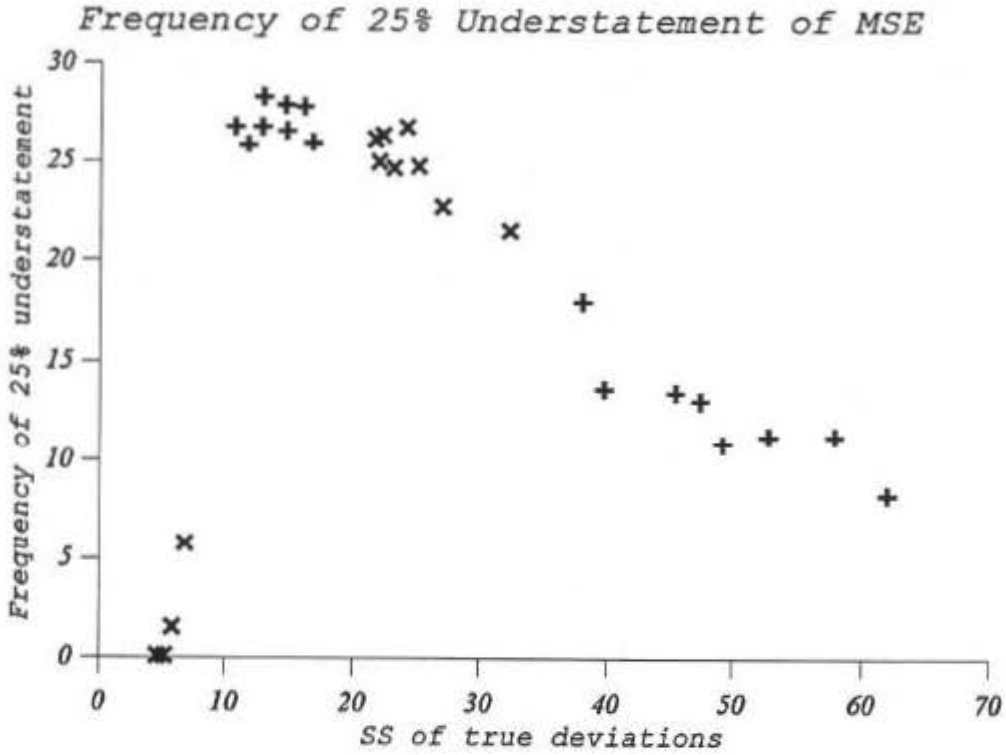


Figure 13 Percent understatement of true MSE by 25 percent or more, estimated variances, equally weighted method, $m = 50$, $p = 4$, 10 obs. per domain.

5. Concluding Remarks

Continued advances in computer technology is certain to have a continued impact on the practice of statistics. Figures 1-13 summarize empirical results that the author would not have had the resources to undertake even a few years ago. Even so, such answers are not yet easily obtained -- for example, each set of points appearing in Figures 1-13 represents about 5 1/2 hrs. of calculation.

The findings, although not generally remarkable, illustrate the subtleties of applying complex estimation methods to practical problems. Features appear that are difficult to anticipate from knowledge of the theoretical results alone. Over time, Monte Carlo

assessment should become even more of a standard to complement theoretical findings.

Substantially more work can and should be done. Section 4.1 outlines a general strategy for useful additional study. As examples, the effect of linkage between θ_i and ψ_i can and should be studied in this manner. Variance generalization has appeared in applications, but what are the consequences of applying a deficient model, i.e., a variance generalization that overpredicts some sampling variances and underpredicts others? What are the consequences of misspecifying (2.1)? How should the variance effects of missing data be taken into account? Issues such as these may have a substantial effect on the behavior of EBLUP procedures, and further Monte Carlo work offers an effective approach.

REFERENCES

Cressie, N. (1992), "REML Estimation in Empirical Bayes Smoothing of Census Undercount," *Survey Methodology,* **18**, 75-94.

Ericksen, E.P., and Kadane, J.B. (1985), "Estimating the Population in a Census Year (with discussion)," *Journal of the American Statistical Association,* **80**, 98-131.

Fay, R.E. (1992), "Inference for Small Domain Estimates from the 1990 Post Enumeration Survey," unpublished manuscript.

Fay, R.E., and Herriot, R.A. (1979), "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data," *Journal of the American Statistical Association,* **74**, 269-277.

Fay, R.E., Nelson, C.T., and Litow, L. (1993), "Estimation of Median Income for 4-Person Families by State," in "Indirect Estimators in Federal Programs," W. L. Schaible and M. A. Gonzalez (eds.), *Statistical Policy Working Paper 21*, Statistical Policy Office, Office of Management and Budget, pp. 9-1 - 9-17.

Ghosh, M., and Rao, J.N.K. (1994), "Small Area Estimation: An Appraisal (with discussion)," *Statistical Science,* **9**, 55-93.

Henderson, C.R. (1950), "Estimation of Genetic Parameters (abstract)," *Annals of Mathematical Statistics,* **21**, 309-310.

Kackar, R.N. and Harville, D.A (1984), "Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models," *Journal of the American Statistical Association,* **78**, 47-59.

Patterson, H.D., and Thompson, R. (1971), "Recovery of Interblock Information When Block Sizes are Unequal," *Biometrika*, **58**, 545-554.

_____ (1974), "Maximum Likelihood Estimation of Components of Variance," *Proceedings of the 8th International Biometric Conference*, Washington, D.C., Biometric Society, pp. 197-207.

Prasad, N.G.N. and Rao, J.N.K. (1990), "The Estimation of the Mean Squared Error of Small-Area Estimators," *Journal of the American Statistical Association*, **85**, 163-171.

Schaible, W.L. and Gonzalez, M.A., eds. (1993), "Indirect Estimators in Federal Programs," *Statistical Policy Working Paper 21*, Statistical Policy Office, Office of Management and Budget.

DISCUSSION

Phillip S. Kott
National Agricultural Statistics Service

For a number of years now, many of us in the survey sampling community have been grappling with the following question:

"What is the proper role of models in survey sampling?"

The answer for survey sampling purists can be found in Hansen, Madow, and Tepping (1983). Their Guiding Principle No. 4 states:

"Models are appropriately used to guide and evaluate the design of probability samples [including the choice of estimators], but with large samples the inference should not depend on the model."

This principle clearly justifies the use of model-assisted methods within a randomization-based framework, which is the basis for Särndal, Swensson, and Wretman's celebrated new textbook (1992). It is in sharp contrast, however, to the approach that Hansen and his colleagues label "model-dependent."

Unfortunately, it is not at all clear how Guiding Principle No. 4 applies to the issue of estimation in small domains. In fact, in Guiding Principle No. 7, Hansen, Madow, and Tepping concede:

"... model-dependent methods may have an advantage with quite small samples, for which probability-sampling many not be appropriate"

This suggests that our original question needs to be turned around:

"What is the proper role of randomization-based inference when estimating small domains?"

To Bayesians like Don Malec and Joe Sedransk, the answer to this question is simple: "none." Others, like Bob Fay and myself, would like to estimate the value θ_i in a small domain i with an estimator t_i that has the following property: as the sample size within domain i grows arbitrarily large (but the sampling fraction stays fixed), t_i approaches θ_i in probability irrespective of the validity of the model used in choosing t_i.

We realize, of course, that the sample size within domain i is not arbitrarily large. In fact, in small domain estimation, the sample size within i is usual so small that a conventional model-

assisted, randomization-based estimator, $t_{i(rb)}$, has an unacceptably large standard error, hence the need for a more creative, small domain estimator in the first place! Still, we would not be happy using an estimator that did not work well when it should; that is, when the sample size within domain i was large.

One can write

$$t_{i(rb)} = \theta_i + s_i, \tag{1}$$

where s_i is the sampling error of estimator $t_{i(rb)}$. Let us assume that the model-assisted randomization-based estimator $t_{i(rb)}$ is (at least) nearly randomization unbiased so that $E_p(s_i) \approx 0$, where the subscript p denotes that the expectation is with respect to the probability sampling process ("nearly unbiased" means that the bias is small because the sample size across all domains is large). Let us also assume that $\theta_{i(rb)}$ is nearly unbiased under a model governing the elements of the population; that is to say, $E_M(s_i) \approx 0$, where the subscript M denotes that the expectation is with respect to the model. Finally, let us assume that the $t_{i(rb)}$ is randomization consistent; i.e. $\text{plim}_{n(i)\to\infty}(t_i/\theta_i) = 1$, where n(i) is the sample size in domain i.

In small domain estimation, it is common to model the behavior of the domain values θ_i as well as of the population's elements. For convenience, let us restrict our attention to the following domain-level "random effects" model:

$$\theta_i = \mu(\mathbf{x}_i) + \epsilon_i, \tag{2}$$

where $\mathbf{x}_i$ is a vector of characteristics for domain i, μ has a known functional form (e.g., linear or logistic) but unknown parameters, and ϵ_i, the random effect, is a random variable with mean zero and positive variance.

Let m_i be a nearly unbiased estimator for $\mu(\mathbf{x}_i)$. An estimator for t_i of the form:

$$t_i^{(g)} = (1 - g)t_{i(rb)} + gm_i$$

is nearly model unbiased. Its mean squared error is (approximately) minimized when

$$g = \frac{\text{Var}(s_i)}{\text{Var}(s_i) + E[(\theta_i - m_i)^2]} . \tag{3}$$

Whether $\text{Var}(s_i)$ is the model or randomization variance of s_i depends on whether one's goal is to minimize the model or design variance of $t_i^{(g)}$. The same holds true for the interpretation of $E[(t_i - m_i)^2]$.

From both a model and randomization-based perspective, $Var(s_i)$ and $E[(\theta_i - m_i)^2]$ are unknown. Särndal, Swensson, and Wretman's estimator for $Var_p(t_{i(rb)}) \approx Var_p(s_i)$ is also a reasonable estimator for the model variance of s_i. A reasonable estimator for $E_p[(\theta_i - m_i)^2]$ is illusive, but a good estimator for $E_M[(\theta_i - m_i)^2] \approx Var(\epsilon_i)$ is not difficult to develop.

Suppose one estimates $Var_M(s_i)$ and $Var(\epsilon_i)$ from the sample, plugs those estimates into equation (3), and then computes $t_i^{(s)}$. Call the result t_i^*. As the sample size in domain i increases, $Var_M(s_i)$ decreases, while $Var(\epsilon_i)$ remains a positive constant. Thus, as n(i) grows arbitrarily large t_i^* converges to $t_{i(rb)}$, making it randomization consistent just like $t_{i(rb)}$. In fact, t_i^* is fully in the spirit with Hansen, Madow, and Tepping's Guiding Principle No. 4: models have been used in the choice of the estimator, but the estimator itself, while biased, is randomization consistent.

Let us now turn to the primary question addressed in the Malec & Sedransk and Fay papers: how should the variance of a small domain estimator like t_i^* be estimated? Both papers take a model-dependent approach. The problem with this approach, of course, is that models can fail. Since Fay's paper deals with simulations, he avoids the problem. Malec & Sedransk do not.

Malec & Sedransk are to be commended for their thoughtful and thorough work in developing complex models at both the element and domain levels that are appropriate for the survey data they are examining. I have absolutely no problems with the determined parts of these models. What bothers me are the random parts. In particular, the authors build in random effects at the county level only. They allow no additional clustering effects within area segments or households. Moreover, they assume county effects are uncorrelated both across adjacent counties and within states. An example of counties in a state likely to be correlated are Kings, Queens, New York, and Bronx Counties -- the four big boroughs of New York City. I suspect that more than one of these counties are represented in the authors' sample.

It should be noted that the goal of the Malec & Sedransk paper is to produce state not county estimators. Their domain-level model is on the county level, however. Thus, they estimate $\theta_{(state)} = \Sigma_{i \in state}\, \theta_i$ with $\Sigma_{i \in state}\, t_{i(MS)}$, where $t_{i(MS)} = m_i$ for counties not represented in the sample. For counties represented in the sample, $t_{i(MS)}$ is similar to the t_i^* discussed above. Nevertheless, because of how the other counties are handled, there is no easy way of modifying a Malec/Sedransk state estimator to make it randomization consistent.

If g were determined from an outside source, the model variance of $t_i^{(g)}$ would be

$$Var_M(t_i^{(g)}) \approx (1-g)^2 Var_M(s_i) + g^2 Var(\epsilon_i). \qquad (4)$$

Once estimators for $Var_M(s_i)$ and $Var(\epsilon_i)$ are computed, an estimator for $Var_M(t_i^{(g)})$ quickly presents itself.

When a g (approximately) satisfying equation (3) is determined from the sample so that $t_i^{(g)} = t_i^*$, it is tempting to simply plug that value into equation (4) along with estimates of $Var_M(s_i)$ and $Var(\epsilon_i)$. A good deal of high powered statistical work has gone into showing why such a practice can be mistaken. I have a more prosaic problem with this approach to variance estimation: it relies entirely on the truth of the model; in particular, on the model for the ϵ_i. It is true that we modeled the ϵ_i in developing the estimator t_i^* in the first place, but to my mind this fact only reinforces a need to be able to evaluate the accuracy of t_i^* in a way that does not require the same model assumptions.

The randomization mean squared error of $t_i^{(g)}$ is

$$MSE_p(t_i^{(g)}) \approx (1-g)^2 Var_p(s_i) + g^2 E_p[(\theta_i - m_i)^2].$$

Let $v(s_i)$ be a randomization-based estimator for $Var_p(s_i)$. One can estimate $E_p[(\theta_i - m_i)^2]$ with $(t_{i(rb)} - m_i)^2 - v(s_i)$. Unfortunately, this estimator is dreadfully unstable. It has, at most, 1 degree of freedom. For many domains, $v(s_i)$ will also be very unstable, since it has, at most, $n(i) - 1$ degrees of freedom.

It may come as a shock, but few users of our statistics are all that concerned with variances. With this is mind, perhaps we should abandon the search for a near perfect variance estimator for t_i^*. We do need to be assured that t_i^* has some minimum degree of accuracy. One possibility is to model $v_p(s_i)$ and $(t_{i(rb)} - m_i)^2 - v(s_i)$ across all the domains and to use the results to derive a conservative indication about the accuracy of t_i^* for each i.

REFERENCES

Hansen, M., Madow, W. and Tepping, B. (1983), "An Evaluation of Model-Dependent and Probability-Sampling Inference in Sample Surveys (with discussion)," Journal of the American Statistical Association, 776-807.

Särndal, C. Swensson, B. and Wretman, J. (1992), Model Assisted Survey Sampling. New York: Springer-Verlag.

DISCUSSION OF SMALL AREA ESTIMATION PAPERS
COPAFS CONFERENCE, MAY 26, 1994

David A. Marker
Westat, Inc.

Both of these papers are important for their general approach to the problem of small area estimation: they attempt to understand the application of new methods through the explicit use of models. Ideally, one would always design surveys to allow for the production of accurate, direct, design-based estimates. However, when such estimates cannot be produced, one is left with only two choices: either don't produce estimates or use models.

Malec and Sedransk present the use of hierarchical Bayes procedures for small area estimation. I find this approach to be more satisfactory than empirical Bayes procedures for at least three reasons. First, hierarchical Bayes procedures do not assume a particular model to be true. To quote George Box, "All models are wrong, but some are useful." Second, by assuming instead that the truth comes from within a class of prior distributions, it is possible to examine the robustness of the estimates; although this is limited to the range of priors contained in the class. Third, hierarchical Bayes allows for the use of informative priors. While Malec and Sedransk do not make use of informative priors, this is a possible area for extending their results. Particularly for repeated surveys such as the NHIS, there is a wealth of historical data that can be used. These data can be incorporated for model selection, as variables in the actual model, or to construct informative priors.

Many authors, including Malec and Sedransk, use the Gibbs sampler to produce hierarchical Bayes estimates. The advantage of the Gibbs sampler is that it allows for computations from complex distributions. However, the experiences relayed by Malec and Sedransk and others indicate that this approach is extremely time intensive, in some cases taking months to produce stable estimates. This raises questions about the actual utility of this approach to produce timely small area estimates.

One additional point is worth making regarding the Gibbs sampler. As mentioned earlier there is an abundance of historical NHIS data from which informative priors could be developed. It would be very interesting to see the results of using the Gibbs sampler when beginning with informative, rather than uninformative, priors.

Malec and Sedransk develop their model using forward stepwise regression. While this is a reasonable approach, it can lead to suboptimal results under complex situations. Therefore, it might be worthwhile to examine alternative model-selection methods. In selecting their model, they disregarded the sampling weights. They reported that the weights would not have had significant impact based on analyses at the national level. My concern is that given state-to-state differences this might not imply that nothing is lost by disregarding weights when producing state estimates.

Fay uses simulation to examine the real situation of computing the accuracy of small area estimates when the variances are unknown. The Prasad and Rao approach that he evaluates is limited to situations in which the mean and variance are independent. Unfortunately in many situations, including the binomial variable used by Malec and Sedransk, this is not true. Prasad and Rao developed a procedure for producing approximately unbiased mean square errors (MSEs) for model-dependent small area estimates. These MSE estimates are, however, conditional on the model.

For government agencies there is a strong interest in producing design-based measures of accuracy, not ones conditional on models. A method for producing design-based small area specific MSEs was introduced by Marker (1993). This approach replaces the average MSE of Gonzalez and Waksberg with a small area specific MSE, where the variance of the model-dependent estimator is computed for each small area i using replicated methods (jackknife or balanced repeated replication). The bias is computed by averaging across small areas.

$$MSE\,(y_i) \quad = \quad var\,(y_i) \;+\; avebias^2\,(y_i)$$

where

$$avebias^2\,(y_i) \quad = \quad aveMSE\,(y_i) \;-\; avevar\,(y_i)$$

This estimator is not completely small area specific, but if the variance term dominates the bias, the root mean square error will provide a useful substitute for the traditional standard error. If the bias term dominates, the small areas can be grouped by expected similar biases. The average bias can then be computed separately for each group

of areas so that the MSE more accurately reflects small area differences. It would be very useful if both Fay and Malec and Sedransk could examine the utility of this approach.

Reference:

Marker, David A., "*Small Area Estimation for the National Health Interview Survey*," Proceedings of the American Statistical Association Section on Survey Research Methods, 1993.

Session 12

NONRESPONSE IN SURVEYS

EXPLORING NONRESPONSE IN U.S.FEDERAL SURVEYS

Maria Gonzalez, OMB; Dan Kasprzyk, NCES; and
Fritz Scheuren, IRS

Section 1: Introduction

This paper is intended to provide a broad summary of nonresponse rate trends in U.S. federal government surveys. We have built directly on the work of a Subcommittee on Survey Nonresponse, commissioned in 1991, by the Office of Management and Budget's Federal Committee on Statistical Methodology (FCSM). A particular debt of gratitude needs to be acknowledged for the role played by Bob Groves (Subcommittee Chair), Mick Couper and the other members of that Subcommittee for their input into what follows (see acknowledgements for a full list of the members).

Highlights of the Subcommittee's efforts have already appeared in the April AMSTAT NEWS (Gonzalez, Kasprzyk, and Scheuren,1994). A more extended treatment will be given in this paper. Still other papers based on the Subcommittee's work will appear in the Proceedings of the 1994 meetings of the ASA.

The present material is organized into four main sections, along with supporting figures, references, acknowledgements and an afterword. First, there is this short Introduction (Section 1); some background considerations come next. These considerations led to the establishment of the FCSM Nonresponse Subcommittee (Section 2).

In Section 3, an overview of the work of the Subcommittee is given, including the principal findings on nonresponse rate trends in federal surveys. Naturally, a discussion is given of limitations as well.

Finally, the recommendations of the Subcommittee are revisited in Section 4 and comments made on the future steps

that we, as federal government statisticians, should take -- both as individuals working on our own surveys and by acting collectively to improve practice as a whole.

Section 2: Some Background on Nonresponse in Federal Surveys

Like the poor, nonresponse in surveys may always be with us. In the days of "representative" samples drawn purposively, nonresponse was present but not visible. (Quota sampling, even today, makes measuring the extent of the actual nonresponse difficult -- maybe impossible). With the ascent of the random sampling paradigm (Bellhouse,1988), nonresponse became a problem that needed to be "solved."

In so far as U.S. Federal surveys are concerned, the turning point in government practice for the randomization paradigm came when Deming invited Neyman to lecture at the U.S.D.A. Graduate School in 1937. Morris Hansen, using Neyman's ideas and his own, and with many collaborators, did the rest (e.g., Hansen, Hurwitz, and Madow,1953).

It seems clear that Hansen and the other early pioneers understood quite well that randomization-based inference was directly challenged by nonresponse. Concerns about bias, for example, were evident from the beginning. In Cochran(1977) there is an example of an early treatment that simply widens the confidence intervals directly to account for the nonresponse bias. This conservative approach was consistent with the main focus of the random samplers of that era who were busy inventing ways to reduce nonresponse to the bare minimum. The U.S.Census Bureau in its Current Population Survey (e.g., Hanson,1978) still continues successfully in that tradition.

Hansen and his collaborators, in addition to a primary emphasis on "prevention," developed designs which called for the subsampling of nonrespondents (e.g., Hansen and

Hurwitz,1946). These were a natural extension of the basic randomization paradigm and called for more thorough fieldwork on a random subsample of nonrespondents. One of the results of this work was to introduce the idea of a weighted response rate. Such samples naturally also had their own nonresponse problems; so this approach too was seen from the beginning as only a partial one. Post-survey adjustment techniques to compensate for flaws in the randomization due to nonresponse were also attacked as well.

For those interested in more information, a special September 1975 issue of the Journal of the American Statistical Association is a recommended reference (Gonzalez, Ogus, Shapiro, and Tepping, 1975). This article provides a useful summary of federal goverment (largely Census Bureau) practices on the reporting of sampling and nonsampling errors, including nonresponse (see also Duncan and Shelton (1978) for still more on the history of sampling in U.S. Federal surveys).

While nonresponse in federal surveys has always been said to be an indicator of the quality of survey data, interest and concern has grown during the last two decades:

- The Panel on Incomplete Data, established by the Committee on National Statistics in 1977, produced three volumes focussing on incomplete data in sample surveys (Madow, Nisselson, Olkin, and Rubin, 1983).

- The Council of American Survey Organizations (CASRO) reviewed response rate definitions with the intent of trying to establish uniformity of definitions across surveys (CASRO, 1982).

- Steeh (1981) and Groves (1989) reviewed trends in the response rates in nongovernment surveys, indicating a decline in response rates over time.

- During the last ten years, the tight federal budget climate has prompted questions about the ability of federal statistical agencies to maintain high response rates with a constant budget.

Theoretical developments in the handling of nonresponse have grown enormously since the mid-1970's. Indeed, the problem has drawn the attention of some of the best statisticians now working on surveys. The National Academy Panel's report on Incomplete Data(1983) was a culmination of sorts. A review of nonresponse adjustment techniques was done by Kalton (1983). Even so, in the ten years since the Panel's report, there has been a lot more done and no end is in sight. The book on nonresponse by Little and Rubin (1986) and a separate book by Rubin (1987) on multiple imputation are perhaps the two most prominent examples of the important work that continues. The treatment of Sarndal et al (1992) and Lessler and Kalsbeek (1992) also are valuable for the way, among other things, they place nonresponse in context of total survey error.

Within this general environment of greater interest in nonresponse, the FCSM decided to sponsor an effort to learn what was known about nonresponse as a source of bias in federal survey estimates. Prominent factors in making this decision were --

- The lack of a systematic review of the topic since the 1983 Committee on National Statistics report.

- A growing perception among the members of the federal statistical community that nonresponse in federal surveys had been increasing over time.

In any event, in 1991 a Subcommittee of the FCSM was formed to study nonresponse in federal surveys. The initial charge of the Subcommittee was, simply stated, to "begin an effort to better understand unit nonresponse in surveys." The proposed approach was to conduct a broad-based review of the level of unit nonresponse rates, currently and over time, in federal surveys. The details of the Subcommittee's work are covered in the next Section.

Section 3: Work of FCSM Subcommittee on Survey Nonresponse

The Subcommittee was specifically charged with the mission to investigate for Federal surveys the levels of response rates, the measures used to compute these response rates, response trends from 1982-1991, perceived correlates of nonresponse, and other related information.

In carrying out its mission, the Subcommittee obtained information from 26 demographic and 21 establishment surveys. These surveys were not selected by probability methods, because no machine-readable listing of Federal surveys with sufficient auxiliary information for appropriate stratification was available. The 47 surveys were chosen, however, to include Federal surveys that differed on a number of key design parameters: those conducted on an ongoing or an intermittent basis, those conducted by Federal agencies, and those carried out by contractors under Federal auspices.

Because of the large differences in the design of surveys to collect establishment versus demographic data, separate questionnaires were constructed for each type and sent to respective survey sponsor or data collection agency. The intent of both questionnaires was to elicit information on a

variety of survey features that earlier literature has shown to affect nonresponse. In addition, information was sought on strategies for post-survey adjustment for nonresponse.

The Study itself incurred no unit nonresponse but did incur a small amount of item nonresponse in its data collection activities. Indeed, it was difficult to get the agencies to respond to the nonresponse questionnaire.

The findings of the Subcommittee span the range from the expected to the surprising. As in any research undertaking, of course, the conclusions drawn from an analysis of the questionnaires should be treated with caution. This point is particularly well taken here given the purposive nature of the sample, the small number of surveys included in the data collection, and the wide variety of design differences that characterize these surveys. Some highlights follow.

Trends in Nonresponse Rates. Despite the prior beliefs of many in the Federal survey community, there was little evidence of declining response rates over time for either the establishment or demographic surveys included in the Subcommittee's study:

- Establishment Surveys. To analyze the response rates of establishment surveys over time, it is more meaningful to limit the analysis to those surveys which reported response rates for several years. For this reason, the analysis of time trends in the response rates for establishment surveys cover only the nine surveys for which both weighted and unweighted data were available for at least six reporting periods between 1981-1991.

Figure 1 shows the average weighted response rate for the nine selected surveys. Figures 1-5 are based on work of the FCSM Subcommittee on Survey Nonresponse. As may be seen, the weighted response rate was only slightly decreasing over the period covered by the data. The average decrease was about 1/4 percent per year. Figure 1 also shows the mean unweighted response rate for the selected nine surveys from 1984-1990. The unweighted rate was slightly increasing, but stable over the period. The average increase was about 1/2 percent per year.

Figure 2 shows weighted response rates for the nine establishment surveys. Five of these weighted response rates are 90 percent or above. Two series have a weighted response rate between 70-90 percent and two series are around 50 percent. More about establishment trends in response will be said in Osmint, McMahon, and Martin.

- Demographic Surveys. Most demographic surveys used unweighted response rates rather than weighted rates for routine monitoring of the data collection process and so we have followed this convention here as well. The analysis of trends over time for demographic surveys was restricted to those surveys with at least 4 data points in the period 1982 to 1991. Only 8 of the 26 demographic surveys included in our data collection met this criterion.

The mean nonresponse rate by year was calculated for these eight surveys from the data provided, along with refusal rates and noncontact rates where available. Although the stimulus for the creation of the Subcommittee was the belief that response rates were

declining over time in demographic surveys, Figure 3 does not support that belief. The mean nonresponse rates for the surveys included in the sample are minimally lower in 1991 than in 1982. This figure shows that refusal rates, a major component of nonresponse, have remained about the same. More about demographic trends in response will be said in Johnson, Botman, and Basiotis.

For the Current Population Survey a longer time series of data is available. Figure 4 shows that the level of nonresponse has been stable for some years. Since the refusal rates seem to have increased, a possible implication is that more effort may have been made to reduce other nonresponse components -- so as to achieve relatively constant overall response rates.

An examination of response rates for the more-frequently fielded demographic surveys reveals large variations across surveys. This variation can be partially understood by separating the studies into two groups (see Figures 5). One group has response rates in the 95 percent range, while a second cluster lies about 10 percentage points lower. The studies in the 95 percent range consist of ongoing studies conducted by the same interviewer corps. The studies in the lower group tend to be less frequently conducted. Neither group exhibits strong trends over time.

In summary, despite the prior beliefs of many in the survey community, there was little evidence of declining response rates over time among either the establishment or demographic surveys included in the study. This could be due to a greater effort in data collection but technological and other survey context changes make this hard to verify. One

final point about these results may be worth making again: There were only a limited set of surveys on which time trends can be measured--just nine establishment surveys and eight demographic surveys.

Other Findings.-- There are other findings from the Subcommittee's work; but only three are highlighted here. These involved issues in the definition of nonresponse, response rate documentation, and post-survey adjustment methods:

- Definitions for Nonresponse. Despite the study's focus on nonresponse rates and despite having contacts in the agencies, major difficulties arose in obtaining consistent information. Just as was found in an early (albeit more general) study,..."rates have different names and different definitions in different places and times." (Bailar and Lanphier, 1978) This issue led, in part, to one of the study's major recommendations (see figure 6, Subcommittee Recommendation 3).

- Response Rate Documentation. Reporting practices for documenting response rate components varied widely across the surveys in the study. Common practice in establishment surveys is in contrast to common practice in demographic surveys. Sponsors of demographic surveys not only were more likely to maintain records regarding a wider variety of nonresponse components but also tended to maintain more historical information. For example, all of the demographic surveys in our data collection included some information about response/nonresponse components. In contrast, for the establishment surveys analyzed, 10 out of 21 did not track any nonresponse components.

- Post-survey Nonresponse Adjustment. Respondents were asked about a number of post-survey adjustment techniques designed to reduce the effects of nonresponse: post-stratification (e.g., simple ratio or raking ratio adjustment), regression modelling of the propensity to respond, and imputation. All surveys in the Subcommittee study used some degree of post-survey nonresponse adjustment. Some of the approaches were very traditional, while others reflected more recent research on estimation strategies.

In the remaining section of this paper we cover the Subcommittee's recommendations and a few ideas on future steps.

Section 4: Some Next Steps for Practice

The Subcommittee made four recommendations that are given in detail in figure 6. Stated briefly the subcommittee recommended:

- Survey practitioners should compute nonresponse rates in a uniform fashion over time.

- In repeated surveys, response rate components should be monitored in conjunction with cost and design changes.

- Agencies that sponsor surveys should publish how they compute response rate and their components in survey reports and their relevance to the quality of the survey results discussed.

- Ongoing research should be conducted on nonresponse adjustment variables, costs and benefits of converting refusals, and similar nonresponse management concerns.

All of these recommendations seem rather obvious. They address some very basic survey management and reporting requirements; furthermore, the suggestions are close in spirit and substance to those made by other groups over the last two decades.

It is true that every survey program examined by the Subcommittee calculated nonresponse rates in some fashion and had some auxiliary information about aspects of nonresponse. It is also true that most survey programs did not have **readily** available what the Subcommittee viewed as "basic" data on nonresponse; nor did repeated surveys have a time series easily available of nonresponse rates and nonresponse components.

What can we expect for the future based on the results of this small exploratory study? Some conjectures follow:

- First, it is unrealistic to assume that the recommendations by yet another subcommittee will be adopted uniformly by the agencies of the Federal statistical system.
- Second, unless mandated, individual survey program managers are likely to remain individualistic and independent with respect to their acceptance and adoption of recommendations concerning their surveys.
- Third, it is important to recognize the diversity of the management of individual survey programs and build on each survey programs' strengths. In other words, these recommendations should be no more than guidelines in any case.
- Fourth, the survey data collection manager and the agency that sponsors the survey need to work together as a team with the interests of the ultimate customer paramount -- recognizing that an information system producing data on nonresponse and its components is mutually beneficial.

One need not assume the points made above are necessarily pessimistic. The underlying theme is the development of a fully professional partnership among data collection managers, agencies that sponsor survey programs, and ultimate customers. Mutual respect and understanding for each other's requirements (given budget constraints) is essential for improving the reporting of nonresponse and nonresponse components. The current theme of "reinventing government" speaks well to the prospect of improving these professional relationships through its team building and customer orientation emphases.

Finally, we can expect incremental improvements in the issues discussed here through the continuing work of the Federal Committee on Statistical Methodology (Gonzalez, 1994) and the National Science Foundation initiated "Program in Survey Methodology" offered by the consortium of the University of Maryland, University of Michigan, and Westat. Both programs are dedicated to the improvement of the quality of Federal survey data. Through these efforts and the individuals involved in Federal data collection programs, progress will be made.

Most of this paper looks inward at the federal statistical system. Obviously, much can be learned by examining private sector experiences and through international comparisons. The companion paper at this session by David Binder and his colleagues from Statistics Canada is an example of what we have in mind. Clearly, too, the Statistics Canada approach to nonresponse rates is worthy of further study by those interested in this area (see Statistics Canada, 1993 and Hidiroglou, Drew, and Gray, 1993)

In the spirit of "reinvention," a systematic benchmarking approach is needed. Some important beginning efforts that bear mention in this regard include the papers by Lyberg and Dean (1992) and Christianson and Tortora (1993).

Acknowledgements and Afterwords

This paper is based on the work of the Subcommittee on Survey Nonresponse whose members included: Robert M. Groves, Chair, BOC, Susan Ahmed, NCES, J. Donald Allen, NASS, David Belli, BEA, Peter Basiotis, HNIS, Steve Botman, NCHS, Eileen Collins, NSF, Mick Couper, BOC, Patricia Guenther, HNIS, Paul Hsen, BLS, Ayah Johnson, AHCPR, Arthur Kennickell, FRB, Paul McMahon, IRS, Jeffrey Osmint, BoM, Antoinette Martin, EIA, Pamela Powell-Hill, BOC, Maria Reed, BOC, Carolyn Shettle, NSF.

One final point may be made about the exploratory nonresponse Study focussed on in this paper. We would like to engage in a dialogue about this study and its recommendations with those who conduct both Federal and non-Federal surveys. Please contact Maria E. Gonzalez (OMB, Statistical Policy Office, NEOB, Room 10201, Washington, D.C. 20503) with your comments or experiences in this field.

References

Bailar, B.A. and Lanphier, C.M. (1978). Development of Survey Methods to Assess Survey Practices. Washington: American Statistical Association.

Bellhouse, David (1988). "A Brief History of Random Sampling Methods," Handbook of Statistics, Vol 6, North-Holland, Amsterdam, pp. 1-14.

Christianson, Anders and Robert D. Tortora (In Press). "Issues in Surveying Establishments: An International Survey," in BG Cox et al (Editors), Survey Methods for Businesses, Farms, and Institutions, John Wiley & Sons. Presented at the International Conference on Establishment Surveys, 1993 Buffalo, NY.

Council of American Survey Organizations (CASRO) (1982). On the Definitions of Response Rates, Port Jefferson, New York.

Duncan, J.W. and Shelton, W.C. (1978). Revolution in United States Government Statistics 1926-1976, U.S. Department of Commerce, Office of Federal Statistical Policy and Standards.

Gonzalez, M.E. (1994). "Improving Data Awareness in the United States Statistical Agencies," The American Statistician. American Statistical Association. Vol. 48, No. 1.

Gonzalez, M.E., Kasprzyk, D., and Scheuren, F. (1994). Nonresponse in Federal Surveys, AMSTAT News, April.

Gonzalez, M.E., Ogus, J.L., Shapiro, G., and Tepping, B.J. (1975). Standards for the Discussion and Presentation of Errors in Census and Survey Data, Journal of the American Statistical Association, 70, 351, Part II.

Groves, Robert M. (1989). Survey Errors and Survey Costs, John Wiley and Sons, New York.

Hansen, M., Hurwitz, W., and Madow, W. (1953). Sample Survey Methods and Theory (2 Volumes), John Wiley: New York.

Hansen, M. and Hurwitz, W. (1943). On the Theory of sampling from finite populations. Ann. Math. Statist. 14, 333-362.

Hanson, R. (1978). The Current Population Survey--Design and Methodology, Technical Paper 40, U.S. Bureau of the Census.

Hidiroglou, M.A., Drew, J.D., and G.B. Gray (1993). "A Framework for Measuring and Reducing Nonresponse in Surveys," Survey Methodology, Vol. 19, No. 1, pp. 81-94.

Johnson, Ayah, Steve Botman, and Peter Basiotis (forthcoming). "Nonresponse in Federal Demographic Surveys: 1981-1991." To be presented at the 1994 annual meetings of the American Statistical Association, Section on Survey Research Methods.

Kalton, G. (1983). Compensating for Missing Survey Data, Research Report Series, Ann Arbor, Michigan: Institute for Survey Research.

Lessler, J. and Kalsbeek, W. (1992). Nonsampling Errors in Surveys, John Wiley & Sons: New York.

Little, R. and Rubin, D. (1986). Statistical Analysis and Missing Data, John Wiley & Sons: New York.

Lyberg, L. and Dean P. (1992). "Methods for Reducing Nonresponse Rates: A Review," presented at the annual meeting of the American Association for Public Opinion Research, St. Petersburg, FL.

Madow, W.G., Nisselson, H., Olkin I., and Rubin D.B. (Editors) (1983). Incomplete Data in Sample Surveys, 3 Volumes, Academic Press, NY.

Osmint, Jeffrey, Paul B. McMahon, and Antoinette Ware Martin (forthcoming). "Response in Federally Sponsored Establishment Surveys." To be presented at the 1994 annual meetings of the American Statistical Association, Section on Survey Research Methods.

Rubin, D. (1987). Multiple Imputation for Nonresponse in Surveys, John Wiley & Sons: New York.

Sarndal, C.E., Swenssen, B., and Wretman, J. (1992). Model Assisted Survey Sampling, Springer-Verlag: New York.

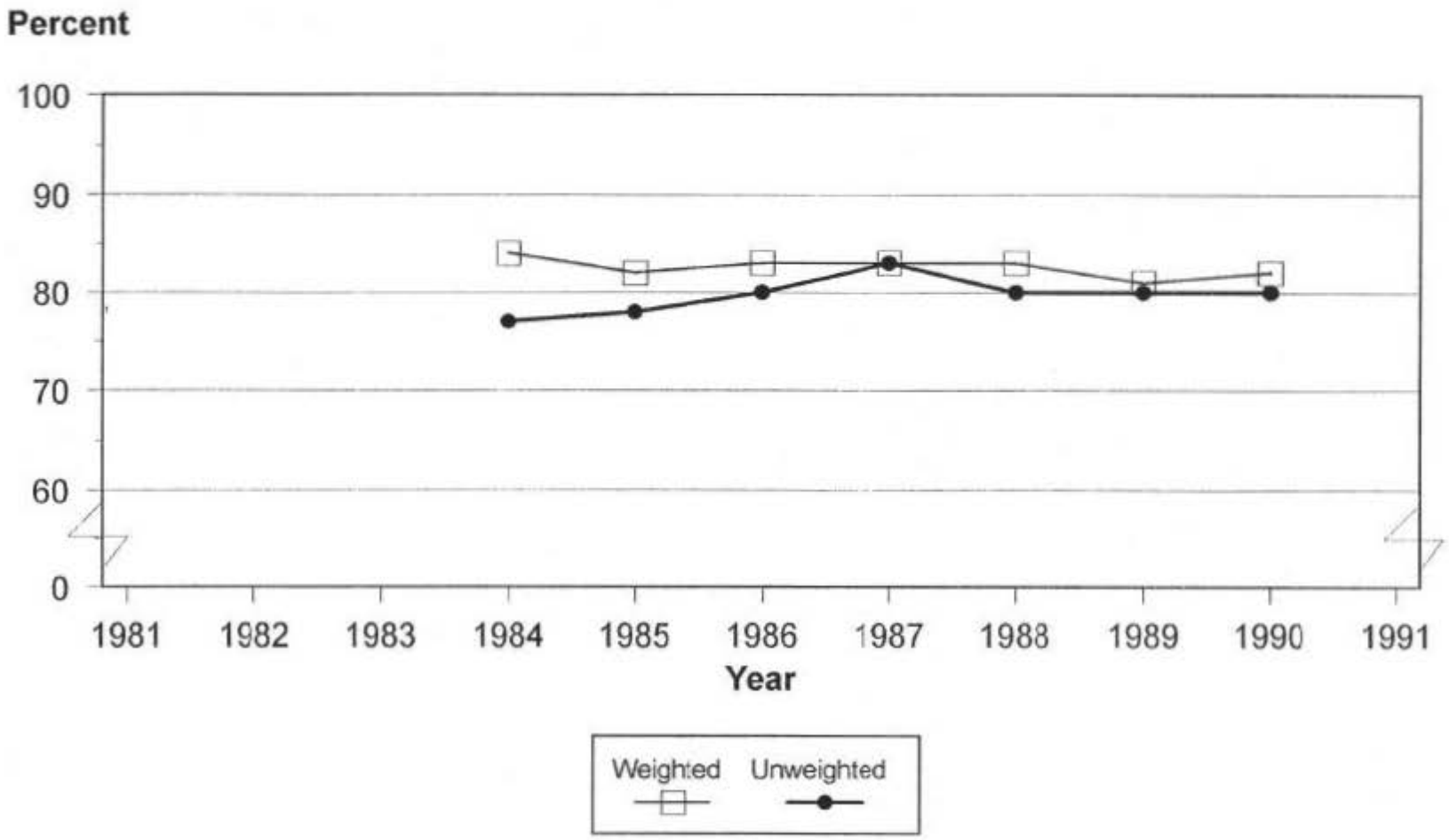
Shettle, Carolyn F., Patricia M. Guenther, Daniel Kasprzyk, and Maria Elena Gonzalez (forthcoming). "Investigating Nonresponse in Federal Surveys." To be presented at the 1994 annual meetings of the American Statistical Association, Section on Survey Research Methods.

Statistics Canada (1993). Standards and Guidelines for Reporting of Nonresponse Rates, Ottawa, Canada.

Steeh, C. (1981). "Trends in Nonresponse Rates," Public Opinion Quarterly, Vol. 45, pp.40-57.

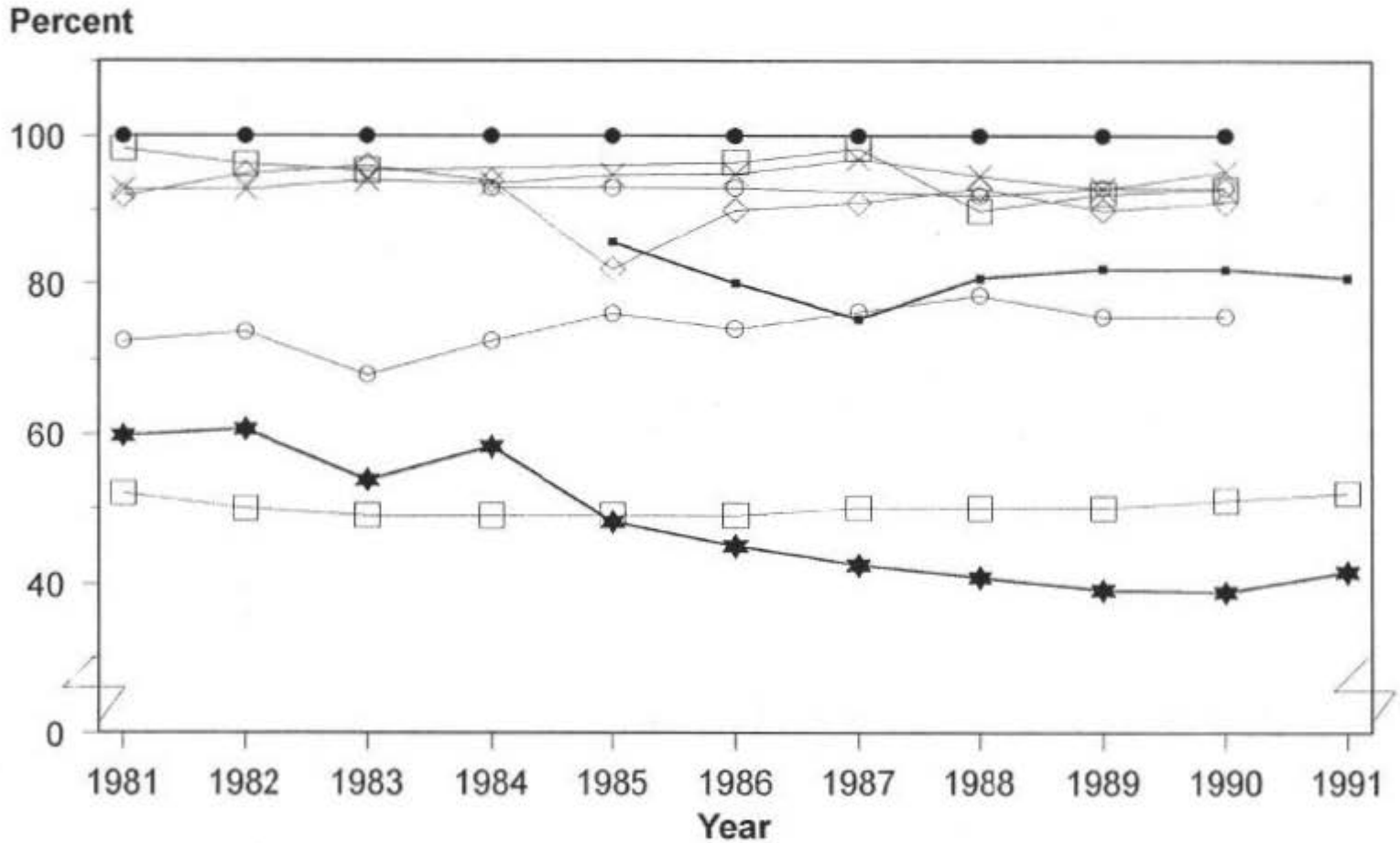
Figure 1.

Mean Response Rates, Economic Surveys



SOURCE: Federal Committee on Statistical Methodology, Subcommittee on Survey Nonresponse.

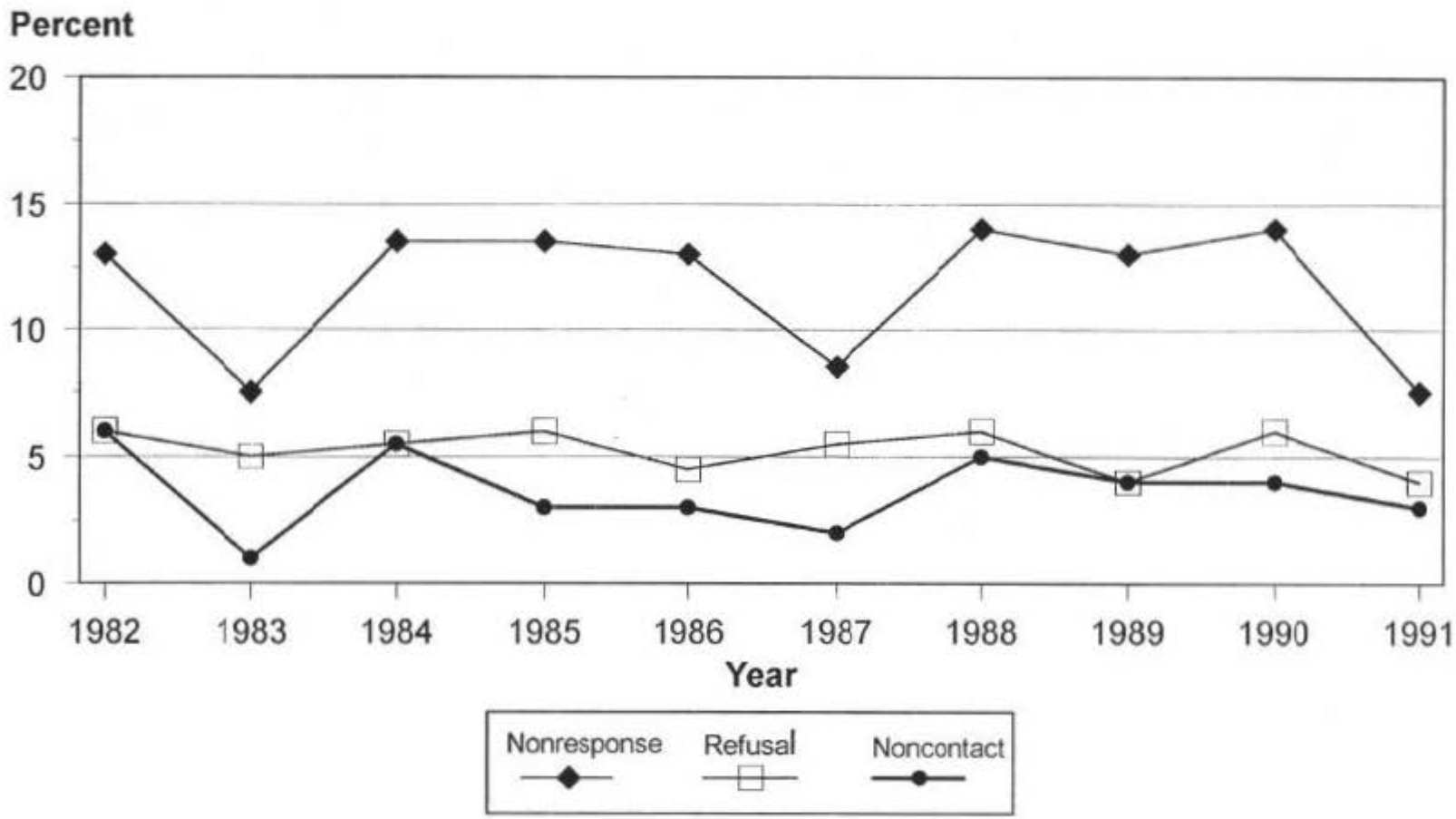
Figure 2.
Weighted Response Rates, Economic Surveys



SOURCE: Federal Committee on Statistical Methodology, Subcommittee on Survey Nonresponse.

Figure 3.

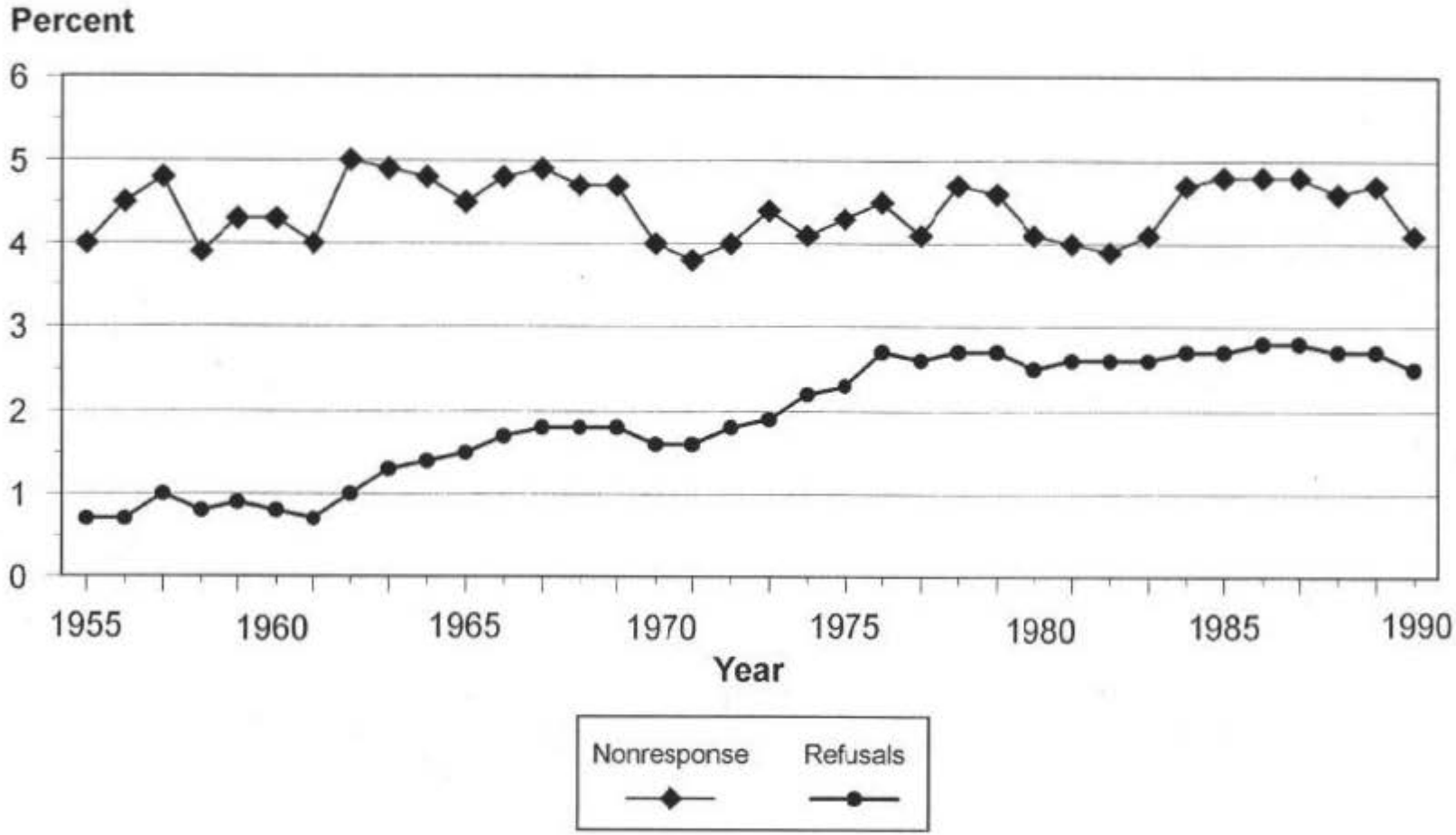
Mean Unweighted Nonresponse Rates, Demographic Surveys



SOURCE: Federal Committee on Statistical Methodology, Subcommittee on Survey Nonresponse.

Figure 4.

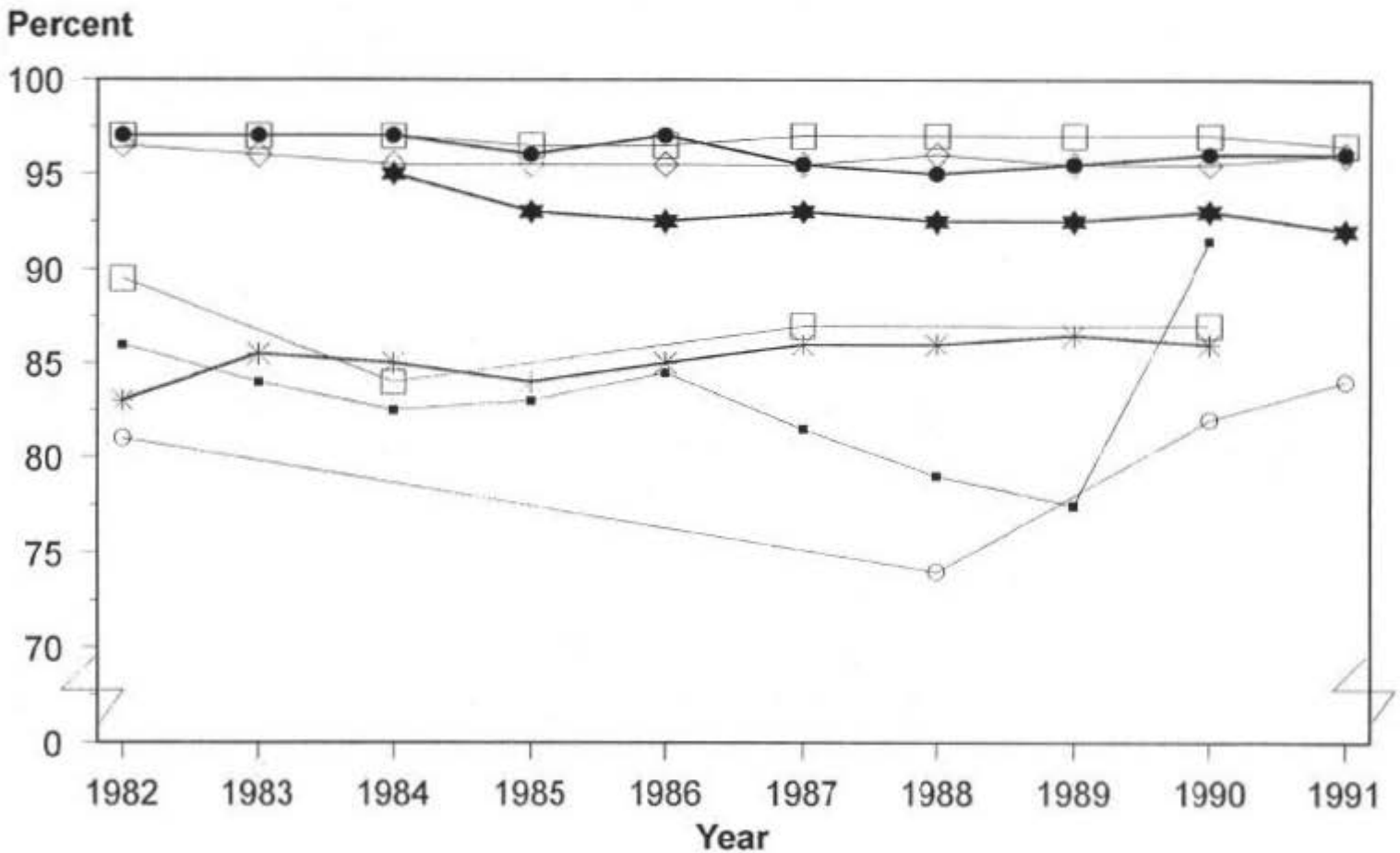
Unweighted Nonresponse Rates, Current Population Survey



SOURCE: Federal Committee on Statistical Methodology, Subcommittee on Survey Nonresponse.

Figure 5.

Unweighted Response Rates, Demographic Surveys



SOURCE: Federal Committee on Statistical Methodology, Subcommittee on Survey Nonresponse.

Figure 6.--Summary of FCSM Recommendations on Survey Nonresponse

Recommendation 1. **Survey staffs should compute response rates in a uniform fashion over time and document response rate components on each edition of a survey.**

The subcommittee chose not to recommend that every survey use the same response rate computations. Other groups have recommended such uniformity (see CASRO, 1982). In the Subcommittee's view, every definition of response rate components offers some useful information. Some response rate definitions inform the designers about the rate of success of measurement of the average sample unit; others focus on different causes of nonresponse. One can distinguish between measures useful as management tools and measures that should be reported to data users so that they can assess the quality of the survey data.

Recommendation 2. **Survey staffs for repeated surveys should monitor response rate components (e.g., refusals, not-at-homes, out-of-scopes, address not locatable, postmaster returns, etc.) over time, in conjunction with routine documentation of cost and design changes.**

The Subcommittee believes that response rate components are useful tools to monitor changes in the quality of survey statistics. Response rates should be easily accessible and timely. By themselves, they are not error measures; however, for repeated surveys, changes in response rate components may signal the need for supplementary study of nonresponse error properties. Such changes can alert the survey designers to changes in the "survey-taking climate" that affect completion of measurement, point to changes in the administrative controls over response rates that may need adjustment, and help measure the effects of any design changes made.

For ongoing surveys, graphs of time series of response rate components, juxtaposed with costs for each collection cycle, and indicators of design changes introduced in that cycle, can be valuable management tools. Survey managers need better tools to diagnose the causes of cost changes in data collection activities. Falling response rates, especially those associated with cases requiring much effort prior to the ultimate nonresponse, magnify cost pressures on surveys. The subcommittee's study did not collect data on survey costs, because comparable cost information across surveys was not believed to be available.

Recommendation 3. **Agencies that sponsor surveys should be empowered to report the response rates of their surveys. The sponsoring agency should explain how response rates are computed for each survey it sponsors. Response rates for any one survey should be reported using the same measures over time, so that**

users may compare the response rates. Response rate components should also be published in survey reports.

An agency that sponsors surveys should compute and explain in its survey publications the response rates for each of the surveys it sponsors. Surveys sponsored over time should report the same measure of response for all data collection periods so that users can compare these measures over time. The actual method used to compute response rates should be described in all publications issued.

The results of recommendations 1 and 2 should be shared routinely with the users of survey data, along with discussions of the relevance of response rates to evaluating the quality of the survey data. An analysis of the characteristics of the nonrespondents should be implemented routinely as part of each cycle of data collection.

Recommendation 4. **Some research on nonresponse can have real payoffs. It should be encouraged by survey administrators as a way to improve the effectiveness of data collection operations.** The Subcommittee believes that areas of research most likely to yield payoffs include:

- Studies of the relative costs of final efforts to raise response rates, through persuasion, repeated callbacks, and other measures. When these costs are compared to number of cases added to the respondent pool, the relative cost per case can be computed. Studies of the effects of these final cases can be made in an effort to assess the cost effectiveness in terms of mean square error of the final efforts.

- Studies of the measurement error properties of information provided by the reluctant respondent cases, relative to the nonresponse bias in statistics that would omit them from computations. This would address a key question in survey design: When data collectors exert great effort to persuade the reluctant to respond, is one type of error, nonresponse, merely exchanged for another type, measurement error? Perhaps, those persuaded to respond may exert less effort at providing accurate data?

- Studies on what variables should be collected to improve post-survey adjustment for unit nonresponse (see Madow et al, 1983: Recommendation 10(2)). When observable or inferred characteristics of nonrespondent units are related to the survey variables and to the likelihood of participation, then collecting and using these variables in post-survey adjustment models might be a cost effective method of reducing overall mean square errors.

MODEL-BASED REWEIGHTING FOR NONRESPONSE ADJUSTMENT

David A. Binder, Sylvie Michaud and Claude Poirier
Statistics Canada, Ottawa, Ontario, Canada K1A 0T6

ABSTRACT

Nonresponse in surveys is inevitable. Much has appeared in the literature on methods of compensating for this source of nonsampling error. There is a growing interest in attempting to understand the causes of nonresponse and studying the differences in characteristics between respondents and nonrespondents. In this paper, we briefly review some related literature, discuss modelling approaches for adjusting for nonresponse and present the research findings for two surveys conducted at Statistics Canada. In both the Survey on Labour Income Dynamics and the Farm Financial Survey, we examine differences in characteristics between respondents and nonrespondents and the suitability of adopting a modelling approach for compensating for nonresponse.

KEY WORDS: Generalized regression estimators; Logistic regression; Response propensity.

1. INTRODUCTION

In virtually every survey, no matter how carefully it is designed, we must accept the fact that some data will be missing. Other than data that is missing by design, such as data from nonsampled units, data can be missing for many reasons; for example, non-contact with the respondent, refusals, late reporting, collection and processing errors, data deletion due to edit failure, undercoverage, etc.

Some measures must be taken to deal with such nonresponse. Over the years, a host of techniques has been developed. The actual choice of technique should depend on a number of factors. These include the method of estimation to be used, the amount of information about the nonrespondents that is available, the extent of other sources of error such as sampling error and response error, the relative importance of the variables to be estimated, the resources available for exploring the problem, the nature of the analyses to be performed and the statistical inferences to be made from the survey, etc.

However, even with all of these criteria, there must necessarily be some subjective judgments on the nature of the nonresponse. As we shall see, many of the methods for coping with nonresponse make use of models, either explicitly or implicitly. Even the most ardent advocates of the pure design-based school will resort to some model assumptions when it comes to adjusting for nonresponse. This presents a new set of problems associated with the statistical inferences, since the randomization distributions on which the inferences are based are no longer purely design-based, unless the nonresponse mechanism can be considered to be part of that design.

In this paper, we shall focus on the implications of the estimation method to be used and the amount of information about the nonrespondents that is available. It will be assumed that the

prime focus of the survey is to obtain estimates of descriptive statistics, such as means, totals, differences and ratios. Often nonresponse is broadly categorized into *unit-level* nonresponse and *item-level* nonresponse. This categorization is often extended, in the case of longitudinal or follow-up surveys, to *wave-level* nonresponse, where wave nonresponse is usually unit nonresponse on a particular survey occasion. In fact, it is for the case of wave nonresponse where we have the richest source of data for the nonrespondents who reported in previous waves.

Unit nonresponse is usually defined as cases where only the frame information is available for the respondents. In practice, this definition is extended to other cases where there is insufficient usable data from the respondents. The usual method for dealing with unit nonresponse is to use an "appropriate" weighting procedure to compensate for the nonresponse. (We define weighting procedures here broadly to include weight adjustments implied by regression, ratio or similar estimation techniques using auxiliary data.)

On the other hand, item nonresponse is handled either through imputation at the item level, or by ignoring the usable information and treating the respondent like unit-level nonrespondents. For wave-level nonresponse to longitudinal surveys, either reweighting or item imputation may be suitable. In this paper, we focus on the methods that use weighting techniques.

In Section 2, we discuss the basic theory underlying many of the adjustment methods and give a brief literature review. In Sections 3 and 4, we give examples of two surveys at Statistics Canada where some of these models have been studied recently. We summarize our findings in Section 5.

2. SOME GENERALITIES

2.1 Estimation

In general, we are interested in means, totals, ratios, etc. of survey variables. We denote the value of the i-th survey variable for the k-th respondent as y_{ik}. In cases where the occasion, t, is relevant, we can use y_{ikt} instead. A sample is selected according to some well-defined sampling plan. The sampling plan is usually based on frame information such as geography, and other classification and size variables. We use s to refer to the selected sample. Unfortunately, in practice, after the k-th respondent is selected, a number of things can go wrong in the process of obtaining and recording the y-values. Some of these are in the general category of response errors, where we obtain data, but they are not the y-values we were seeking. In this paper, we ignore these types of errors, except to point out that if these errors lead to large biases, the resources for nonresponse concerns may need to be trimmed in order to address the larger problem. The problem that we are addressing here is the case where the y-values are unobtainable. We denote by $s' \subseteq s$ the set of units for which we obtain usable y-values. (The subscript t is implied, where appropriate, for longitudinal surveys.)

First, we describe the estimators in the case of no nonresponse. Associated with each sampled unit, k, we have a survey weight given by

$$w_k(s) = g_k(s)\,\pi_k^{-1},$$

where π_k is $\Pr(k \in s)$, the usual first-order inclusion probability, and $g_k(s)$ is a weight adjustment that makes use of auxiliary frame data, such as poststratification, regression and ratio adjustments, etc.; see, for example, Särndal, Swensson, and Wretman (1992). We assume that the estimator of a total for a y-variable on the t-th occasion is given by

$$\hat{Y}_{it} = \sum_{k \in s_t} w_k(s_t)\, y_{ikt}. \tag{2.1}$$

Note that this estimator could be made more general, if necessary, to allow for composite estimators and multiphase samples which can depend on y-values that are observed on other occasions, but we do not introduce this complexity here. Sufficient conditions for (2.1) to be asymptotically design consistent are:

1) the probability distribution of s depends only on the auxiliary data but not directly on the y-values for the current occasion, (2.2.1)
2) the limiting expectation of $g_k(s)$ is unity, (2.2.2)
3) the variance of $\hat{Y}$ is asymptotically zero. (2.2.3)

We now consider the implications of nonresponse. Formally we assume that, given the sample, s, the set of responding units, s', follow a probability distribution $p(s'|s)$. This is completely general, allowing for correlated response patterns. It also allows for the classical case, where it is assumed that the response behaviour is nonrandom and is an inherent attribute of the selected respondents, just like the survey variables. We now consider methods of nonresponse adjustment which we refer to as *generalized reweighting methods*. Associated with each responding unit, k, we have an adjusted weight given by

$$w_k'(s', s) = g_k'(s', s)\, w_k(s),$$

where $g_k'(s', s)$ is a weight adjustment that makes use of auxiliary frame data, as well as other information that may be available for the nonresponding units. This allows the weight adjustment to depend on survey values that were observed on previous occasions from a longitudinal survey. We assume that the estimator of a total for a y-variable on the t-th occasion is given by

$$\hat{Y}_{it}^{(GR)} = \sum_{k \in s_t'} w_k'(s_t', s_t)\, y_{ikt}. \tag{2.3}$$

We let $\rho_k(s)$ be $\Pr(k \in s' | s)$. In addition to (2.2.1) to (2.2.3) above, sufficient conditions for (2.3) to be asymptotically consistent with respect to the original design and the response probabilities are:

1) the probability distribution of s' given s depends only on the auxiliary data and the survey data from previous occasions, but not directly on the y-values for the current occasion, (2.4.1)
2) the limiting expectation of $g_k'(s', s)$ is $\{E[\rho_k(s)]\}^{-1}$, (2.4.2)
3) the variance of $\hat{Y}^{(GR)}$ is asymptotically zero. (2.4.3)

If (2.4.2) is violated, then the expectation of $\hat{Y}^{(GR)}$ is

$$\sum E[g_k'(s', s)] E[\rho_k(s)] y_k. \tag{2.4.4}$$

The form of this bias is important, because if one were to impose model assumptions on the y-variables, it is possible that the model-bias becomes small. However, for those who wish to make the fewest model assumptions, it is clear that one should restrict attention to adjustment methods which yield condition (2.4.2) as closely as possible. This implies that the weight adjustment should reflect the propensity to respond as nearly as possible. Of course, the probability mechanism generating these response probabilities are generally unknown, so the weight adjustment must necessarily be model-based.

Another important feature of (2.4.2) is that if there are some "hard-core" nonrespondents -- that is, units where $\rho_k=0$ -- there would be no consistent estimates.

2.2 Examples from the Literature

The most basic form of reweighting for nonresponse that may lead to acceptable results is to simply use $g_k(s')$ instead of $g_k(s)$ in (2.1). This implies that

$$g_k'(s', s) = g_k(s') / g_k(s) .$$

This was suggested by Bethlehem (1988) for the case of the generalized regression estimators. In this case we have that the bias of the estimator is $X\beta^* - Y$, where β^* is the expected value of the estimated π-weighted regression coefficient with no nonresponse adjustment. We see then that even though this estimator is generally biased, if the regression model is reasonable, the bias can be small.

Oh and Scheuren (1983) discussed weighting class adjustment methods, which is a poststratified estimator using weighting classes as poststrata. We see that this is consistent under the assumption that the response propensities are equal within weighting classes. In practice, this technique is in widespread use; see, for example Chapman, Bailey, and Kasprzyk (1986). It was

extended to generalized regression estimators by Särndal and Swensson (1987).

One of the difficulties with weighting class adjustment methods is that there may be too many weighting classes to control. Binder and Théberge (1988) showed that with a multiplicative model for response propensities, raking ratio estimators will yield unbiased estimates. This is consistent with (2.4.2). More complex weighting schemes are proposed by Alexander (1987) and Deville, Särndal, and Sautory (1993). These could be justified under various model assumptions for the response propensities.

Many authors have proposed the use of logistic regression models to explain the nonresponse mechanism. This is a commonly used model for binary dependent variables. Examples of this can be found in Ekholm and Laaksonen (1991), Folsom (1991), and Lepkowski, Graham, and Kasprzyk (1989). In the latter paper, the logistic regression model is compared to weighting class adjustment methods, where the weighting classes are determined through some data analytic searching methods.

In Iannacchione, Milne, and Folsom (1991), after weights are included to reflect the estimated propensity to respond, the weights are fine-tuned so that certain estimates correspond to the estimate obtainable with the nonrespondents included. This is possible for wave nonresponse where certain estimates can be made for a previous wave using either the previous wave respondents or the current wave respondents. This technique should generally improve the estimates. The differences in the estimates can also be used as an diagnostic tool for the model.

Judkins and Lo (1993) and Eltinge and Yansaneh (1993) used logistic regression to model the nonresponse propensities, but then created weighting classes based on the fitted values and used weighting class adjustment methods to reweight. One of the drawbacks of the weighting class adjustment methods is that the appropriate weighting classes are not always obvious, so that such data modelling is used to help define the classes. It is expected that this method should yield results that are similar to the weights based on the logistic regression. However, if the logistic model is correct, the method will tend to introduce a small bias since (2.4.2) will be violated. In practice, though, the logistic regression model is only an approximation to the true probability mechanism.

As we can see, reweighting methods have a strong base in the literature. The theory we have given in Section 2.1 indicates that the validity of these methods are model-based. Therefore it can be important to study the characteristics of the nonrespondents to develop the most suitable model. In Sections 3 and 4, we perform such studies on each of two surveys. We see that the models help our understanding of the factors that contribute to nonresponse.

An important side benefit of such studies is to help the survey manager pinpoint areas for improvement in the data collection phase.

3. SURVEYS OF LABOUR AND INCOME DYNAMICS AND LABOUR MARKET ACTIVITY

3.1 Introduction

Statistics Canada launched a major panel survey of households in 1994 called the Survey of Labour and Income Dynamics (SLID). The survey follows individuals and families for six years, collecting information on their labour market experiences, income and family circumstances. Its origins are in several surveys, including the Labour Market Activity Survey (LMAS). The LMAS served both as a longitudinal and as a cross-sectional survey. Two panels have been conducted to date, a two-year panel during 1986-1987 and a three-year panel during 1988-1990. For each longitudinal panel, respondents who participated in the first wave were interviewed and traced. All persons living with them in the following waves were also interviewed but not traced. Different studies are currently being conducted on nonresponse to the LMAS in hopes of finding approaches that will minimize the impact of nonresponse on the SLID data. Here we discuss our study on model-based reweighting.

Similarly to its predecessor (LMAS), the longitudinal sample for SLID is selected from the sample of dwellings that participated in the Labour Force Survey (LFS) in January 1993. The LFS has a response rate of 95%. Out of those respondents close to 90% agreed to participate in SLID. This sub-sample of respondents, comprising 15,000 households, is defined as the longitudinal sample, representative of the Canadian population as of January 1993. The longitudinal sample will be interviewed for six years, with two interviews carried out each year. Note that a sub-sample of LFS respondents who had refused to participate to SLID has been selected for evaluation purposes. If they respond in subsequent years, we may be able to determine how different they are from the rest of the sample. Preliminary analysis could not find systematic differences in the LFS characteristics between the nonrespondents and the respondents. More studies will be done by linking the full sample to administrative files to be able to evaluate if there are differences in terms of income characteristics.

Attritional nonresponse will be compensated with a weighting adjustment. Imputation will be used to compensate for some nonresponse; for example, nonresponse that is non-attritional. The weighting will include the following steps:

i) calculation of the initial weight based on the sample design,
ii) nonresponse adjustment,
iii) post-stratification by province, age groups, and sex to the 1993 population estimates.

The longitudinal panel of LMAS has been used as the research vehicle for the nonresponse modelling and weighting adjustments.

3.2 LMAS Survey Design and Nonresponse

For the first interview of the panel, LMAS is conducted as a supplement to the January Labour Force Survey (LFS). All eligible respondents from the LFS are included in the LMAS sample. In the

subsequent waves, for the longitudinal component of LMAS, all respondents to the first wave are interviewed in January of the following year(s). People are traced if they have moved.

LFS uses a multiple stage sample design. A stratum is defined based on geographic variables. At least two distinct PSU's (primary sampling units) are selected within each stratum. LFS initial weights go through a series of adjustment factors at the stratum level to produce a sub-weight. This sub-weight is then adjusted to population estimates by province/age-group/sex groups, plus an adjustment by Economic Region and Census Metropolitan Area, to produce a final weight. More details may be found in Singh, Drew, Gambino, and Mayda (1990).

For the LMAS longitudinal sample, nonresponse adjustment is done at the stratum- component level, corresponding to a PSU or a group of PSU's, as defined for the LFS. A poststratification is then done to adjust the nonresponse adjusted weights to population estimates at the province/age-group/sex level.

When the LMAS file was evaluated, it was found that nonresponse was quite different among certain groups:

- movers, including people that could not be traced, had a nonresponse rate of close to 20% while nonresponse for non-movers was about 2%. This was by far the characteristic that presented the most differences,
- based on characteristics from Wave 1, persons that were employed in Wave 1 had higher response rates after three years than those who were unemployed in Wave 1,
- similarly, persons that were married in Wave 1 had higher response rates in Year 3, compared to those who were single in Year 1,
- persons who lived in non-urban areas in Year 1 had higher response rates after three years.

The different characteristics between respondents and nonrespondents suggested that nonresponse adjustments should be done at some level different than stratum-component. Logistic regression was used to model the nonresponse behaviour. The multiple logistic response function is

$$\text{logit}(p) = \log[p/(1-p)] = \boldsymbol{\beta}'\boldsymbol{x},$$

where p is the probability of response to the 1987 survey for a 1986 survey respondent, $\boldsymbol{\beta}$ is the column vector of regression parameters, and $\boldsymbol{x}$ is the vector of independent variables.

3.3 Modelling the Response Probabilities

The dataset for the 1986/87 panel of LMAS consisted of 66,817 individuals, of which 3,385 (5%) were nonrespondents to the 1987 interview. Demographic variables that were likely to be related to nonresponse were chosen from the 1986 LMAS master file as possible independent variables for the model.

The variables examined for inclusion in the nonresponse model were:

Province at 1986 interview
Urban/Rural area indicator at 1986 interview

Household size at 1986 interview
Type of dwelling (house; other) at 1986 interview
Status of dwelling (owned; rented) at 1986 interview

Sex
Age at 1986 interview
Marital status at 1986 interview
School attendance (full time; part time; none) in 1986
Highest level of education at 1986 interview

Any employment in 1986
Any unemployment in 1986
Any out-of-labour-force in 1986
Number of jobs in 1986
Any short tenure jobs (< 2 years) held in 1986
Any long tenure jobs (2 years or more) held in 1986
Any absences from work in 1986
Industry of job(s) in 1986

Average weekly income (over all jobs) in 1986
Received any unemployment insurance in 1986
Received any welfare in 1986
Moved (changed address between 1986 interview and 1987 interview)

All the categorical variables were converted to groups of dichotomous variables. The differences between respondents and nonrespondents with respect to the independent variables were analyzed. The correlations between all pairs of these variables were examined to find any potential multicollinearity.

First, a stepwise linear regression procedure was used to identify potentially useful variables for the modelling. This reduction in the choice of variables resulted in fewer variables to be entered into the logistic procedures saving considerable computer resources. The variables given in the STEPWISE procedure were entered into the SAS procedure PROC LOGISTIC with the BACKWARD and FAST options. These options allowed LOGISTIC to use an approximate backward elimination method to eliminate nonsignificant variables. Different logistic regression models were fitted to the full dataset using combinations of the most significant variables identified from the sample file. A consideration in choosing the model was the number of variables. It was desired to have a model with a small number of variables so that utilizing the model would be simple.

The model is used to make adjustments to the weights of the respondents in the second year (1987). For this model, the dependent variable was total nonresponse, and the independent variables were characteristics observed the previous year (1986) plus the current year's information (1987) on whether or not the person moved.

The BACKWARD option of PROC LOGISTIC was used with the sample file to identify eight variables related to nonresponse.

Male	(MALE)
Single	(SINGLE)
Rented dwelling	(RENT)

Any employment	(ANYEMP)
Highest education=secondary	(EDUCSEC)
Moved since 1986 interview	(MOVED)
Household size, to a maximum of 8	(HHS)
Age	(AGE)

Before fitting the models on the full dataset, the two continuous variables (household size and age) were examined for linearity on the logit scale. As with the prediction model, the age variable was replaced by two dichotomous variables for age: AGE1 for persons aged 25-54, AGE2 for persons aged 55-69 - the survey was conducted for persons aged 16-69 - and a transformation was applied to household size (HHSTRANS=|HHS-4.5|).

Four models were fitted to the full dataset: (1) using all eight variables; (2) using all except RENT; (3) using all except EDUCSEC; (4) using all except EDUCSEC and AGE. Although all eight variables were significant using the sample file, when the models were fitted to the full data file, certain ones no longer appeared important. However, it was decided to retain them in the models anyway. The statistics for evaluating the fit of the models indicated few differences between the four models. The Pearson residuals were plotted against the fitted values and the residual plots were examined. The residuals from Model (3) indicated a slightly better fit with fewer extreme values. Again using the sample file, the data were examined for the presence of two-way interactions between the variables in the model. Two sets of interactions were added to the model: the (AGE1 AGE2)*HHSTRANS and (AGE1 AGE2)*SINGLE. A summary of the fitted values for this model is given below. Note that the age and single variables as well as their interactions are not statistically significant. Nevertheless, when a model was fitted with these variables removed, it was found that there were more extreme values in the residuals.

Table 1

Parameter Estimates for Weighting Final Model.

Variable	$\hat{\beta}$	s.e.	χ^2
INTERCEPT	-3.81	0.14	702.59
HHSTRANS	0.13	0.06	4.97
MALE	0.25	0.04	41.98
RENT	0.23	0.04	29.14
SINGLE	0.11	0.16	0.43
MOVED	2.31	0.04	3065.95
AGE1	-0.15	0.17	0.75
AGE2	-0.19	0.15	1.65
AGE1*HHSTRANS	0.02	0.07	0.07
AGE2*HHSTRANS	0.05	0.06	0.55
AGE1*SINGLE	0.13	0.18	0.52
AGE2*SINGLE	0.11	0.17	0.40

Using the estimated parameters from the final model, predicted probabilities of nonresponse were calculated for all respondents to the 1987 interview and a nonresponse adjustment was made. Finally, a poststratification adjustment to population control totals at the province-sex-agegroup level, yielded the 1987 final weight.

3.4 Evaluation of the Weights

If the nonresponse weighting adjustment is adequate, there should be no difference in estimates obtained from the 1986 respondents and estimates obtained from the 1987 respondents when tabulating on 1986 characteristics. A number of demographic and labour-related characteristics were evaluated. Estimates were calculated using the 1986 weights, the 1987 model-adjusted weights, and the 1987 regular weights, including a ratio-adjustment at low geographic levels for nonresponse adjustment. For each characteristic a 95% confidence interval was calculated for the estimate based on the 1986 weights. The two 1987 estimates were compared for differences to the 1986 estimates as well as differences to each other. Tables 2 and 3 below show some of the results. Table 3 incorporates the poststratification adjustment, which, in general, improves the estimates.

Table 2

Comparison of the estimates with the two non-response adjustments, before the post-stratification, tabulated on 1986 characteristics.

	1986 estimate	95% c.i. for 1986 estimate	1987 model-based estimate	1987 regular estimate
Marital Status Married	64.6%	(64.1,65.1)	65.1%	65.7%
Single	26.7%	(26.3,27.0)	26.3%	25.7%
Widowed	3.1%	(2.9,3.3)	3.0%	3.0%
Divorced	5.7%	(5.4,6.0)	5.6%	5.5%
Highest Education Grade 0-8	14.7%	(14.2,15.2)	14.6%	14.6%
Secondary	50.3%	(49.7,50.9)	50.0%	50.0%
Some Post-Secondary	10.1%	(9.8,10.4)	10.2%	10.1%
Post-Sec. Cert./Dip.	12.9%	(12.5,13.3)	13.1%	13.1%
University Degree	12.0%	(11.6,12.4)	12.2%	12.2%
Weeks Employed in 1986 0 weeks	22.8%	(22.4,23.2)	22.6%	22.5%
1-26 weeks	12.0%	(11.7,12.3)	11.7%	11.6%
27-48 weeks	12.2%	(11.9,12.5)	12.1%	12.0%
49-52 weeks	53.0%	(52.4,53.6)	53.6%	54.0%

Of all the characteristics compared, only one 1987 estimate was outside the 1986 confidence interval: weeks employed=49-52 using the regular weighting. One pattern was clear, however. The estimates using the model-based weights were consistently closer to the 1986 estimates than those using the regular method of weighting.

4. FARM FINANCIAL SURVEY

4.1 Introduction

The Farm Financial Survey (FFS) has been a regular agricultural survey since 1980. The objective of the survey is to gather financial information on Canadian farmers. The survey collects information on revenues, expenses, assets and liabilities. Crop and livestock information are also collected to measure physical characteristics of the farms. Due to the collection of sensitive data, a low response rate has always been observed for the survey. A study was initiated on the 1992 survey data to identify the causes of nonresponse and possible solutions to reduce its impacts on the estimates.

The population of interest consists of all Canadian farms active for the reference year, excluding the multi-holding companies, the institutional farms, the community pastures, the farms on Indian Reserves and the farms with less than $2,000 in sales. The survey population is represented by a list frame and an area frame. The 1992 list frame was a register of all of the 1986 Census farms without the farms defined by the above exclusion rules. The list frame was stratified within each province by farm type and by farm size. The farm size was defined by the total farm assets derived on the Census.

The area frame was used to compensate for the undercoverage due to the Census itself or caused by new farms which started their activities since 1986. Basically, the area frame was a list of land segments outlined on topographic maps. Stratified replicates of segments were selected from the area frame. All farmers operating some land in the sampled segments were enumerated, and a register was created. There were 1,153 area frame farms that did not appear on the list frame. They were all contacted for the FFS as for other agricultural surveys. In addition to the area frame farms, a stratified sample was selected from the list frame to obtain a overall sample of about 12,000 farms. See Britney and Poirier (1992) for more details on the 1992 FFS sample design.

Domain estimation within each stratum was performed to obtain estimates of level from both the list and area samples. The simple expansion estimator was used on the 1992 list sample. The initial weighting was done by stratum using the population size over the <u>observed</u> sample size, so that a nonresponse adjustment is made at the stratum level. For the area frame, the estimation was done separately by replicate. For a given replicate, the data were aggregated at the segment level by applying to the farm data, factors corresponding to the proportion of the farms within the segment. Then, the segment totals received expansion weights (π) to represent the population. When nonresponse occurred for an area farm, the respondents within the same segment were reweighted on an area basis to compensate the farm land for which data were unavailable. For both the list and area units, partial nonresponses were donor imputed and used the same way the regular respondent were. Details are given by Maranda (1989).

The nonresponse observed in the 1992 Farm Financial Survey was relatively important. The FFS questionnaire was relatively long

with many sensitive questions related to the financial balance sheet. The resulting total unit-level refusal rate of about 15% across the country was the highest of our agricultural surveys. In addition to the total refusals, the no-contacts represented another 5% of the sample. Some provinces presented higher nonresponse rate than others. In Saskatchewan, data were unavailable for almost 30% of the sampled farms. Table 4 shows the nonresponse distribution across the country.

Table 4

1992 Nonresponse Distribution

Province	Sample Size	Total Refusal	No-Contact
Newfoundland	211	20	16
P.E.I.	528	64	14
Nova Scotia	668	74	11
New Brunswick	537	48	18
Quebec	1311	124	51
Ontario	1513	250	84
Manitoba	1756	321	109
Saskatchewan	1880	424	126
Alberta	1868	312	109
B.C.	1448	175	138
Total	11720	1812	676

4.2 Nonresponse Models

A part of our study was first to identify the causes of nonresponse. This could help taking decisions related to the collection methods to increase the response rate. It also allowed the identification of factors that may be considered in any nonresponse reweighting models. Since, the no-contacts and the refusals were possibly caused by different factors, they were kept separate in all of the hypotheses we made. The potential causes that were studied on the 1992 FFS data are:

1 - **The frame origin:** This corresponded to whether or not the farm was selected from the list frame. Since the area frame farms were conceptually missed by the Census, they probably showed characteristics that trend to generate nonresponse. Also, since the area sample is being used by many agricultural surveys, the area frame farms might refuse because of their response burden.

2 - **The farm size:** The capability or the will to respond could depend on the farm organisation and on its size.

The size was evaluated using the farm assets and sales obtained from the 1986 Census of Agriculture. This size was available only for the list units.

3 - **Geography:** The geographic location aimed to identify the interviewer effect and the impact of farmer associations which could boycott government surveys because they were not benefited by their programs. Census divisions were used to verify this hypothesis.

4 - **Farm type:** The farmer's availability depends on the type of his farm. Seven categories of farm type were used to differentiate the farms.

5 - **Response burden:** Because the large number of agricultural surveys held in a short period of time, the response burden became important for some farmers. The overlaps with the December Stock Survey and the January Livestock Survey (JLS) were both studied to verify its impact on the response rates. These surveys were both conducted less than two months before the FFS. The effect of the overlap with the previous FFS, held in 1990, was also investigated.

6 - **Age of operator:** The age of an operator could affect its will to cooperate, but the data available to verify this hypotheses were not reliable enough to do any studies.

Tests of independence were conducted to verify if any of the above factors could affect the response status: 'completed', 'no-contact' and 'refusal'. The partial refusals were included with the completed questionnaires. The statistic used to conduct the independence tests was the weighted Pearson statistic χ^2 with the Fellegi (1980) correction to take into account the design effect. This test is known to be conservative.

The farm assets and sales, which were both indicators of the farm size, were replaced by categorical variables defined using the estimated quartiles. The census divisions representing the geographic location were grouped into a maximum of 9 classes within each province. This ensured a minimum number of observations within each cell of the cross classification with the response status.

In some cases, where dependence was detected between the factors and the response status, additional tests were conducted to identify the nature of the dependence. This was done through statistical tests on proportions. The most important conclusions are described here but more details can be found in Poirier (1994).

The independence tests, conducted with a confidence level of 5%, identified certain causes of nonresponse. First, within each province except Ontario, the farm type had a high impact on nonresponse. Also, the farm size measured in term of assets affected the response rates in most of the provinces, but no significant impact was due to the sales variable. The geographic location and the response burden generated by the previous FFS survey significatively affected the probability to respond in three provinces. Finally, the frame origin and the overlap with the January Livestock Survey or the December Crops Survey seemed to not affect the response status at all.

As in Section 3, we modelled the nonrespondent behaviour with logistic regression modelling using the SAS procedure LOGISTIC. We performed the analysis separately by province. Using frame origin as an independent variable, the results confirmed the previous conclusions of no frame effect. Since some variables were not available for the area sample and since the frame origin did not seem to affect the response, the remaining analyses were performed only on the list units, which represented more than 90% of the whole sample. In the rest on this paper, the results applied for the list units only.

For the purposes of this study, the following variables were included in the model:

i) Assets (1 if assets are smaller than the median, 0 otherwise),
ii) Sales (1 if sales are smaller than the median, 0 otherwise),
iii) $Type_i$ (1 if in the ith farm type, 0 otherwise),
iv) $Area_i$ (1 if in the ith geographic area, 0 otherwise),
v) FFS (1 if in the 1990 FFS sample, 0 otherwise),
vi) JLS (1 if in the 1992 JLS sample, 0 otherwise),

The farm types are (1) crop farms, (2) dairy farms, (3) cattle farms, (4) hog farms, (5) poultry farms, (6) sheep farms, and (9) unknown type of farm.

The variables that were found more significant by the BACKWARD option within the provinces were kept in the model. The most commonly selected variables were the farm types and the FFS variables. Table 5 shows the resulting estimated parameters corresponding to the variables kept in the model. It also provides the attached χ^2 values.

Table 5

Nonresponse Logistic Parameters with their χ^2 Values

Prov.		Intercept	FFS	Assets	Type4	Type5	Type6	Type7	Area1	Area2	Area3	Area5	Area6	Area7
NFLD	B	1.37	1.05	.	.	-1.78	.	.	.	.	.	.	.	.
	χ^2	37.23	5.75			6.76								
PEI	B	1.75	.	.	.	.	.	.	.	.	.	.	.	.
	χ^2	204.18												
NS	B	1.93	.	.	.	.	.	.	.	.	.	.	.	.
	χ^2	275.05												
NB	B	1.65	0.71	.	.	.	.	.	.	.	.	.	.	.
	χ^2	98.87	6.79											
QUE	B	2.05	.	.	.	-0.58	.	.	.	-0.45	.	.	.	.
	χ^2	392.98				5.11				4.03				
ONT	B	0.93	0.53	.	.	.	1.07	.	.	.	.	.	.	.
	χ^2	71.00	14.18				4.09							
MAN	B	0.94	0.44	.	.	-0.45	.	-1.16	0.59	.	.	.	.	.
	χ^2	113.35	13.34			4.21		38.90	14.05					
SASK	B	0.94	0.51	0.31	.	.	.	-2.40	.	.	.	.	-0.61	-0.32
	χ^2	78.44	18.67	6.85				134.31					12.20	4.22
ALB	B	1.04	0.69	0.26	-0.33	.	.	-1.12	.	.	.	.	.	.
	χ^2	79.25	30.12	4.30	5.51			40.48						
BC	B	0.88	.	.	.	.	.	.	0.93	0.86	0.47	0.43	.	.
	χ^2	74.82							15.29	21.01	5.85	5.13		

Variables that were not significant for any of the provinces have been removed from this table. From the χ^2 results, it appears that the FFS overlap and the farm Type7 (representing the sheep farms) have the most important impacts on nonresponse. The positive FFS parameters mean that farms overlapping the previous FFS tended to have higher response rates, whereas the negative sheep farm parameters (Type7) imply they tended to respond less often.

Weighted regressions were also fitted to the data using the WEIGHT statement of the LOGISTIC procedure. The weighting variable was defined at the stratum level as the design weight adjusted to the overall sample size. Stratum level adjustments were not performed. The resulting estimated parameters were very close to the first set of estimates which, as we explained in Section 3, is highly desirable.

4.3 Evaluation of the Weights

To evaluate the nonresponse adjustment, the 1992 frame values representing farm assets were estimated from the sample. Assets levels were estimated for each province with the corresponding

coefficient of variation (CV), including the nonresponding units. Then, estimates based only on respondents were produced, using the original weight, adjusted for nonresponse at the stratum level only. By comparing both set of estimates we could derive the nonresponse bias introduced by the current method. Finally, regression adjusted estimates were produced from the above logistic model.

If we denote by $\hat{Y}_0$ and CV_0 estimated level and coefficient of variation, respectively, from the full sample, and $\hat{Y}_{adj}$ the corresponding adjusted estimates based only on respondents, we can estimate the bias associated with the adjustment model. In Table 6 we show the results.

Table 6

Comparison of the Adjustment Models for 1992 Frame Value of Farm Assets

Prov.	Y_0	CV_0 (%)	Stratum Adjusted Weight	Logistic Adjusted Weight
			BIAS (%)	BIAS (%)
NFLD	7.7 E07	2.91	2.34	1.80
PEI	6.7 E08	0.76	-0.35	0.19
NS	7.9 E08	0.68	-1.15	-0.27
NB	6.1 E08	0.82	-0.16	0.46
QUE	8.5 E09	0.53	-0.28	-0.04
ONT	2.1 E10	0.56	-1.16	-0.85
MAN	9.0 E09	0.57	0.21	0.45
SASK	2.7 E10	0.52	-0.40	-0.89
ALB	2.7 E10	0.57	0.53	-0.12
BC	5.6 E09	0.62	-2.32	-2.24
TOTAL	1.0 E11	0.25	-0.35	-0.54

We see that the logistic adjusted weight generally performs better, but not consistently so. In fact the bias increases for NB, MAN, SASK, and the Total. To improve the model, inclusion of some interaction factors like size and farm type, or size and geography was tried but they were rarely kept in the model and when they were, the resulting effects were small and their impact was negligible.

4.4 Conclusion

The selected model did not consistently provide the expected bias adjustment. This may be caused by a low number of factors included in the model or by the fact that significant factors were used in the frame stratification. Future work might include looking for more interactions using the Automated Interaction Detection method used in Section 3. Also, now that the 1993 data are available, the study could be extended.

5. SUMMARY

Nonresponse adjustment through reweighting is now in common use. We have shown that the success of this technique generally depends on having available variables that can be used as good predictors of the nonresponse behaviour. Having such variables, various models can be used to adjust the estimates based on the predicted response propensities. This seems to be the best general approach. Other approaches include using estimation methods such as regression estimators to compensate for the deficiencies of the sample. We have seen that if the regression models are valid, the nonresponse bias vanishes.

We have concentrated here on asymptotic biases. However, there are still many unresolved issues for estimation of variances and construction of confidence intervals. As well, we have not properly addressed the issue of whether or not to use the sampling weights when fitting the nonresponse models. In our examples, the weighted and unweighted versions of the estimated response models gave similar results. This is highly desirable since it confirms the validity of the model.

Nonresponse problems will not go away. A better understanding of the response mechanisms will lead to better survey practices in the long run.

ACKNOWLEDGEMENTS

We are especially grateful to Graham Kalton, who most kindly provided us with a comprehensive list of references on this topic. We would also like to thank Ralph Folsom and Steven B. Cohen for sending us some of their recent work.

REFERENCES

Alexander, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology* 13, 183-198.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics* 4, 251-260.

Binder, D.A. and Théberge, A. (1988). Estimating the variance of raking-ratio estimators. *Canadian Journal of Statistics* 16, 47-55.

Britney, H. and Poirier, C., (1992). 1992 Farm Credit Corporation: Design documentation, Internal Paper, Agriculture Section, Business Survey Methods Division, Statistics Canada.

Chapman, D.W., Bailey, L., and Kasprzyk, D. (1986). Nonresponse adjustment procedures at the U.S. Bureau of the Census. *Survey Methodology* 12, 161-180.

Deville, J-C., Särndal, C-E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association* 88, 1013-1020.

Ekholm, A. and Laaksonen, S. (1991). Weighting via response modeling in the Finnish Household Budget Survey. *Journal of Official Statistics* 7, 325-337.

Eltinge, J.L. and Yansaneh, I.S. (1993). *Weighting adjustments for income nonresponse in the U.S. Consumer Expenditure Survey*. Technical Report No. 202, Department of Statistics, Texas A&M University, College Station, Texas.

Fellegi, I.P. (1980). On adjusting the Pearson chi-square statistic for cluster sampling. *Journal of the American Statistical Association* 71, 665-670.

Folsom, R.E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. *Proceedings of the Social Statistic Section*, American Statistical Association, 197-202.

Iannacchione, V.G., Milne, J.G. and Folsom, R.E. (1991). Response probability weight adjustments using logistic regression. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 637-642.

Judkins, D. and Lo, A. (1993). Components of variance and nonresponse adjustment for the Medicare Current Beneficiary Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Lepkowski, J., Graham, G., and Kasprzyk D. (1989). Weighting adjustments for partial nonresponse in the 1984 SIPP Panel. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 296-301.

Maranda, F. (1989). Proposal for NFS estimation, Internal Paper, Agriculture Section, Business Survey Methods Division, Statistics Canada.

Oh, H.L. and Scheuren, F. (1983). Weighting adjustments for unit nonresponse. In *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies* (W.G. Madow, I. Olkin and D. Rubin, eds.), 143-184. Academic Press, New York.

Poirier, C. (1994). The causes of non-response in the context of agricultural surveys. Internal Paper, Agriculture Section, Business Survey Methods Division, Statistics Canada.

Särndal, C-E. and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review* 55, 279-294.

Särndal, C-E., Swensson, and Wretman, J. (1992), *Model Assisted Survey Sampling*. Springer-Verlag, New-York.

Singh, M.P., Drew, J.D., Gambino, J.G., and Mayda, F. (1990). *Methodology of the Canadian Labour Force Survey 1984-1990*. Statistics Canada, Catalogue 71-256, Ottawa.

NONRESPONSE DISCUSSION

J. Michael Brick
Westat, Inc.

I would like to thank the authors of both papers for presenting insights on very different aspects of the nonresponse problem. The first paper, "Exploring Nonresponse in U.S. Federal Surveys" by Gonzalez, Kasprzyk, and Scheuren, reports on the activities of a Subcommittee on Nonresponse for the Federal Committee on Statistical Methodology (FCSM). The main focus of this paper is on trends in response rates for federal surveys over the last 10 years. The second paper, "Model-based Reweighting for Nonresponse Adjustment" by Binder, Michaud and Poirier, reports on research in a particular approach to reduce the bias due to nonresponse.

In "Exploring Nonresponse in U.S. Federal Surveys," the authors describe recent investigations of the levels of response rates and trends in nonresponse and refusal rates in federal surveys conducted from 1982-1991. The other parts of the mission of the subcommittee include reporting on correlates of response rates and other related matters. We look forward to reports on those activities later this summer.

The findings on trends in nonresponse rates are especially welcome. In general, these data show that there has been no increase in nonresponse or refusal rates over the last decade. This finding contradicts the conventional wisdom that nonresponse rates, and refusal rates in particular, have been rising due to the increases in the burden on respondents. It clearly shows that it is risky to make general conclusions about this type of phenomenon based on limited data.

Given that response rates seem to be fairly stable over the last decade, the question still remains about the value of judging the quality of survey data over time from this type of data. After all, we are really interested in the quality of the data, not the response rates themselves. However, response rates are the only measure of quality typically produced from surveys. While having something that is measurable is useful, response rates are frequently not very informative of the quality of the data.

Using what Deming calls the "modern" approach to quality control, it is important to recognize that the response rates are really a <u>product</u> characteristic of a survey. Survey methodologists, on the other hand, must concentrate on the <u>process</u> of collecting data. The danger is that examining product variables alone can mask the process related to the quality of the data. For example, if the population surveyed changes so that a group that has generally high response rates is now sampled, then we might expect the response rates to increase. Thus, response rates might change even if respondents willingness to participate remained the same. Similarly, unweighted response rates may be affected by changes in the sample design, such as changes in the sampling rates for different segments of the population.

For survey methodologists, the process must be the most important part of their job. They are responsible for producing survey data of the highest possible quality within the constraints of the survey. In many ways, this paper only peripherally addresses this set of core users, because it does not discuss the process itself. To do this requires more process data, much of which is not comparable across surveys.

The focus on response rates is also limited for other reasons. The important relationship between response rates and costs is not explored in this paper. Cost data are crucial, since the cost of obtaining the same level of response rates may have increased or decreased. Unfortunately, cost data are extremely difficult to obtain and formulate in a manner that is useful for comparison

purposes. Developing a general method for analyzing cost data remains, in my opinion, one of the most important unresolved problems in survey research.

As the authors of this paper indicate, improving response rates does not always improve the quality of the survey estimates. This has been misinterpreted by some researchers to mean that response rates are not important. They are important and reasonable efforts should be made to eliminate nonresponse. The real question is related to how much effort should be placed on improving response rates versus reducing other sources of error in a survey. The answer to this question is not always clear, but there are guidelines for reasonable practice. In general, the resources devoted to a source of error (be it sampling or nonsampling) should be proportionate to the size of that error relative to the sum of sampling and nonsampling errors.

In one of their recommendations, Gonzalez, Kasprzyk, and Scheuren state that research should be encouraged by survey administrators as a way to improve the effectiveness of data collection operations. They include a specific recommendation that studies of collecting items to improve post-survey adjustments should be encouraged. I think this is a very interesting choice for the recommendation. It supports collecting items rather than exploring new methodologies. I applaud this direction. Research on collecting additional items is a sound direction to improve our ability to decrease nonresponse bias.

The paper by Binder, Michaud and Poirier, "Model-based Reweighting for Nonresponse Adjustment" is very consistent with the recommendations of the FCSM Subcommittee. Since the type of nonresponse discussed in this paper is not planned, randomization does not apply and a modeling approach must be used to reduce the bias. The authors discuss the results from applying a specific form of modeling nonresponse for two different surveys.

Their efforts to reduce the bias in the estimates for the Survey of Labour and Income Dynamics (SLID), closely parallel work done on the U.S. Survey of Income and Program Participation (SIPP). The Census Bureau sponsored several efforts to examine methods to reduce attrition bias in the SIPP and many of the papers from these efforts will be presented this summer. In the SIPP research conducted at Westat, we used many of the same techniques to model the nonresponse as Binder, Michaud and Poirier, with much the same result.

I would like to make two specific comments on the paper prior to returning to a discussion of the importance of collecting items for reducing nonresponse bias. The first deals with the use of the logistic regression model used to predict response propensity. This work, and our own efforts in the SIPP research, suggest that these types of models are not the most conducive to uncovering relationships between variables when faced with a large number of potential predictor variables. Categorical search algorithms seem better equipped at identifying nested relationships that might be important for reducing nonresponse bias. In this paper, these relationship were uncovered with the search algorithms and added to the logistic regression model. We have used the output from the search algorithms to form cells for nonresponse adjustment directly. Both methods seem to work reasonably well.

The second comment is really a question. The item most related to nonresponse was mover status. People who moved were far more likely to be nonrespondents. This characteristic has obvious appeal and could be of great importance in nonresponse bias reduction. The question is how nonrespondents were defined as movers or nonmovers based on the data available. In most surveys, we do not have these data for nonrespondents. I think it would be useful for other survey methodologists to see how the authors made this classification for SLID.

If an item such as whether or not the person has moved is available, then it can be used to estimate response propensity. Having items that are good estimators of response propensity is critical for reducing bias. I have found that different nonresponse adjustment methods, when

properly applied, often have approximately the same performance. This may happen because almost all the methods estimate individual response propensities conditional on auxiliary data. Estimating a response propensity for each unit is very difficult since only one observation is available (the unit either responded or did not respond). Items related to the likelihood of responding at the individual level are, therefore, extremely valuable to improve this estimation.

The authors take an approach that seems most sensible. They include as large a number of items as possible in the model without allowing the adjustments to become so variable as to unduly increase the variance of the estimates. Since nonresponse bias is a function of the covariance between the estimate and the response propensity, items related to both should be entertained. As suggested in this research, the first items included should be those related to response propensity. If it were possible to model these response propensities precisely, then the nonresponse bias for all the estimates would be eliminated. Since the modeling is imperfect, additional items highly related to key statistics from the survey should also be included, where possible.

While longitudinal surveys and studies in which there are frames with substantial data on the sampled units may be able to use this method to reduce the bias due to nonresponse, the more typical situation may be that faced in the Farm Financial Survey. In this case, the data on the frame are limited and do not appear to be predictive of either response propensity or the key estimates. The methods used are not very effective on reducing the nonresponse bias, as might be expected. If the nonresponse is large enough in a survey of this nature, the recommendation of the FCSM to develop a mechanism for collecting items useful for nonresponse adjustment should be seriously considered. This common situation reinforces the importance of keeping nonresponse to a minimum.

DISCUSSION

Joseph L. Schafer
Pennsylvania State University

1. Comments on "Exploring nonresponse in U.S. Federal Surveys"

This paper by Gonzalez, Kasprzyk, and Scheuren (GKS) is worthwhile reading. Section 2 gives a nice historical overview of the development of statistical methods for survey nonresponse. As this section clearly demonstrates, many of the great strides in the practice of nonresponse adjustment, and survey sampling in general, have come about as a direct result of personal interaction between statisticians in federal agencies and in those in academic circles. It is our hope that this type of fruitful interaction will continue in the years ahead.

In the world of surveys, there are many different types of nonresponse. Unit nonresponse arises when the entire vector of survey variables for a sample unit is missing. Item nonresponse arises when individual elements of the vector are missing. As pointed out in the other paper in this session, panel surveys often suffer from wave nonresponse, which occurs when the entire vector of survey variables is missing for a unit at a particular occasion or wave. The GKS paper has introduced a new type of missingness: nonresponse nonresponse, which occurs when members of the federal statistical community fail to report their nonresponse rates to members of the FCSM subcommittee. An even worse type of nonresponse nonresponse occurs when producers of federal surveys do not report basic information on nonresponse--such as nonresponse rates and methods of adjustment--to the users of their data.

It is worthwhile to ask whether nonresponse nonresponse is ignorable, in the sense defined by Rubin (1987). In the FCSM subcommittee's study, the nonresponse nonresponse would be ignorable if the nonresponding statisticians' surveys were representative of all federal surveys in terms of missingness rates, methods of adjustment, etc. Chances are, the nonresponding statisticians had desks that look like mine--paperwork stacked up a foot high all around--and they didn't find the time to return the questionnaire. Or, perhaps they didn't respond because they hadn't been tracking basic information such as nonresponse rates, and compiling the information would have required an unusual amount of effort. Whether those characteristics of the nonrespondents are systematically related to the basic study variables is anyone's guess. In the common, non-technical sense of the word, however, this nonresponse to the FCSM subcommittee's study is

probably not ignorable; we shouldn't ignore it. If members of the subcommittee found it so difficult to obtain even basic information on nonresponse from the very people who produce the survey, imagine how hard it must be for the average data user to do the same thing.

It was astonishing to see that almost half of the establishment surveys in the study had no tracking at all of the various components of nonresponse. We hope that in the future, producers of federal surveys will take to heart the recommendations of this report and devote a little more time and effort to studying and documenting the levels and causes of nonresponse. At any rate, it is comforting to learn that the basic perception that many of us had--that nonresponse rates have been rising over the past few years--may be only a perception, and the sum total of the various nonresponse components may not have changed very much.

Reading this paper made me think a lot about the future, and what directions we should take in our research on nonresponse and nonresponse adjustment. The suggestion near the end of the paper--that data collectors and statisticians work together as a team, sharing information in a way that is mutually beneficial--is especially thought-provoking. If this were done, it would open up entirely new avenues for developing improved methods of nonresponse adjustment. As shrinking budgets force us to re-allocate our resources, it may not be possible to continue to spend so much money chasing after nonrespondents, trying to get them to hand their data vectors over to us. It's becoming increasingly likely, for example, that the Census Bureau will not have the resources in the year 2000 to follow up on every housing unit that doesn't mail back its census form with a personal interview. Statisticians and data collectors should start to think long and hard about which nonrespondents they should attempt to follow up, and how persistently they should attempt to do so. They need to weigh the costs and benefits involved in converting refusals and not-at-homes to responses, so that they may decide whether the money might be better spent elsewhere.

In regard to this point, I would like to make two basic observations. The first observation is that <u>not all nonrespondents are equally important</u>. In industrial settings, statisticians have to carefully design their experiments. Trial runs are typically expensive, and they have to decide under what values of X they should collect their next value of Y. They know that certain, carefully chosen combinations of Xs will help them to "nail down" the quantities of interest much better than other combinations of Xs, so they choose their values of X carefully before spending additional money to collect another Y.

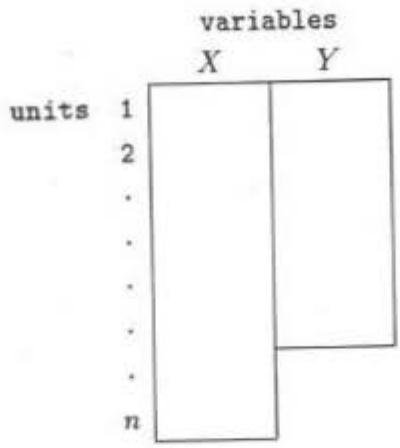


Figure 1: Data after the initial phase of collection

In a survey setting, there is typically an initial phase of data collection in which every unit in the sample has been contacted once, or at least an attempt has been made to contact them once. After this initial phase there will be a pool of nonrespondents, and there will be (a limited amount of) money to spend on attempts to convert some of them to respondents. Which nonrespondents should we go after first? To answer this question, we should look at their Xs. At this point, the data will resemble the diagram in Figure 1. Variables available from the sample frame, denoted by X, will known for all sample units. Survey variables, denoted by Y, will be available for the respondents only. Let us partition Y as $Y = (Y_{obs}, Y_{mis})$ where Y_{obs} denotes the values of the survey variables for the respondents, and Y_{mis} the values of the survey variables for the nonrespondents. It should be possible to build a regression model for the distribution of Y_{obs} given X, and then use this model to guess or predict what the values of Y_{mis} will be. These predictions may then form the basis for ranking the nonrespondents according to the priority with which they should be followed up.

Let $P(Y_{obs}|X,\theta)$ denote the form of the model fit to Y_{obs} given X, where θ represents some unknown parameters. Let y_i^* denote an element of Y_{mis}. To the classical survey statistician, y_i^* is a fixed, unknown constant. To the Bayesian, however, it is unknown and therefore a random variable. The uncertainty about y_i^* can be expressed Bayesianly as

$$V(y_i^*|X,Y_{obs}) = EV(y_i^*|X,Y_{obs},\theta) + VE(y_i^*|X,Y_{obs},\theta),$$

where the outer moments are taken with respect to $P(\theta|X,Y_{obs})$, the posterior distribution of the unknown model parameters. The first component on the right-hand side, $EV(y_i^*|X,Y_{obs},\theta)$, represents the residual variation of y_i^* about its predicted value from the regression model. The second component,

$VE(y_i^* | X, Y_{obs}, \theta)$, represents the uncertainty in the regression prediction itself. This decomposition suggests that if we want to minimize our uncertainty after a limited amount of followup, we should target as high priority those units that (i) have a high amount of residual variance, and (ii) have high leverage for estimation of θ. In other words, we should try to follow up the units (i) whose values of Y cannot be predicted well by our model, and (ii) whose values of X are unusual and thus, if they were converted to respondents, could greatly improve our ability to predict the missing Y values for the other nonrespondents.

In addition to the predictive variance of y_i^*, we also need to consider the probability that a nonresponding unit can successfully be converted to a respondent. Even if the predictive variance for a particular unit is high, it may not make sense to attempt followup if the followup operation is likely to be unsuccessful. This suggests construction of another regression model to predict the probability of successful followup--or, perhaps, the cost of successful followup in terms of number of attempts, field worker time, etc. Data for fitting this model might come from similar survey operations of the past, perhaps updated by data from the current survey as they become available.

As data collectors begin to share information with statisticians on an ongoing basis, one can imagine the development of a continuous-loop feedback system in which the field-operations unit provides data on respondents as they become available, and the statistical unit processes the information, updates the parameters of its regression models, and decides which of the remaining nonrespondents should be designated for followup.

The second general observation that I would like to make is that not all nonresponse mechanisms are the same. From a theoretical standpoint, it is useful to classify nonresponse mechanisms into two categories: mechanisms that are ignorable and mechanisms that are nonignorable. Using the notation developed above, an ignorable mechanism is one in which the probabilities of response do not depend on Y_{mis} after accounting for dependence on X and Y_{obs}. Ignorable nonresponse mechanisms tend to be easier to deal with than nonignorable ones, and virtually all methods of nonresponse adjustment in use today make some implicit assumptions of ignorability.

From a practical standpoint, however, nonresponse mechanisms should probably be classified into a slightly different dichotomy: mechanisms that are known to be ignorable, versus mechanisms that are not known to be ignorable. Mechanisms that are known to be ignorable include those in which the missing data are missing by design. Surveys that employ double sampling,

matrix sampling, etc. result in rectangular datasets with patches of missing data that are missing by design; the data are unrecorded because the data collector never intended to collect them. When data are missing by design, ignorable missing-data techniques may be applied without fear of introducing bias. The more insidious type of missingness mechanism is the unknown type. When the nonrespondents are a self-selecting subsample, we do not really know how strongly the selection process may be related to the missing data Y_{mis}. When faced with missingness of this type, the only thing that a practitioner can usually do is to apply some ignorable missing-data technique and hope for the best--i.e. pray that any biases incurred by nonignorability will not be severe.

As long as resources for data collection are finite, a certain amount of missing data will be inevitable. But the point that I want to make is this: By intelligent allocation of resources in the followup operation, we may be able to convert a substantial amount of the missing data that would ordinarily be of the type "unknown" to the type "ignorable." In a typical followup operation of today, attempts are made to follow up nonrespondents in a rather haphazard (i.e. unplanned) fashion until the resources run out, at which time the data collectors close out their operation and get on with their lives. Decisions about which nonrespondents are to be followed up are not made by a central decision-making unit, but are made in the field by supervisors or by the interviewers themselves. It may be that the field staff is placing high priority on the nonrespondent units that appear to be easy to get, thereby attempting to minimize the number of nonrespondents that remain after closeout. Although minimizing the nonresponse rate is a laudable goal, the end result is that all of the nonresponse that remains after closeout is of the type "unknown." From a statistician's point of view, a better strategy may be to concentrate one's resources on obtaining data for a probability sample of nonrespondents, a sample that is guaranteed to be representative of the nonrespondent pool. Even if data for these units are expensive to obtain--e.g. requiring a large number of call-backs--and the overall rate of missingness in the end is higher than it would have been if the followup decisions were made by field staff, the end result will be that the missing data for nonrespondents that are not included in the followup sample will be of the type "ignorable."

As a scientist, I would be willing to trade a few percentage points of missingness for a guarantee that (at least most of) the missing data are ignorable. I suspect others would as well. The tradeoff between the cost of missing data versus the benefit of knowing the missingness mechanism is a subtle but important issue to which statisticians ought to pay more attention

in the future.

2. Comments on "Model-based reweighting for nonresponse adjustment"

This paper by Binder, Michaud, and Poirier (BMP) discusses in detail the methods used by StatCanada to model response propensities in two of its ongoing surveys. By reading this paper, and examining the tables of regression coefficients, one develops an excellent sense of what factors may be related to nonresponse in demographic and establishment surveys. It is interesting to note that in regard to these two surveys, BMP and StatCanada seem to be following the recommendations of the authors of the previous paper: clearly documenting the rates of nonresponse, the factors related to nonresponse, and the methods used for nonresponse adjustment.

Throughout this paper, the value of adopting a model-based approach to nonresponse adjustment clearly shines through. By constructing an intelligent model for the nonresponse mechanism, one is able to carry out a nonresponse adjustment using many more explanatory variables than would otherwise be possible using a more traditional approach. In a more traditional approach, one would form adjustment cells by crossing the classes of every explanatory variable. This would be equivalent to building a response-propensity model that includes all possible interactions among the explanatory variables, whether or not those interactions have much predictive power (and they often don't). The modeling approach adopted by BMP allows them to exclude unimportant high-order interactions, and instead include main effects for a larger number of explanatory variables.

One issue that may deserve a little more attention is what should be done with the estimated response propensities once they are calculated. On this point, statisticians north of the U.S.-Canada border tend to use the reciprocals of these probabilities as factors in the nonresponse weighting adjustment. Statisticians south of the border tend to form classes--e.g. by dividing the estimated propensities into quintiles--and reweight the observations within these classes. Little and Rubin (1987) comment that the latter may sacrifice a little bias for the sake of reduced variance and robustness against model failure. I wonder if anyone has done a comparison of the two methods in a realistic setting to see which one tends to perform better.

Another important issue, which is perhaps beyond the scope of the BMP paper, relates to the underlying philosophy of response-propensity weighting. Response-propensity weighting attempts to control and reduce nonresponse bias. The theory of propensity scores (Rosenbaum and Rubin,

1983) says that if reweighting could be performed on the basis of the actual (as opposed to estimated) propensity scores, then the reweighted distribution of the respondents would not be systematically any different from that of the respondents--in other words, nonresponse bias would be eliminated. But of course, bias is only one component of error, the the other being variance. As pointed out by Little (1986), response-propensity weighting may do very little to control variance. One might also want to consider forming weighting classes on the basis of variables that are highly correlated with the survey variable of interest (if any are available), so that variance might also be reduced. Perhaps forming weighting classes on the basis of two variables--a linear predictor of the response propensity, and a linear predictor of the survey variable of interest--may be a reasonable approach.

The methods described in the BMP paper also raise a number of theoretical issues that deserve a closer look by statisticians in the future. One question involves the use of complex, automated variable-selection procedures to choose a model for nonresponse adjustment. Whenever the form of the model is chosen through examination of the sample data, the procedure used to select the model should be considered a part of the overall method of nonresponse adjustment. Any calculations of bias and variance--whether carried out analytically, or by simulation, jackknifing, etc.--should recognize that the model itself is sample-dependent and therefore random, and the model selection procedure must therefore be included in the calculation.

Another, perhaps more basic, issue pertains to the criteria used for selecting the model. BMP emphasized the principle of parsimony, eliminating variables whose effects were not significantly different from zero. They also included variables tended to reduce the number of extreme residuals. In the end, they were left with models that had very few parameters relative to the size of the dataset. I was left with the feeling that they could have included more variables, provided that the model-fitting could be accomplished in a reasonable amount of time. The usual criteria given by textbooks on regression modeling--high R^2, low prediction error, all coefficients statistically significant, and so on--are usually appropriate when the goal is to acquire some scientific understanding of how the response variable is related to the pool of potential predictors. When the goal is not necessarily scientific understanding, but adjusting for nonresponse, however, it is not yet clear what model-selection criteria statisticians ought to be using.

Finally, another issue that deserves further investigation is the proper

role of sample-design information in the construction of response-propensity models. I suspect that many statisticians would attempt to include design information by fitting logistic regressions with standard software, including the case weights (inverses of the sample-selection probabilities) in the fitting procedure. The correctness of such a procedure is not at all clear. If the goal were to estimate regression coefficients for predicting nonresponse for the entire population, then including the case weights would be appropriate. The goal in response-propensity modeling, however, is to estimate the probability of response for the units in the current sample. To the extent that this response propensity is related to covariates describing the sample design (e.g. stratum or cluster indicators), those covariates ought to be included in the model somehow. But merely weighting the cases by their inverse probability of selection is probably not sufficient to guarantee that the special features of a dataset that arise from complex sampling, such as clustering effects, are appropriately described.

Additional references

Little, R.J.A. (1986), "Survey nonresponse adjustments for estimates of means," International Statistical Review, 54, 139-157.

Rosenbaum, P.R. and Rubin, D.B. (1983), "The central role of the propensity score in observational studies for causal effects," Biometrika, 70, 41-55.

TOWARD AN AGENDA FOR THE FUTURE

OLD DIRECTIONS, NEW DIRECTIONS: CONCLUDING REMARKS

Norman M. Bradburn
NORC

I have been asked to take on the unenviable task of giving an overview of the last two days and, on the basis of that overview, to give some thoughts on future directions. In order to have some empirical idea of what the directions--new or old--have been, I looked at the topics covered in this symposium and compared them with the topics covered in the first symposium four years ago.

If we view the programs of the conferences as unobtrusive measures of what are the central research directions in statistical methodology, at least in the minds of the organizers of these conferences, we can get an idea of how things have changed in the last four years. For comparison, I coded the session topics into four themes: 1) conceptualization of statistical measures--that is, what we are measuring; 2) design and analysis of statistical surveys; 3) operational issues in data collection; and 4) dissemination of data to users. The result of the comparison is seen in Table 1.

Table 1

Themes of Papers Given at 1990 and 1994 Seminars

	1990	1994
Conceptualization	0	1
Design/Analysis	5	4
Operations	6	5
Dissemination/Use	1	2
Total	12	12

Six (or seven) of the 12 sessions in each year are on the same topics. They are: 1) Administrative Records, 2) Longitudinal Surveys, 3) Data Editing, 4) Disclosure Limitation, 5) Computer Assisted Data Collection, 6) Cognitive Testing of Questions. The seventh, Incomplete Surveys, could be considered at least partially the same topic, i.e., Coverage Problems in 1990 and Non-Response in 1994.

Without looking at the new topics, one might view this as little change; but when we look at the particulars of those new topics, it seems clear that there are major shifts in direction. New session topics were: 1) Customer Surveys, 2) Respondent

Incentives, 3) Small Area Estimation, 4) Time Series revisions, and 5) Economic Classification Revision.

What is it about the new topics that leads me to think we are moving along some significant new directions? I see the primary theme of these new topics as the need to rethink some old problems in order to cope with three types of change: 1) changes in social conditions, 2) growth in demand for data and 3) technological change. Let me comment on these in turn.

1. **Changes in social conditions.** Society is becoming more diverse along many lines. It is not just the increasing cultural diversity that is so much remarked upon but also the increasing diversity in household structures, in a changing occupational structure, and in the globalization of the economy. Our traditional ways of collecting data and the statistical categories we have used are creaking and in need of revision. Surveys are harder to carry out, so we become concerned about questionnaire development and incentives for respondents. It is even necessary to rethink how we conduct the decennial census. Our classification schemes for economic and social categories are no longer adequate, so we rethink the SIC and SOC and the problems of revising time series when you change the measuring instruments.

2. **Growth in demand.** It is a truism these days that we live in an information society. The demand for data grows apace with the proposals for health care reform and the health data network and the increasing demand for evaluation of social services and social programs. Concern for international economic competitiveness leads us to rethink the categories we use to collect economic statistics, such as how we measure the trade balance and what industrial and occupational classification scheme we use. It is important to have an economic statistical system that is consonant with that of other countries.

There are two implications for statistical methodology of this increase in demand for data. The first is that it requires a strong customer orientation to be responsive to the changes in demand. We need to know what the consumers of statistical information require in order to carry out their work efficiently. Second, because of the high cost of data collection in a world of tight budgets, we need to develop alternative ways of doing things--not just tinkering around the edges, but some fundamental changes ("reengineering," to use the current buzz word).

3. **Technological change.** CATI, CAPI, CASI, Audio-CASI. computer-assisted everything. The ability to handle large data files cheaply has changed the economics of data analysis, if not of data collection. Data are accessible on Internet; you can call in for SIPP data. All of these advances whet the appetite for more data--more quantitatively, in more detail, and more easily accessible.

At the same time, increased availability of large data sets increases the risk of disclosure of data that compromises the privacy and confidentiality promised to respondents and upon which our ability to get accurate data depends. Confidence in the statistical system must be maintained, and that means preventing disclosure on individually identifiable data for non-statistical purposes. Hence our renewed concern for data disclosure limitation principles and methods.

Where does all this change lead us? First, I think it leads us to a fundamental rethinking of some of our major data systems. We have already seen the revision of the CPS. We have heard here about proposed revisions in the Standard Industrial Classification system and the Standard Occupational Classification. These are truly major changes in our statistical thinking.

Work has begun on a revision of the CPI. We are rethinking the way we conduct the census. Dare we think of an administrative records enumeration in 2010 with the long form data supplied by continuous measurement? Can the national accounts be recast to include supplemental accounts for the environment (so-called Green Accounting) and recast for other non-market sectors as well? Will data bases become public utilities that can be used for many purposes? For example, the basic health care file, if it were to come to pass, could be a prime source of data for an administrative records census or a data base for assessing the outcomes of health care services. A Master Address File that was available to everyone, the Census Bureau, the Post Office, state and local governments, the private sector, would be a true common good.

A final theme that cuts across all of our sessions here is that we must take a more international perspective. We can learn from other countries that have been confronting some of the same problems. There is also increased pressure to have consistent statistical measures across the developed countries at least, so that international comparisons can be made.

The spur of competition can be good. The State governors adopted a set of Educational Goals for the year 2000. One of these is for the United States to be first in science and math education by the year 2000. In the Economist magazine's annual rankings of governmental statistical systems, the United States has ranked about 5th for the last few years. Perhaps we should take as our goal to replace Statistics Canada as first in the world among statistical systems by the year 2000.

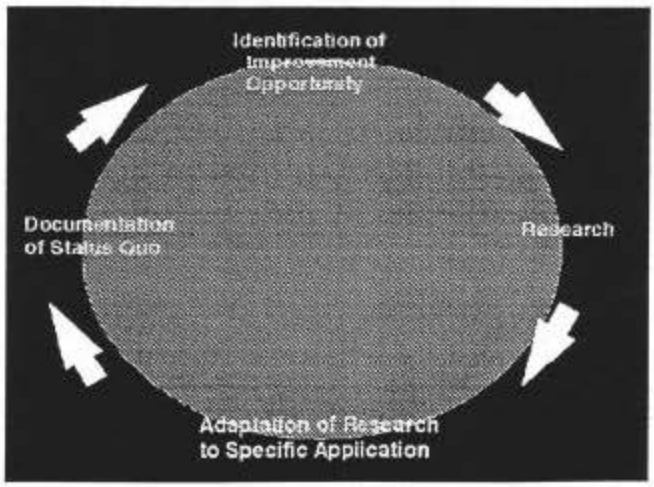
Future directions? From what we've heard here, we may not be able to know precisely how to proceed, but we will have to move forward with an international perspective, a willingness to rethink our methods, our tools, and our purposes.

Toward an Agenda for the Future

Robert M. Groves
University of Michigan and Joint Program in Survey Methodology

It is obvious that the rate of change of the society and economy being measured by statistical agencies is increasing. We have only to observe the ubiquity of merger, acquisitions, and turnover of firms, partly a reflection of the ongoing creation of novel combinations of services and goods. In the demographic domain, we are faced new living arrangements, nontraditional household definitions, and homelessness at levels of occurrence that challenge traditional methods of measurement. It now seems clear that the rate of change of social and economic measurement systems needs to approximate that in the society. We have heard over the last two days much evidence of this.

Figure 1 is a version of charts that are popular among those in the quality movement. It is used to describe the process of continuous improvement key to that philosophy. As "reinvention," "restructuring," "re-engineering," "customer orientation," "total quality management," and related concepts make their way into statistical agencies, it might be useful to reflect how the technical research and development presented over the last two days relate to the process described in such a chart.



The process begins, even prior to the contents of this chart, with the identification of customer needs and desires. Given that as a basis, first the organization needs to document what it does and how the processes of the organization are now conducted. This documentation leads to identification of potential areas of improvement, "identification of improvement opportunity." Whether the process can or cannot be improved is a matter to be resolved in a "research" step. Then the results of the research are applied in "adaptation of research to specific application." At this point, the process begins again, this time with the production process operating at a higher level of overall efficiency, but still subject to improvement.

One aspect that differentiates research in government statistical agencies from research in academia is the last step, application of research findings in practical settings. This fact is accompanied by organizational roles that perform these activities in conjunction with the researcher. The researcher is joined by a manager, who is charged with the responsibility of directing the processes undergoing change.

It's useful to review each of the four steps from the vantage point of a program manager of a data series and a researcher or scientist, both located in a government statistical agency.

Documentation

Documentation is often viewed as the bane of those who develop and conduct data collection, processing, and estimation programs. It is most often not done (or not done completely) for many statistical series. When done, it is often done long after the introduction of a new process, and then completed by someone who was not part of the creation of the process. The lack of good documentation on design, processing, and estimation features of statistical series can be a nontrivial impediment to their improvement. For example, when there is clear evidence that a process is not performing as desired (e.g., nonresponse rates too high, edit processes yielding contradictory data), improvement is often delayed by lack of clear documentation of exactly what is being done. When lack of documentation is joined by high turnover among staff and computer based systems, it is often the case that workers implementing a computer-based process truly don't know what they are doing.

From the management perspective, documentation allows knowledge of what processes are the staff are performing. It permits the manager to look for opportunities to streamline processes, combining steps and improving flow. For the researcher or scientist, documentation is the equivalent of gathering observations about possible causes of outcome variables. It is the inductive part of the scientific method -- assembling information that suggest hypotheses for sources of weaknesses (and thereby opportunities for improvement).

Why is documentation an important step in the process of continuous improvement? Improvements on undocumented systems are limited to those performing the *status quo* activities because only they know what is being done. This limits the set of persons who can be called upon for suggestions for improvement. This limitation is exacerbated in organizations where staff implementing programs do not have rudimentary testing and experimentation skills. It seems clear that without documentation of best practice,

agencies can't get to "the cutting edge" that the benchmarking aspect of the quality movement espouses.

Opportunities for Improvement

The use of Pareto curves prompts managers to focus on the problems that are causing the greatest loss of quality or efficiency in production. This is a vehicle to reallocate budgets to save money on expensive processes by investing money in design and retooling phases. Given most government budget systems, an investment in a research and development project to improve one component of a statistical series generally means that some other components will incurred reduced funding. The manager needs the courage of his/her convictions that such an investment can lead to cost efficiencies at a later point.

For the scientist the "identification of improvement opportunities" step often also means a search for experiments on design features common to many surveys. Often important weaknesses of one statistical series are shared by other statistical series. If the researcher can identify such cross-series problems, then single research projects can offer improvements to several data series at once.

In one sense, this seminar can fulfill a similar function. If one agency discovers improvements that are applicable to the work of another, significant research savings are possible.

Research

The research step from the management perspective is a search for independent evidence that change involving risk was well-founded. The research step tests some model of the real change in the production system. To the extent that the research step perfectly mimics the real production setting, the inference from the experiment will apply to the production process. In this sense, research is a risk reduction tool for the manager. Change without research runs higher risks of no improvement or even loss of quality in statistical series.

For the scientist, only experimentation offers the proper grounding for application of some discovery. The scientific method underscores the need for explicit contrast under similar conditions of the new process with the old, before recommendations about change can be well-founded. Much research in social and economic statistics is "applied," with its findings relevant to ongoing statistical series. Applied research is the type most often practice in statistical agencies (and reported in conferences like this one). The cheapest research (in the long run) is theory-driven, often motivated by basic research. This is the research that offers solutions for large sets of surveys and census operations. Choosing the right blend of support for basic and applied research in government agencies takes great wisdom. One without the other is vacuous.

Adaptation of Research to Specific Application

This step is typically the focal point for program managers. From the standard managerial perspective, the real work (and associated risks) of change begin here.

The issues of concern to the manager always concern whether the costs of the change are smaller than the benefits. Sometimes real courage on managers' parts is required in organizational cultures where failed change is punished by withdrawal of the support of superiors.

From the scientific perspective, this step typically requires speculation about the inferential limitations of the research into alternative solutions. Scientists are generally well-trained in hypothesis generation, but not particularly well-schooled in decision-making with imperfect information. In the world of statistical agency programs, there is rarely or never perfect information. Indeed, in any application of a scientific discovery there is some inferential leap that goes beyond the research. Many scientists are uncomfortable with such leap, but scientists in statistical agencies, to be responsive and useful to their agencies, must develop these skills. Clearly, this is most easily accomplished when they work in partnership with their colleagues on the program side.

All of the steps in Figure 1 are difficult, but this step seems most fraught with difficulties and failure to communicate between the managerial and the scientific cultures within statistical agencies.

Implications for Future Symposia, the Working Paper Series, the System

What do these observations have to do with a symposium like the one we have all just experienced? Let's examine the content of this symposium. There was clearly more attention on documentation and adaptation of research to applications, than on methods used to identify opportunities for improvement in data series, or in the findings of more basic research. This is an appropriate mix, I believe, for the purposes of the conference - - an attempt to disseminate new developments of applied utility.

Given that focus, future symposium might reconsider the format of a scientific conference, where a researcher presents his/her work and a discussant critiques the scientific merits of the work. Perhaps for the purposes of application of new discoveries, workshops and didactic seminars might be more useful.

One method of doing that is to use the Federal Committee on Statistical Methodology to identify agencies using "best" practices in various components of statistical work, and to encourage them to mount such workshops, perhaps with some collaboration with scientists and practitioners outside the Federal statistical system.

It is clear that the Federal statistical system can profit from the successes of integrating science and management, of adapting research discoveries to practice that exist in some agencies, but need to be spread more widely across all agencies. What the system does not need is more papers assessing ideas alone. It does need more papers describing how new ideas were assessed, molded to a specific problem, and systems changed to incorporate them. Problems of statistical system are not solved by ideas alone, but by ideas implemented to produce innovation in data series. Presentations in future symposia might be collaborations of program managers and scientific staff, addressing the four steps of Figure 1 for a specific survey or set of statistical activities, from their two

different perspectives.

I feel more strongly that the symposium should continue annually, than that it should take on a particular format or offer particular content. Regardless of its format, it is a gathering of the clan within the statistical system, allowing them to meet and converse with one another, to compare techniques, to try out new ideas removed from their organizational cultures, to renew their commitments to improved quality in statistical activities. This will always have value.

TOWARD AN AGENDA FOR THE FUTURE

Katherine K. Wallman
Office of Management and Budget

Over the past two days, we have heard many ideas and suggestions for how we can build upon and improve the efforts we are undertaking through the Federal Committee on Statistical Methodology (FCSM). The "new directions in statistical methodology" we might pursue range from updating and expanding our technical contributions to broadening and deepening the understanding and application of our work. Underlying all of the proposals are two fundamental and, I believe, shared perceptions: first, that the FCSM's activities to date have been of considerable value, and second, that there are opportunities we should pursue to further the FCSM's goal of improving the quality of data produced by the Federal statistical system for our community of users within and outside the government.

In his keynote address for this seminar, Graham Kalton reminded us of the four functions for the FCSM that Margaret Martin outlined at our 1990 Seminar on the Quality of Federal Data. These functions do indeed serve as a useful framework for considering the committee's future agenda, and I would like to return to them in highlighting in particular a few of the suggestions that have been made for ways we might broaden and deepen the understanding and application of the FCSM's work.

The first of these functions is the exchange of knowledge, techniques, and experience among committee members. As has been noted, the members of the Federal statistical community who serve on the FCSM and its subcommittees often carve out time for these activities from their already overly full calendars. The result may well be that they must focus rather narrowly on the task at hand and may forego the opportunity for more informal dialogue. We need to think of ways to give these professionals "permission" to engage in less structured conversation. As a first step, the leaders of the statistical agencies should be encouraged to view the work of the FCSM as a priority for their staff members. In part, we may be able to address this matter by familiarizing the agency leaders more fully with the committee's products.

The second function outlined for the FCSM is the provision of "state of the art" reports to encourage best practice among a broader group. With respect to the "state of the art" aspect of the committee's work, I think there is some considerable agreement that while an FCSM report may meet this challenge when it is issued, in some cases the art is changing so quickly that a working paper may become if not obsolete at least outdated more rapidly than we would hope. Perhaps we should consider new forms of publication -- "loose leaf" printed versions and/or electronic versions that could have sections updated, rather than waiting

several years to revise the entire paper. The challenge of sharing our work with a broader group is key. On the one hand, as Graham Kalton elaborated, we need to reach out beyond our Federal walls in developing our working papers to capitalize on the expertise resident in the private sector, foreign governments, and other arenas. His remarks provide a number of useful suggestions the FCSM should pursue. On the other hand, it is absolutely essential that we continue to strengthen efforts that have been undertaken to disseminate the FCSM's products far more widely, both within and outside the Federal Government. This means not only distributing our product to additional audiences, but also extending the seminars, workshops, and other forums where the subject matter of the papers can be explored in greater detail and become more useful to those less familiar with the content.

In terms of the third function suggested by Margaret Martin, recommending areas for improvement and needed directions for research, once again the need for greater outreach has been highlighted, and some useful paths to pursue have been suggested. At the same time, we need to take some care in reaching out that we do not ask the FCSM to be "all things to all people." As has been noted, the members' plates are quite full already, and we should not overload them. Yet there should be new ways to meet reasonable demands, and we as a community should put some of our energy into brainstorming about alternative ways of operating to address these needs.

The consensus building role suggested as a fourth function for the FCSM is surely an area where the committee can make a contribution. In my view, however, to suggest that the FCSM should be responsible for obtaining consensus on issues such as definitions, concepts, and classifications moves us to an arena that is beyond its mandate. Many of these activities involve domains of expertise beyond the specialties of the FCSM membership, as well as exceptionally labor intensive tasks. Through the cooperation of the statistical agencies and other relevant components of the Federal Government, we are pursuing a number of these challenges. We would welcome the suggestions for further work in this area that might arise as a result of the FCSM's deliberations.

In closing, let me note our appreciation to the Council of Professional Associations on Federal Statistics for bringing us together once again, to the session organizers and presenters for their substantial contributions to the success of this seminar, and last but not least to the members of the FCSM and its devoted leader Maria Gonzalez. We look forward to planning an encore to be held in 1996.

Reports Available in the Statistical Policy Working Paper Series

1. Report on Statistics for Allocation of Funds (Available through NTIS Document Sales, PB86-211521/AS)
2. Report on Statistical Disclosure and Disclosure-Avoidance Techniques (NTIS Document sales, PB86-211539/AS)
3. An Error Profile: Employment as Measures by the Current Population Survey (NTIS Document Sales PB86-214269/AS)
4. Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics (NTIS Document Sales, PB86-211547/AS)
5. Report on Exact and Statistical Matching Techniques (NTIS Document Sales, PB86-215829/AS)
6. Report on Statistical Uses of Administrative Records (NTIS Document Sales, PB86-214285/AS)
7. An Interagency review of Time-Series Revision Policies (NTIS Document Sales, PB86-232451/AS)
8. Statistical Interagency Agreements (NTIS Documents Sales, PB86-230570/AS)
9. Contracting for Surveys (NTIS Documents Sales, PB83-233148)
10. Approaches to Developing Questionnaires (NTIS Document Sales, PB84-105055/AS)
11. A Review of Industry Coding Systems (NTIS Document Sales, PB84-135276)
12. The Role of Telephone Data Collection in Federal Statistics (NTIS Document Sales, PB85-105971)
13. Federal Longitudinal Surveys (NTIS Documents Sales, PB86-139730)
14. Workshop on Statistical Uses of Microcomputers in Federal Agencies (NTIS Document Sales, PB87-166393)
15. Quality on Establishment Surveys (NTIS Document Sales, PB88-232921)
16. A Comparative Study of Reporting Units in Selected Employer Data Systems (NTIS Document Sales, PB90-205238)
17. Survey Coverage (NTIS Document Sales, PB90-205246)
18. Data Editing in Federal Statistical Agencies (NTIS Document Sales, PB90-205253)
19. Computer Assisted Survey Information Collection (NTIS Document Sales, PB90-205261)
20. Seminar on the Quality of Federal Data (NTIS Document Sales, PB91-142414)
21. Indirect Estimators in Federal Programs (NTIS Document Sales, PB93-209294)
22. Report on Statistical Disclosure Limitation Methodology (NTIS Document Sales, PB94-165305)
23. Seminar on New Directions in Statistical Methodology (NTIS Document Sales, PB95-182978)

Copies of these working papers may be ordered from NTIS Document Sales, 5285 Port Royal Road, Springfield, VA 22161 (703)487-4650