Does Length Really Matter? Exploring the Effects of a Shorter Interview on Data Quality, Nonresponse, and Respondent Burden

Scott Fricker, Brett Creech, Jeanette Davis, Jeffrey Gonzalez, Lucilla Tan, Nhien To
Bureau of Labor Statistics

March 16, 2012

1. INTRODUCTION

1.1 Background

The U.S. Consumer Expenditure Survey is an ongoing monthly survey conducted by the U.S. Bureau of Labor Statistics (BLS) that provides current and continuous information on the buying habits of American consumers. The Consumer Expenditure Survey consists of two independent components: The Quarterly Interview (CEQ) Survey and the Diary (CED) Survey. For the CEQ, interviewers visit sample households five times over the course of thirteen consecutive months. Each interview is conducted with a single household respondent who reports for the entire household. The first interview establishes cooperation, collects demographic information, and bounds the interview by collecting expenditure data for the previous month. This 'bounding' interview is designed to limit forward telescoping, which is the process by which respondents remember and report events or purchases as taking place more recently than they actually occurred. The four remaining interviews are administered quarterly and ask about expenses incurred in the 3-month period that just ended.

The CEQ survey presents a number of challenges for both interviewers and respondents. The interview is long, the questions detailed, and the experience can be perceived as burdensome. In part because of these challenges, there is a widespread belief that some CEQ data are underreported. Underreporting has been variously attributed to recall error, panel conditioning, respondent fatigue, satisficing, and other causes. The length and perceived burden of the CEQ survey may also have deleterious effects on response rates.

1.2 Study Objectives

This study was the first in a comprehensive and ongoing effort to examine alternative data collection strategies for the CEQ that may improve data quality, maintain or increase response rates, and reduce data collection costs. In particular, this study assessed the effects of administering a shorter CEQ questionnaire on respondent burden, data quality, and nonresponse error. A separate condition in this study examined the extent to which using a 1-month (versus a 3-month) reference period affected underreporting due to recall errors. The study design enabled BLS to perform data quality analyses using both direct measures (e.g., number of expenditure reports, expenditure amounts) and indirect measures (e.g., response rates, measures of perceived burden, item nonresponse, etc.), and to estimate

nonresponse bias by comparing response rates, sample composition, and expenditure estimates across treatment conditions. The results from this study will be used to inform future CEQ research activities and decisions about how to redesign the production survey.

2. PREVIOUS RESEARCH

2.1 Survey Length

Survey organizations routinely limit the length of their surveys under the assumption that longer surveys can negatively impact a number of survey quality outcomes. The empirical literature examining this issue has primarily focused on the effect of length on nonresponse, and the results from these studies are mixed. For example, some studies have found that longer surveys or more frequent survey requests decrease response rates (e.g., Collins et al., 1988; Dillman et al., 1993), increase drop-out rates (e.g., Haraldsen, 2002), and reduce respondents' willingness to respond to future surveys (e.g., Apodaca et al., 1998; Groves et al., 1999). In contrast, other studies have found that longer interviews are associated with higher response rates and panel-survey sample retention (e.g., Champion and Sear, 1969; Branden et al., 1995) or have no association at all (e.g., Sharp and Frankel, 1983; McCarthy, Beckler, and Qualey, 2006). Although data conflict regarding whether survey length increases various forms of nonresponse, the evidence *in toto* suggests that there is at best a weak positive association. One reason for these equivocal findings is that respondent motivation to participate is affected not only by length, but also by a variety of other factors such as topic interest or the survey sponsor.

Motivation may additionally affect data quality more broadly. Individuals' motivation to respond in a thoughtful manner may decrease over the course of a long survey due to respondent fatigue or boredom. Although there has been less empirical attention to the impact of survey length on data quality than nonresponse, several studies provide evidence that respondents in longer surveys have greater likelihood of straight-line responding (Herzog and Backman, 1981), increased rates of item-nonresponse (Galesic, 2006; Peytchev and Tourangeau, 2005), and provide fewer survey reports (Backor, Golde, and Nie, 2007) than those in shorter surveys. Data quality also has been shown to deteriorate over the course of an interview, with increases in item nonresponse, 'don't know' reports, and response order effects, and less time spent on each

question, the longer the duration of the interview (e.g., Krosnick, 1999; Peytchev, 2005; Roberts *et al.*, 2010). Taken together, these findings corroborate the received wisdom that interview length should be kept to a minimum, both to avoid the potential for nonresponse and to reduce satisficing behavior that can jeopardize survey data quality.

2.1.1 Split Questionnaires

The practical reality is that some surveys are excessively long, and often it is not feasible to simply cut items from a questionnaire to achieve reductions in respondent burden. Survey organizations may need to ask a large set of questions to meet stakeholder analytic objectives and to accommodate periodic requests to add new questions to an existing instrument. One method that has been developed to shorten surveys while still achieving the analytic needs of the organization is the use of split questionnaires (also referred to as multiple matrix sampling; see, e.g., Raghunathan and Grizzle, 1995). In one implementation of a split-questionnaire survey design, the original survey is divided into one 'core' component containing high-priority questions (e.g., socio-demographic variables) and a number of subcomponents containing approximately equal numbers of the remaining items. The full survey sample is likewise split into distinct subsamples, and each subsample of respondents completes the core component plus a randomly assigned subcomponent. Figure 1 provides an illustration of a split questionnaire survey design with a core component and three subcomponents.

Figure 1. Split Questionnaire Design with Three Components

Respondent Subsample	Questionnaire Split			
	Core Component	Subcomponent A	Subcomponent B	Subcomponent C
A				
B				
C				

Split questionnaire designs reduce the length of the survey while still collecting the necessary information from at least some of the sample members, but they also result in missing data. The goal is to minimize the amount of information lost relative to the complete questionnaire, and appropriate decisions must be made at various phases of the survey process to aid optimal implementation and estimation. Survey designers must determine how to best construct the questionnaire splits (e.g., random allocation of items to subcomponents; grouping logically-related items within a subcomponent; distributing highly correlated items to different

components; use of a core component or not). They must decide which subset of the full sample will receive any given questionnaire component(s) (e.g., through random or predictive assignment). And they must select techniques for analyzing the resultant data (e.g., available case method; single imputation; multiple imputation; adjustments to calibration weighting). Gonzalez and Eltinge (2007a; 2007b; 2008; 2009) provide in-depth treatments of these issues and the foundational discussions of design, implementation, and analysis considerations for a potential application of split questionnaire design to the CEQ.

A relatively small but growing literature suggests that carefully implemented split questionnaire designs can be effective in producing key population and subpopulation estimates, and for reducing respondent burden, compared to full questionnaires. For example, Navarro and Griffin (1993) investigated the possible application of this approach for the 2000 Decennial Census, and found that it achieved adequately reliable small-area population estimates as well as reductions in respondent burden. The seminal paper by Raghunathan and Grizzle (1995) demonstrated that a split questionnaire design, coupled with multiple imputation to produce a complete dataset, could obtain estimates (means and regression coefficients) similar to those derived from the full dataset. And more recently, Wedel and Adiguzel (2008) found that a split questionnaire design yielded parameter estimates that were very close to the complete-data estimates, and that respondents who were administered split questionnaires had more favorable reactions to the survey (e.g., shorter perceived duration, lower ratings of boredom and fatigue, etc.) than those who received the complete questionnaire.

2.1.2 Use of Global Items

For surveys that ask a series of detailed questions about a given topic – as the CEQ does for household purchases across a variety of expenditures categories – another option for shortening the length of the interview is to replace some of the detailed questions with global items. Global questions ask about topics at a more aggregated level. For example, rather than asking separate questions about how much a household spent on shoes, pants, shirts, jackets, etc., a global question might simply ask what was spent on clothing, full stop. Global questions could replace detailed questions for the entire sample or for subsets of the sample as way to collect some information on expenditures without imposing the burden associated with administering the full

set of detailed items. The information obtained from the global reports could then be used to derive estimates at a more detailed level (e.g., in a split-questionnaire design in which respondents' global answers are used as inputs to imputation models).[1]

The gains achieved by the use of global questions (i.e., reductions in survey length and potentially respondent burden) have to be considered against the loss of detailed information and its impact on the needs of the survey stakeholders. In addition, the decision to use global questions needs to be informed by an understanding of their impact on respondent error. For example, because global questions lack the specificity of their more detailed counterparts, they may also fail to provide the definitional clarity and retrieval cues required to elicit full and accurate responses (e.g., Conrad and Schober, 2000; Dashen and Fricker, 2001; Hubble, 1995), which in turn may actually increase respondents' burden. Additionally, some studies suggest that global questions produce overestimates and are less reliable than more detailed questions (e.g., Battistin, 2003; Zimmit, 2004). On the other hand, there is empirical evidence that decomposed (detailed) questions also can lead to increases in measurement error, for example when the granularity of the question does not match the way events are stored in memory, or when respondent fatigue induces satisficing (e.g., Belli *et al*, 2000; Menon, 1997; Shields and To, 2000). As these results suggest, the effectiveness of global questions will vary because information about different topics is encoded and stored in memory in different ways (e.g., depending on its salience, frequency of occurrence and use, and its contextual associations).

2.2 Reference Period

The selection of the length of the survey reference period ideally should be based on a number of factors. First, there are analytic and operational considerations. Survey designers must consider the operational costs associated with different reference period designs, and examine these designs in light of the required levels of precision of the estimates. For example, shorter reference periods may necessitate more frequent interviews, which under a fixed budget would result in smaller sample sizes and less statistical precision. Additionally, if shorter reference periods result in more frequent interviews, respondent burden and the likelihood of survey

[1] Notwithstanding their treatment in subsequent sections of this report, global items likely would not be used by CE as simply direct substitutes for detailed items. However, an in-depth exploration of imputation models incorporating global reports was beyond the scope of this project.

nonresponse may increase (e.g., Bradburn, 1978; Apodaca *et al.*, 1998). Second, designers should understand how information on topics covered in the survey is encoded and structured in respondents' memory. Finally, ideally there needs to be an awareness of the error properties associated with the response processes under different reference period implementations, with the selection of a reference period that minimizes those errors.

The literatures on memory, cognition, and survey response processes indicate that for most surveys no single reference period will be optimal for all items. Memory decays over time and, on the whole, short reference periods may improve recall relative to long reference periods (e.g., Miller and Groves, 1985). But, forgetting occurs at different rates for different events (e.g., Bradburn, Rips, and Shevell, 1987). Respondents tend to forget events that are infrequent, irregular, or not salient, all else being equal (e.g., Menon, 1994), so shorter reference periods should aid recall of these events. Longer reference periods may be more appropriate when asking about salient events or regular events that vary little across time or about which the respondent has abstracted and stored some generalized information (e.g., *I usually spend $10 at the laundromat on Fridays*.). In either case, as noted earlier, the granularity of the survey question should match the information stored in respondents' memory. For example, if a question using a short reference period asks respondents to enumerate and provide information about individual events but respondents' memory reflects more aggregated, summary-level information about those events (e.g., as is common for frequently occurring but mundane purchases), reporting errors can occur.

The length of the reference period can also impact another source of recall errors known as *forward telescoping*. In forward telescoping respondents erroneously report events as having occurred during the reference period when in fact they occurred prior to it, a phenomenon that generally leads to overreporting[2]. This effect may be caused by respondents misperceiving the length of the reference period (e.g., respondents given a 3-month reference period actually think about the last 3.5 months) or uncertainty about when a target event occurred (Tourangeau, Rips, and Rasinski, 2000). As with errors of omission, there is not a simple relationship between reference period length and telescoping errors. Early models suggested that shorter reference

[2] The countervailing effects of backward telescoping – placing in-scope events outside the reference period – are thought to be weaker than those of forward telescoping because memory for older events generally has greater temporal imprecision than memory for more recent events.

periods should reduce recall loss and increase forward telescoping, but a number of studies provide evidence that the occurrence and magnitude of telescoping depends on respondents' ability and motivation, the specificity of the question format, the salience of the event being recalled, and the availability of additional temporal cues such as those provided in bounding interviews (e.g., Groves *et al*., 2004; Neter and Waksberg, 1964; Prohaska, Brown, and Belli, 1998). In one of the few telescoping studies that used verification data to check the veracity of respondents' reports, Huttenlocher, Hedges, and Prohaska (1988) found no differences in the amount of forward telescoping for 'long' and 'short' reference periods (an academic year and a academic quarter, respectively). Respondents in this study were more accurate when reporting for the shorter reference period, but the authors attributed this finding to steeper forgetting curves for older events and to more effortful memory search under the shorter reference period.

3. OVERVIEW OF THE PRESENT STUDY

3.1 Study Design Issues

The primary purpose of this study was to investigate the effects of shortening the length of the CEQ interview (by implementing a split questionnaire design that incorporated global questions) and the length of the CEQ reference period on survey nonresponse, data quality, and respondent burden. In order to achieve these objectives, staff from the BLS Branch of Research and Program Development (BRPD), Office of Survey Methods Research (OSMR), and Branch of Information and Analysis (BIA) formed the Measurement Issues Study (MIS) Team to plan, implement, and analyze data from a small-scale field test of a modified CEQ.

The study utilized a basic experimental design in which respondents were randomly assigned to a control group which received no treatment, a test group that was administered a shortened version of the questionnaire, or a test group that was administered a shortened reference period. As with any field test, an effort was made to mirror as many of the CEQ survey procedures and conditions as possible (e.g., use of a panel design that incorporated a bounding interview, use of CEQ questions and materials), but there were a number of significant departures. First, the budget for the project prevented in-person data collection so we relied instead on centralized computer-assisted telephone interviews

(CATI). As a consequence of changing from in-person to phone-administered interviewing, the Team decided to shorten the overall length of the survey because of concern that study respondents would not accept a phone survey lasting 60 minutes or more on average. To do this, we eliminated questions about a number of CEQ expenditure categories to develop a basic study instrument with a completion duration target of 30 minutes (less for the treatment groups)[3].

Another procedural departure for the MIS was that respondents in the shorter reference period condition were interviewed once a month, not once a quarter as in the current CEQ (and in the MIS control group). This was a necessary consequence of the study objectives, since we wanted to aggregate data from the three monthly interviews that used the 1-month reference period and compare those estimates to estimates derived from the control group's standard 3-month reference period, as well as examine potential differences between the control group and this treatment group in nonresponse and respondent burden. Finally, for the shortened interview treatment group we implemented a basic split questionnaire design. We divided our full study questionnaire (i.e., the 30-minute 'basic' version) into one 'core' component and two subcomponents, split the treatment sample into two random subsamples, and then administered each subsample the core component plus one of the subcomponents. In addition to the detailed expenditure questions in their assigned subcomponent, respondents were asked a smaller number of global expenditure questions to augment the loss of detailed information from the remaining expenditure categories (i.e., those covered in the unassigned subcomponent). This allowed us to derive section-level expenditure estimates for all expenditure categories for both of the shortened interview subsamples, and to examine whether the global items produced data of sufficient quality to replace detailed questions.

3.2 Defining the Key Study Outcome Concepts and Measures

The MIS was designed to shed light on three key concepts – respondent burden, data quality, and nonresponse error – defined as follows.

Respondent Burden – Bradburn (1978) identifies four factors that contribute to respondent burden: (1) length of the interview; (2) effort required by the respondent; (3) amount of perceived stress experienced by the respondent; and, (4) periodicity of the interview. We administered

[3] Additional details about this and other design issues can be found in the Method section of this report.

questions covering these four factors and used respondents' answers to determine the effect of each of a shorter questionnaire and a one-month reference period on respondent burden.

Data Quality – The CE Program Office operates under the premise that "more is better," suggesting that respondents who report more expenditures (in terms of both number of items and absolute dollar amount) have higher data quality than those who report fewer expenditures. In the survey methodological literature, the term "data quality" often is used to refer to multiple error sources (e.g., measurement and sampling) and dimensions (e.g., timeliness and accessibility of data). Therefore, the CE Program Office conceptualization assesses only one component of data quality, namely measurement error, and we adopt this perspective here. Given the design of the MIS, the true value of the household expenditures is unknown, so we assessed data quality by examining six indirect indicators: number of expenditure reports, average expenditures, record usage, information book usage, combined expense reporting, and the amount of "don't know" and "refusal" responses.

Nonresponse Error – Unit nonresponse is the failure to obtain any measurements from a sampled unit. In longitudinal surveys (such as the CEQ and the MIS), it can arise in the form of panel attrition if sample members respond to the first and/or several consecutive interviews, but fail to respond to the remaining interviews (these are often referred to as *dropouts*). Nonresponse error occurs when the values of statistics computed based only on respondent data differ from those based on the entire sample data (Groves, et al. 2004). To assess the potential for nonresponse error in this study, we examined response rates, panel attrition rates, and changes in respondent sample composition across the waves of the survey.

3.3 Analysis Overview

Given the substantial design differences between the CEQ and the MIS and the relatively small sample size of this study, we do not address comparisons between the study data and CE production data in this report. Instead, our primary focus is on statistical comparisons between the study control group and each of the study treatment groups on the dimensions of respondent burden, data quality, and

nonresponse error. Specifically, we investigated the following hypotheses suggested by the literature reviewed in Section 2[4]:

1a. A shorter interview achieved by splitting the questionnaire will reduce respondent burden.

1b. A shorter interview achieved by splitting the questionnaire will increase data quality.

1c. A shorter interview achieved by splitting the questionnaire will reduce nonresponse error.

2a. The 1-month reference period treatment will increase respondent burden.

2b. The 1-month reference period treatment will improve data quality.

2c. The 1-month reference period treatment will increase nonresponse error.

3. Global expenditure questions will increase data quality[5].

In the next section of the report we provide further specification about the study design and method.

4. STUDY METHODS

4.1 General

This MIS investigated the aforementioned issues using a truncated CEQ interview and a restricted panel design. There were three test conditions in this study (see Table 1).

Control Group (C)

In the C condition, sample units completed a bounding interview in wave 1. The bounding interview used a 1-month reference period and consisted of items taken from nine sections of the current CEQ instrument plus a "core" set of questions (e.g., demographic items) that were

[4] For each of the hypotheses the relevant comparison is the study control group.

[5] 'Data quality' here is defined solely as higher reported average expenditure amounts and lower incidence of 'don't know' and 'refusals;' this study cannot address whether such responses are valid. Other data quality metrics (e.g., number of reports, combined reporting) were not available for the global questions, by definition. In addition, our study design made it impossible to examine the unique effects of global questions on respondent burden or nonresponse error since the global items were a part of the overall shortened questionnaire treatment.

administered across all study conditions. These same C group sample units were contacted again three and six months later to complete two additional interviews using the same "core + nine sections" questionnaire with a 3-month reference period. The C condition paralleled the existing CEQ survey procedures and served as the basis of comparison for the other experimental conditions.

Shortened Questionnaire (SQ)

In the SQ condition, sample units completed the same full bounding interview in wave 1 as the C condition cases, and then were randomly assigned to one of two subsamples that were administered subcomponents of the full questionnaire in waves 2 and 3. Subsample A (SQ-A) received sections 6, 14, 16, 18, and 20, the "core" questions, and a small number of global expenditure questions from sections 9, 12, 13, and 17. Subsample B (SQ-B) received sections 9, 12, 13, and 17, the "core" questions, and a small number of global expenditure questions from sections 6, 14, 16, 18, and 20. Within each subsample group, respondents were split into two groups. One group received the global expenditure questions prior to the detailed expenditure questions, and the other group received the global questions after the detailed items. This counterbalancing allowed us to control for and examine potential order effects stemming from the placement of the global questions. Both the SQ subsample assignments (SQ-A or SQ-B) and the presentation order for the global questions were fixed for waves 2 and 3.

The process of determining which sections to allocate to SQ-A and SQ-B was determined by examining intra-sectional correlations, average duration per section, incidence rate, and potential data quality concerns (e.g., PCE-CE comparisons, imputation/allocation rates). In looking at intra-sectional correlations, we picked sections that had the highest number of "significant" correlations (i.e., 0.1 or greater) with other sections. Then, for each of those sections we identified the section with which it was most highly correlated, and allocated those two sections to different subsamples of our SQ condition. We also attempted to keep the total interview duration similar in the two subsamples. We then examined the incidence rates and CVs for summary variables in the selected sections, and checked our split against one used in Ghosh and Vogt (2000).

Table 1. MIS Test Conditions

Condition	Wave 1		Wave 2	Wave 3
Control (C)	Bounding Interview (1-month recall) "FULL" Interview (Core + 9 sections)	→	2nd Interview (3-month recall) "FULL" Interview (Core + 9 sections)	3rd Interview (3-month recall) "FULL" Interview (Core + 9 sections)
Shortened Questionnaire (SQ)	Bounding Interview (1-month recall) "FULL" Interview (Core + 9 sections)	Respondents randomly assigned to one of two sub-samples – (a) or (b) →	2nd Interview (3-month recall)	3rd Interview (3-month recall)
			(a) Core + sections 1 – 4[6]	(a) Core + sections 1 - 4
			(b) Core + sections 5 - 9	(b) Core + sections 5 - 9
Reference period (RP)	4 Consecutive 1-month Interviews			
	Bounding Interview (1-month recall) "FULL" Interview (Core + 9 sections)	2nd Interview (1-month recall) "FULL" Interview (Core + 9 sections)	3rd Interview (1-month recall) "FULL" Interview (Core + 9 sections)	4th Interview (1-month recall) "FULL" Interview (Core + 9 sections)

Reference period (RP)

In the RP condition, sample units received the same "full" bounding interview that was administered to the wave 1 C and SQ respondents. They then received three consecutive monthly interviews using the same "full" questionnaire with a 1-month reference period (rather than the 3-month reference period used in the C and SQ interviews). Table 2 shows the six possible interview types.

[6] The section numbers referenced in the four SQ cells of this table do not correspond to the original section numbers in the CEQ; they are for illustrative purposes only (see Table 2 for corresponding CEQ sections).

Table 2: Interview Content by MI Study Treatment Group for Interviews After the Initial Interview

	Front	Control Card	Housing	Global (9, 12, 13, 17)	Global (6, 14, 16, 18, 20)	Global Before Detailed	6	9	12	13	14	16	17	18	20	Back
Control Group	x	x	x				x	x	x	x	x	x	x	x	x	x
SQ-A, Version 1	x	x	x	x			x				x	x		x	x	x
SQ-A, Version 2	x	x	x	x		x	x				x	x		x	x	x
SQ-B, Version 1	x	x	x		x			x	x	x			x			x
SQ-B, Version 2	x	x	x		x	x		x	x	x			x			x
Reference period	x	x	x				x	x	x	x	x	x	x	x	x	x

4.2 Data Collection

Mode and Fielding Period

All data in this study were collected by computer-assisted telephone interviewing (CATI) conducted by the Census Bureau's Tucson Telephone Center (TTC) staff. The overall fielding period for this study was nine months, beginning June 1, 2010 and ending in February 18, 2011, but varied across the four treatment conditions. The C and SQ groups consisted of three quarterly interviews, each with a 1-calendar-month fielding period. The RP condition consisted of four consecutive monthly interviews with a three-week fielding period each wave (i.e., the 1st through the 21st of each month).

The sample release was staggered for each treatment group such that one-third of the cases assigned to each group were interviewed in month n, one-third in month $n + 1$, and one-third in month $n + 2$. This approach provided a more manageable case workload for TTC and spread out data collection to minimize potential monthly or seasonality effects. This staggered schedule was carried forward throughout all subsequent interviews, based on when the case was originally released and the appropriate reference period for the condition. These sample segments are henceforth referred to as "panels." For all treatment conditions, a sample unit's eligibility for

continued participation in its survey panel was contingent on its completion of the first interview; nonrespondents at wave 1 were dropped from the remainder of the study.

Sampling Frame

Census developed a nationally representative sampling frame for a target of 8,100 completed interviews across all study treatments and interview waves. The Demographic Statistical Methods Division (DSMD) used the CEQ reserve cases from the address-based unit frame and matched them to known telephone numbers using a telematch procedure. The address-to-telephone number match enabled survey advance materials to be sent to sample members prior to CATI contact. DSMD achieved a 31% telematch rate; non-matches were excluded from the study sample. Census provided the sampling frame, conducted the telematch procedure, purged the frame of known nonresidential units and nonworking numbers, and drew the sample.

Advance Materials

Prior to the start of each interview wave, the Census Bureau's National Processing Center (NPC) mailed advance materials to sample members with an address match. (Table 3 provides the list and scheduled mail outs for these materials; the complete documents are available upon request.)

Table 3. MIS Advance Materials Distribution

1st Interview		**2nd - *nth* Interview**
1st Mailing	2nd Mailing	
Advance letter (modified Form CE-303-L1); "Tracking Your Spending Behavior" brochure.	Modified Information Booklet (CE-305(C))	Advance letter (modified Form CE-303-L2); Modified Information Booklet (CE-305(C))

The MIS Team worked with DSMD staff to modify the existing CEQ advance letters (Form CE-202-L1 – L5) and the CEQ Information Booklet (CE-305(C)). The two biggest changes to the advance letter were that respondents were asked to participate in the "Consumer Expenditure Telephone Survey" and told that "the average interview takes about 25 minutes."[7] The revised MIS Information Booklet

[7] The MIS Team estimated that interviews would take 25 minutes to complete *averaging across treatment groups and interview waves*. Pre-tests indicated that the wave 1 "full" interview took an average of 30 minutes to complete in each condition; waves 2 and 3 C interviews took 27 minutes each; waves 2 – 4 RP interviews took 25 minutes each; and wave 2 and 3 SQ interviews took an average of 18 minutes to complete.

eliminated sections that were not administered in the study and added examples for the global expenditure category questions, but otherwise was identical to the production CEQ Information Booklet.

Within-Household Respondent Selection

As in the CEQ, any adult member of the sampled household age 16 or older could serve as a MIS respondent, but an attempt was made to collect household spending information from the most knowledgeable adult household member (e.g., the owner/renter or their spouse). Changes in respondents between survey waves were allowed and tracked by the instrument when they occurred.

Survey Instrument

Census modified the existing CEQ interview Blaise source code to develop and implement the MIS survey instrument. Table 4 outlines the sections that were taken directly from the CEQ instrument, the set of new questions that the MIS Team provided Census for integration into the Blaise instrument, and the distribution of section/item assignments for the subsamples in the SQ condition. In each interview, respondents were asked about their household purchases in each of the expenditure categories over the

Table 4. MIS Questions: Subject, Origin, and SQ Allocations

CEQ Section	Subject	Question Origin	SQ Section Allocation
FRONT	Case Management	Existing CEQ	Core
CONTROL	Demographics/Roster	Existing CEQ	Core
BACK	Contact Information/CHI	Existing CEQ	Core
n/a	Rent/Mortgage	New – MIS-provided	Core
n/a	Income	New – MIS-provided	Core
6	Appliances	Existing CEQ	SQ-B
9	Clothing	Existing CEQ	SQ-A
12	Vehicle Operating Expenses	Existing CEQ	SQ-A
13	Insurance (non-health)	Existing CEQ	SQ-A
14	Hospital/Health Insurance	Existing CEQ	SQ-B
16	Education Expenses	Existing CEQ	SQ-B
17	Subscriptions/Entertainment	Existing CEQ	SQ-A
18	Trips	Existing CEQ	SQ-B
20	Expense Patterns/Food	Existing CEQ	SQ-B
n/a	Global Expenditure Questions	New – MIS-provided	SQ-A & SQ-B
n/a	Post-Survey Assessment Questions	New – MIS-provided	Core

reference period. In addition, in their final interview (wave 3 for the C and SQ groups, wave 4 for the RP group), respondents were asked a set of post-survey assessment questions (PSAQs) that measured how burdensome they found the survey experience to be, their interest in the survey content, the perceived difficulty of responding to the survey questions, perceived appropriateness of interview length and frequency of survey requests, their estimate of the interview length, and their use of MIS recall aids. Formal systems and verification tests of the instrument were carried out by the MIS Team and Census prior to the start of data collection to ensure that instrument navigation, flow, edits, and database capture and output met study specifications.

Interviewer Staffing, Training, and Monitoring

Approximately 30 TTC CATI interviewers and supervisors worked the MIS data collection over the course of the study fielding period. Census was responsible for the staffing assignments and produced monthly reports for the MIS Team on survey operations. The MIS Team and Census jointly developed an extensive set of interviewer training materials for this study based on the existing CEQ training documentation. However, because the existing materials were designed for CEQ field representatives (not centralized-CATI interviewers) and also did not cover topics and procedures unique to the MIS study, considerable revisions were necessary. Training was developed in two formats: MIS self-study and MIS classroom training. In addition, we provided interviewers with a MIS-tailored Interviewer Manual. The 2-day classroom training was conducted at the TTC facilities during the week of May 10, 2010. The sessions were lead by TTC supervisors and attended by members of the MIS Team who answered study-related questions as required. Throughout the data collection period, TTC supervisors randomly monitored interviewers to ensure that they asked questions as worded, probed effectively, and recorded respondents' answers accurately. In addition, MIS Team members routinely monitored interviews from the remote observation facility at Census, and provided corrective feedback to TTC interviewers when appropriate. Finally, the MIS Team conducted an interviewer survey in September, 2010 to identify any potential problems in survey administration or interviewer understanding of the MIS concepts or procedures. Neither the regular monitoring nor the debriefing survey revealed significant issues that would have negatively impacted survey administration or the quality of the MIS study data.

Data Collection Procedures

MIS cases were assigned to individual interviewers using a WebCATI control system that Census has employed on other CATI surveys (e.g., ACS). At initial contact, the MIS interviewer verified that they had contacted the correct address and attempted to complete the interview. If the respondent agreed to participate, the interviewer proceeded to collect household roster and demographic information (in wave 1; this information was simply verified and updated in subsequent waves) and to administer the expenditure questions appropriate to the MIS treatment group. The control system's set of integrated checks helped to minimize errors (e.g., out-of-range responses, inappropriate skips). In the event of a refusal, the case was reassigned to a supervisory interviewer or refusal conversion specialist; after two refusals WebCATI removed the case from the interview queue and coded it as a noninterview. At the end of each fielding period (three weeks or one month, depending on treatment group), all cases were assigned one of three CATI outcome code types: interview, noninterview, or ineligible for CATI (e.g., incorrect address or telephone numbers). If a sample unit was assigned noninterview or ineligible at the end of the wave 1 fielding period, no further interviews were attempted with this unit.

5. FINDINGS

5.1 Overall Response Rates and Sample Sizes

For this study, we calculated the response rate for each treatment group and interview wave using the AAPOR RR#4 formula (with 0.33 as the estimated proportion of "eligibility unknown" cases assumed to be eligible):

$$\text{Response Rate} = \frac{\text{Interviews}}{\text{Interviews} + (\text{Refusals} + \text{Others} + \text{Non-contacts}) + e(\text{Unknown Eligibles})}$$

Table 5 shows the final response rates for the treatment conditions, averaging across wave and panel.

Table 5. Final Response Rates by Treatment, Averaging Across Wave and Panel

Treatment Group	Response Rate	Sample Size (n)
Control Group (C)	51.7%	3,951
Shortened Questionnaire (SQ)	52.9%	8,092
SQ-A	51.2%	3,906
SQ-B	54.3%	4,186
Reference period (RP)	49.1%	6,525

Table 6 outlines the number of completed interviews by MIS wave.

Table 6. Number of Completed Interviews by MIS Condition and Wave

Condition	Wave 1			Wave 2	Wave 3
Control (C)	805			533	477
Shortened Questionnaire (SQ)	1,686		Total	(n = 1,067)	(n = 1,036)
			SQ-A	487	474
			SQ-B	580	562
	Wave 1	**Wave 2**	**Wave 3**	**Wave 4**	
Reference period (RP)	1,102	607	606	598	

5.2 Verifying Random Assignment to Treatment Groups

Wave 1 Sample Composition

We next examined data from wave 1 completed interviews to compare the frequency distributions of CU size, respondent's age, gender, race, education attainment, and housing tenure across the four treatment groups (see Table 7). This served as a manipulation check of our random assignment to the study treatments (i.e., under random assignment of sample units, the group attributes of the different treatment groups should be roughly equivalent). There were no statistical differences between treatment groups in CU size, respondent's age, gender, race, or

educational attainment. There was evidence of association between treatment group and housing tenure, $\chi^2(3, n = 3{,}580) = 15.5$, $p < .01$. The RP group contained more owners in wave 1 than the C and SQ groups.

Table 7. Demographic Comparisons by Treatment Condition (1st Interview)				
	TREATMENT GROUP			
	CONTROL	RECALL	SQA	SQB
N	805	1,102	774	912
	Percent Distribution			
Number of CU Members				
1	25.3%	25.2%	26.4%	21.3%
2	38.8	37.4	39.0	41.8
3+	35.9	37.4	34.6	37.0
Age Group				
<25	1.2	1.5	1.6	1.8
25-34	4.1	5.3	5.4	4.6
35-64	55.8	54.3	55.9	56.4
65+	37.4	37.7	36.3	35.9
Female	61.1	58.6	59.7	58.1
Race				
White	83.6	86.0	85.1	84.4
Black	7.7	6.4	6.3	6.8
Asian	4.2	4.1	4.1	4.1
Marital Status				
Married	60.5	61.7	61.6	64.7
Widowed	14.9	15.2	12.0	11.6
Divorced	11.8	12.3	11.8	10.5
Separated	1.4	1.6	1.8	1.0
Never married	9.7	7.6	11.8	10.4
Education				
< HS	8.7	7.6	8.9	6.6
HSgrad	22.1	24.6	24.7	23.7
Some college	27.7	26.2	27.1	26.0
Undergrad	22.2	21.6	20.8	24.1
Postgrad	17.5	18.8	17.4	18.1
Own/Rent*				
Own	87.6	89.7	86.2	91.8
Rent	10.3	8.8	12.0	6.9
No rent or mort	2.1	1.5	1.8	1.3
Received Info Book?				
No	33.0	29.6	32.3	32.9

* $p < .01$

Wave 1 Outcome Measures

Since MIS wave 1 interviews were identical for all treatment groups, we can also compare key data quality outcome measures across the groups. Again, given random assignment and the lack of compositional differences between groups observed in Table 7, we would also expect there to be no differences in these measures, and that is what we found (see Table 8). The four treatment groups obtained very similar average expenditures, number of expenditure reports, and incidence of combined expense and 'don't know' reporting.

Table 8. Wave 1 Outcome Measures Verifying Random Group Assignment

Variable	SQ-A	SQ-B	RP	C
$\bar{X}$ **total expenditure ($)**	1,154.50	1,299.60	1,233.30	1,351.00
$\bar{X}$ **# of reports**	7.22	7.49	7.43	7.14
$\bar{X}$ **# of combined reports**	0.18	0.18	0.21	0.18
$\bar{X}$ **# of DK reports**	0.54	0.71	0.60	0.64

The availability of MIS frame data additionally allowed us to compare wave 1 respondents and nonrespondents in each of the four treatment groups on characteristics of area poverty, urbanicity, and Census Region. To the extent that these frame variables are correlated with one or more key data items collected in the MIS, differences between nonrespondents and respondents may indicate the potential for nonresponse bias. The bolded cells in Table 9 show the values that reached statistical difference between respondent and nonrespondent on these variables within each condition. The most consistent finding is that households living in high poverty areas appear to be underrepresented in the MIS respondent pool (by 2.3% to 5.2%, depending on the treatment group). In addition, an examination of the relative magnitudes of the difference estimates across the treatments suggests that the C group is at greatest risk for nonresponse bias (i.e., its difference scores are generally larger than those in the other groups). These results, however, do not address the issue of treatment effects on nonresponse bias since the different treatment manipulations were not implemented until wave 2. We explore these analyses in subsequent sections of the report.

Table 9. Nonrespondent – Respondent Differences on Frame Variable in Wave 1

	SQ-A		SQ-A DIFF	SQ-B		SQ-B DIFF	RP		RP DIFF	C		C DIFF
	Type A	Resp		Type A	Resp		Type A	Resp		Type A	Resp	
20% + in poverty	13.1%	9.7%	**3.4%**	12%	8.8%	**3.2%**	11.5%	9.2%	**2.3%**	14.5%	9.3%	**5.2%**
Urban area	87.9	85.7	**2.2**	83.9	84.2	**-0.3**	86.4	84.6	**1.8**	86.8	83.4	**3.4**
Census Region												
NE	23.2	26.2	**-3.0**	24.6	23.8	**0.8**	24.6	25.8	**-1.2**	25.3	23.5	**1.8**
MW	24.0	25.6	**-1.6**	25.6	26.3	**-0.7**	24.8	28.3	**-3.5**	23.9	27.7	**-3.8**
S	31.6	29.3	**2.3**	34.2	32.9	**1.3**	32.5	31.3	**1.2**	35.9	29.9	**6.0**
W	21.2	18.9	**2.3**	15.5	17.0	**-1.5**	18.0	14.6	**3.4**	14.8	18.9	**-4.1**

5.3 Effect of a Shortened Questionnaire

We next examined the effect of our SQ group on data quality, respondent burden, and nonresponse error. The questionnaire for the two SQ treatment groups was a shortened form of the C questionnaire through the use a split questionnaire design with global questions for a subset of expenditure categories. The reader will recall that in SQ-A the global questions were asked in place of detailed questions in sections 9 (clothing), 12 (vehicle operations), 13 (health insurance), and 17 (subscriptions), and in SQ-B global questions were asked in sections 6 (appliances), 14 (non-health insurance), 16 (education), 18 (trips), and 20A (regular weekly expenditures – i.e., grocery shopping). Since the effects of the global questions may differ by expenditure category, and there may be distinct effects resulting from the unique composition of the detail-global item combinations in each SQ subgroup, we conducted and present separate analyses comparing SQ-A to the Control group and SQ-B to the Control group for each of our analytic dimensions.

Data preparation

Since the source variables required for the C and SQ group comparisons differed depending on whether expenditures were collected from detailed or global questions, we created new analysis variables based on the appropriate source in order to analyze group differences. (Documentation on the mapping between source variables and analysis variables for these group comparisons is available upon request). In addition, before creating the analysis variables, we zero-filled the

source variables since the sample mean (i.e., the average expense incurred per category across all sample units) is the statistic of interest.

5.3.1 SQ – Data Quality

Recall that the hypothesis was that the SQ treatment would result in better data quality than the C group (as defined by higher average expenditure amounts overall, a greater number of detailed expenditure reports, and fewer instances of combined reporting and use of DK and REF among the detailed items). Table 10 presents the results of our data quality analysis comparing SQ-A and C groups. We found no significant difference between the SQ-A and C groups in total expenditures in either wave 2 or 3, although in both interviews the SQ-A group produced estimates that were approximately 20% higher than those obtained under in the C (i.e., in the hypothesized direction). Restricting our analysis to estimates derived only from the detail questions, we found that SQ-A and C performed essentially the same in terms of the number of valid reports, combined (or aggregated) reports, and "don't know or refused" responses (see Table 10). In addition, when we drilled down further to examine section-level comparisons between these two groups for their common detailed sections (6, 14, 18, and 20), we found that they were similar in dollar expenditures, number of valid reports, number of combined (or aggregated) reports, and number of "don't know/refuse" in both waves 2 and 3 (data not shown). Stated differently, the total average expenditure amounts were higher in the SQ-A group (though not significantly so) in both interviews solely because respondents reported higher expenditures in response to the global expenditure questions than they did to the detailed questions. In one sense, the lack of effect of the SQ-A treatment on respondents' detailed reports – no reduction in the number or amount of reporting, and no increase in combined reporting or 'don't know/refusals' – coupled with the higher overall dollar spending estimates, offers some support to the hypothesis that the SQ treatment should produce better data quality.

Table 10. Comparison of Aggregate Data Quality Measures for SQ-A and CONTROL

Variable	SQ-A (A)	Control (C)	Difference (A-C)				
			Mean	95% LCL	95% UCL	SE	p-value for t-test
Wave 2 (quarterly recall)							
$\bar{X}$ **Total expenditures ($)**	3318.00	2752.20	565.80	-58.24	1189.80	318.02	0.0769
$\bar{X}$ **# of valid reports (detailed)**	3.26	3.30	-0.04	-0.43	0.36	0.20	0.8527
$\bar{X}$ **# of combined reports (detailed)**	0.03	0.02	0.01	-0.01	0.04	0.01	0.2215
$\bar{X}$ **# of DK/REF responses (detailed)**	0.18	0.23	-0.05	-0.12	0.02	0.04	0.1771
Wave 3 (quarterly recall)							
$\bar{X}$ **Total expenditures ($)**	2,968.20	2,516.70	451.49	-107.10	1,010.00	284.61	0.1131
$\bar{X}$ **# of valid reports (detailed)**	3.41	3.57	-0.16	-0.55	0.23	0.20	0.4100
$\bar{X}$ **# of combined reports (detailed)**	0.03	0.01	0.01	-0.01	0.03	0.01	0.2418
$\bar{X}$ **# of DK/REF responses (detailed)**	0.18	0.24	-0.06	-0.14	0.01	0.04	0.0903

The trends that existed in the SQ-A to C comparisons were even stronger in the SQ-B to C comparisons (See Table 11). The average overall total expenditure amount was significantly higher in the SQ-B group than the C group in both waves 2 and 3, and SQ-B had more expenditure reports overall than the C group in the detailed question sections for both waves, as well. Moreover, when we examined the indicators of poor data quality (i.e., use of combined reports and "don't know/refused" reports) – we found no difference between the SQ-B and C groups, and note that in both groups incidence of these behaviors is exceedingly low.

As before, we also examined our data quality metrics at the section level for the detailed sections common to both SQ-B and C (9, 12, 13, and 17). There were no section-level differences in dollar expenditure amounts between SQ-B and C in wave 2, but the SQ-B group did have significantly more expenditure reports for vehicle operating expenses (section 12) (1.52 vs. 1.31, $p < .05$) and entertainment (section 17B) (2.12 vs. 1.83, $p < .01$), as well as significantly more

"don't know/refused" reports to questions about non-health insurance policies (0.14 vs. 0.08, $p < .05$). Similarly, in wave 3, SQ-B respondents reported significantly higher dollar expenditures for vehicle operations than C respondents ($263.2 vs. $177.3, $p < .01$), as well as significantly more reports in this category (1.39 vs. 1.13, $p < .01$). There were no wave 3 differences between SQ-B and C groups in combined reports or "don't know/refused" reporting. So, here again we see the impact of global items inflating the total expenditure amounts, but there is also evidence that the shortened interview in the SQ-B group had some independent, additive effect.

Table 11. Comparison of Aggregate Data Quality Measures for SQ-B and CONTROL

Variable	**SQ-B (B)**	**Control (C)**	**Difference (B-C)**				
			Mean	95% LCL	95% UCL	SE	p-value for t-test
Wave 2 (quarterly recall)							
$\bar{X}$ **Total expenditures ($)**	3,955.60	2,674.20	1,281.40	645.17	1,917.60	324.25	**<.0001**
$\bar{X}$ **# of valid reports (detailed)**	9.37	8.41	0.96	0.22	1.71	0.38	**0.0107**
$\bar{X}$ **# of combined reports (detailed)**	0.35	0.33	0.02	-0.06	0.10	0.04	0.6367
$\bar{X}$ **# of DK/REF responses (detailed)**	0.28	0.28	0.00	-0.10	0.10	0.05	0.9717
Wave 3 (quarterly recall)							
$\bar{X}$ **Total expenditures ($)**	4045.40	2,442.90	1,602.60	909.03	2,296.10	353.43	**<.0001**
$\bar{X}$ **# of valid reports (detailed)**	9.99	8.98	1.01	0.16	1.86	0.43	**0.0197**
$\bar{X}$ **# of combined reports (detailed)**	0.27	0.31	-0.04	-0.12	0.03	0.04	0.2667
$\bar{X}$ **# of DK/REF responses (detailed)**	0.18	0.20	-0.02	-0.10	0.05	0.04	0.5921

In addition to comparing average expenditures between the SQ and C treatment groups, we also examined the distributions of expenditure shares between the treatment groups. Expenditure shares are a common way of representing how total expenditures are allocated to the different components of spending. Changes in relative shares can impact the CPI cost weights, so we

wanted to explore potential treatment effects on this measure[8]. To test for differences in shares between the SQ and C groups, we used the Chi-square test of homogeneity in proportions (where the null hypothesis is that different treatment groups have the same proportion of consumer units (CUs) in the expenditure categories) and the adjusted Rao-Scott chi-square test statistic which accounts for the complex sample design. The analysis was implemented with Proc SurveyFreq in SAS v 9.1.3, with the CU as the unit of observation (cluster), and the CU's expenditures in each category as the weights. Non-positive expenditures such as those for reimbursements were dropped from the analyses, and for the SQ groups both the detailed and global items served as source variables.

Expenditure shares were calculated as follows:

Aggregate expenditure on category *j* for group *g*, , is the sum of expenditures on category *j* by households *i* in group *g*:

Total expenditures for group g:

The relative share of category *j* for group *g*: —

Table 12 shows the expenditures shares for the SQ and C groups for waves 2 and 3. As the distributional differences in Table 11 and the associated chi-square results indicate, there was a large and significant treatment effect for both the SQ-A and SQ-B groups relative to the control group. Although the differences between the SQ and C groups' expenditure shares were relatively small for some expenditure categories (e.g., health insurance and trips in SQA – C waves 2 and 3), they were quite large for others (e.g., health insurance and trips for SQ-B – C in waves 2 and 3). This variability may be due in part to the influence of the global questions (e.g., heath insurance and trips were measured by detailed questions in SQ-A, but by global questions in SQ-B).

[8] Our method of calculating expenditure shares differs from the current BLS methods computing expenditure shares. The MIS did not account for various weighting steps used by CPI, and our expenditures base is different because we excluded a number of CEQ sections.

Table 12. Relative Expenditure Shares for SQ and Control Groups for Waves 2 and 3

	Wave 2 (column %)			Wave 3 (column %)		
	SQ-A	**SQ-B**	**CONTROL**	**SQ-A**	**SQ-B**	**CONTROL**
Appliances	9.8	6.9	13.1	11.6	9.1	14.8
Clothing	8.3	6.4	8.5	12.0	7.5	9.8
Vehicle operations	9.2	6.8	8.1	10.7	6.5	7.0
Non-Health insurance	25.2	13.8	17.7	27.9	14.7	20.6
Health insurance	5.0	15.5	4.6	5.9	17.9	6.1
Education	26.8	21.2	24.4	16.5	23.9	22.1
Subscriptions – Entertainment	6.2	6.7	9.0	7.8	6.6	9.5
Trips	3.9	19.7	7.8	1.2	10.6	2.6
Weekly groceries	5.6	3.0	7.0	6.4	3.1	7.4
Total	100.0	100.0	100.0	100.0	100.0	100.0

Test of Homogeneity of SQ expenditure relative shares against CONTROL

- SQA Wave 2: Rao-Scott Chi-Square Test Statistic = 21.23(df=8, p =0.0066)
- SQB Wave 2: Rao-Scott Chi-Square Test Statistic =74.60 (df=8, p<0.0001)
- SQA Wave 3: Rao-Scott Chi-Square Test Statistic =23.44 (df=8, p =0.0028)
- SQB Wave 3: Rao-Scott Chi-Square Test Statistic =70.32(df=8, p<0.0001)

We also examined respondents' use of recall aids (records and the MIS Information Booklet) in their final interview (questions on recall aid usage were only administered in the SQ and C conditions in wave 3). As can be seen in Table 13, the prevalence of information booklet use (before or during the interview) was quite similar across the SQ-A, SQ-B, and C treatments, with slightly more respondents in the two SQ groups than the C group using the Information Booklet to prepare prior to the interview, and slightly fewer in the SQ groups using the Booklet during the interview.[9] Record use trended higher for the C group than either SQ group, but this effect was not significant.

[9] However, the high missing rate for the later variable – over 50% in all three treatment groups – suggests that there may have been administration problems with this question.

Table 13. Use of Recall Aids for SQ and C Respondents, Wave 3 (column percent shown)

	SQ-A N=474	**SQ-B** N=562	**Control** N=477	**Chi-sq p-value**
Information book use to prepare before interview				0.0774
missing	8.9	8.0	13.0	
Yes	44.7	44.0	43.4	
No	46.4	48.0	43.6	
Information book use during interview				0.0437
missing	55.3	56.1	56.4	
Yes	35.0	33.3	36.3	
No	9.7	10.7	7.3	
Record use				0.4124
missing	0.21	0.36	0.21	
Yes	31.22	33.99	37.11	
No	68.57	65.66	62.68	

5.3.2 SQ – Nonresponse Properties

To assess the potential for nonresponse error, we began by comparing SQ and C response rates by interview wave and selected characteristics (see Table 14). When each MIS wave is treated as independent, both SQ groups achieved higher response rates in the final wave than the C, though this result failed to reach statistical significance (p=0.8843). There also was no indication of a treatment effect in the distribution of response rates by geographic characteristics (Census region, percent of poverty in the area, and urban area).

Table 14. Response Rates for SQ and C by Selected Characteristics

Characteristics	C		SQA		SQB		Chi-sq p-value
	No. Eligible[1]	Response rate %	No. Eligible[1]	Response rate %	No. Eligible[1]	Response rate %	
Interview wave							0.8843
1	2,019.75	39.9	1,973.43	39.2	2,087.33	43.7	
2	756.11	70.5	714.75	68.1	856.08	67.8	
3	735.05	64.9	699.37	67.8	839.03	67.0	
Percent of population in poverty in the area							0.5235
20% or more	396.39	42.4	372.64	41.1	374.03	45.2	
Less than 20%	3,114.52	52.9	3,014.91	52.5	3,408.41	55.3	
Census region							0.0505
North-East	833.90	51.9	850.18	52.1	915.66	53.1	
Mid-West	926.88	55.5	842.74	54.0	980.53	56.7	
South	1,127.87	47.2	1,028.77	48.3	1,263.60	52.8	
West	622.26	54.0	665.86	51.1	622.65	55.4	
Urban area							0.4258
Rural	546.03	54.8	453.35	57.1	599.73	55.0	
Urban	2,964.88	51.1	2,934.20	50.3	3,182.71	54.2	

[1] The proportion of eligibility among cases with "unknown" final disposition was assumed to be e=0.33

We then calculated the cumulative response rates for the C and SQ groups, where the response rate at each wave is conditional on eligibility in wave 1 (see Table 15). This provides a cleaner picture of the potential impact of longitudinal burden on response rates and controls for the initial take rate in each treatment group. The cumulative response rate at wave *t* was computed as:

$$\frac{\text{Interviews at wave t}}{\text{[Interviews + (Refusals + Others + Non-contacts) + e(Unknown Eligibles)] at wave 1}}$$

Table 15. SQ and C Group Cumulative Response Rates by Wave

Wave	Response Rate Conditional on Eligibility at Wave 1 (%)		
	C	**SQ-A**	**SQ-B**
1	39.9	39.2	43.7
2	26.4	24.7	27.8
3	23.6	24.0	26.9

Recall that all wave 1 interviews were identical and respondents were administered the SQ treatment for the first time in wave 2. Thus, changes in the SQ cumulative response rates

between wave 1 and 2 are unlikely to be the result of a treatment effect. Treatment effects are more likely to occur in wave 3 given respondents' experience with the full wave 2 interview. If hypothesis 1c is correct, we would expect the SQ groups to have lower attrition rates than the C group. That is what we found: the attrition rates between wave 2 and wave 3 for SQ-A (-0.7%) and SQ-B (-0.9%) were substantially lower than the one observed for the C group (-2.8%).

Another way of assessing potential effects of the SQ treatment on nonresponse error is to compare this group's sample composition in the final interview to that of the C group. As discussed in section 5.2, there were essentially no differences between these groups in wave 1 (the only significant difference was a higher proportion of homeowners in SQ-B than in the C and SQ-A groups). Here we found no evidence of differential changes in sample composition between the SQ and C groups over the life of the panel, suggesting that the magnitude of potential nonresponse bias was at least no greater in the SQ conditions than in the C group. Although the SQ-B group continued to have more homeowners than the other two groups in wave 3, the association was not significant at the final wave (p=0.1142); the distribution of other characteristics also were similar between the groups.

Table 16 presents the estimated relative nonresponse bias for each expenditure category (and associated 95 percent confidence interval) for nonrespondents at the final interview wave. The following formula was used to compute this estimate for wave 3 nonrespondents:

$$\text{Relative Bias} = \frac{\hat{\bar{Z}}_{R,j} - \hat{\bar{Z}}_{T,j}}{\hat{\bar{Z}}_{T,j}}$$

where:

- $\hat{\bar{Z}}_{R,j}$ = mean expenditure estimate for expenditure category j in wave 3 from the total sample. Where there was a nonresponse in wave 3 on category j, the expenditure value was substituted from wave 2 (if reported); if it was not reported in wave 2, then the wave 1 value used for wave 3.
- $\hat{\bar{Z}}_{T,j}$ = mean expenditure estimate for expenditure category j from respondents in wave 3.

Variance estimates of the relative nonresponse bias for each expenditure category were computed using the random groups method (Wolter, 1985), and the data were weighted using the base weights provided by DSMD.

Table 16. Estimated Relative Nonresponse Bias for SQ and C in the Final Wave (Using Base Weights)

	SQ-A			SQ-B			CONTROL		
	Relative bias (%)	95%CI		Relative bias (%)	95%CI		Relative bias (%)	95%CI	
Appliances	**24.2**	5.4	43.0	**26.0**	3.3	48.7	**18.8**	2.4	35.2
Clothing	6.5	-8.5	21.5	-6.4	-18.6	5.8	6.4	-9.5	22.3
Education	14.0	-23.7	51.6	0.6	-19.8	21.1	1.4	-32.4	35.2
Health insurance	-3.3	-33.6	27.1	**-52.4**	-58.8	-46.0	**-22.7**	-44.4	-1.0
Non-health insurance	**-24.6**	-31.0	-18.3	7.3	-1.8	16.4	-5.3	-11.7	1.0
Weekly groceries	**65.5**	61.1	69.9	**55.4**	50.3	60.6	**53.9**	50.3	57.5
Subscriptions & entertainment	**-42.8**	-54.6	-31.0	**-36.1**	-48.3	-23.8	-12.7	-25.7	0.3
Trips	24.8	-17.2	66.8	-20.0	-40.2	0.1	34.4	-15.3	84.1
Vehicle operations	10.3	-6.2	26.7	**18.1**	5.9	30.4	**21.0**	8.6	33.3
Total expenditures	6.1	-4.8	17.0	-3.6	-9.9	2.6	**9.7**	1.7	17.7

A negative value for the relative nonresponse bias indicates that by using only data collected from final wave respondents we would underestimate the expenditure (assuming no other sources of error); conversely, a positive value for the relative nonresponse bias suggests that on average, final wave respondents report higher expenditures than nonrespondents. If zero is included in the 95 percent confidence interval of the estimated relative nonresponse bias, it indicates that nonresponse bias is not affecting the estimated expenditure for that item.

The evidence presented in Table 16 suggests that, due to nonresponse, we may be over-estimating final wave SQ-A expenditures for weekly grocery shopping by 66 percent (95CI: 61.1% to 69.9%) and appliances by 24 percent (95CI: 5.4% to 43.0%), but under-estimating non-health insurance by 25 percent (95CI: -31.0% to -18.3%) and subscriptions and entertainment by 43 percent (95CI: -54.6% to -31.0%). However, nonresponse bias does not appear to affect the SQ-A total quarterly expenditures estimate significantly (estimated relative bias of 6.1%, 95CI: -4.8% to 17.0%).

There is some indication of nonresponse bias at the expenditure section level for the SQ-B group, as well. We appear to be over-estimating final wave expenditures on vehicle operations by 18 percent (95CI: 5.9 %to 30.4%), groceries by 55 percent (95CI: 50.3% to 60.6%), and appliances by 26 percent (95CI: 3.3% to 48.7%), but under-estimating health insurance by 52 percent (95CI: -58.8% to -46.0%), and subscriptions and entertainment by 36 percent (95CI: -48.3% to -23.8%). However, again, nonresponse bias does not appear to affect the total expenditure estimate (estimated relative bias of -3.6%, 95CI: -9.9% to 2.6%).

We see a similar pattern of section-level nonresponse bias in the C group, as well. Final wave expenditure estimates for vehicle operating costs, appliances, and groceries appear to be significantly over-estimated in our respondent pool, whereas health insurance estimates appear to be significantly under-estimated More troubling, the total quarterly expenditure estimate for the C group appears to be positively biased by 9.7 percent (95CI: 1.7% to 17.7%).

We can also examine Table 16 to compare the relative bias measures between the C and SQ groups to assess the effects of our treatment. That is, where there is evidence of bias in the C group, we can look to see if the SQ treatment alleviated, eliminated, or added to the bias. Conversely, we can identify instances where the SQ group may introduce nonresponse bias not present in the C group. For example, nonresponse in both SQ groups appears to exacerbate the bias existing in the C group in the expenditure estimates for appliances, subscriptions and entertainment, and weekly groceries, and to reduce the nonresponse bias in estimates of vehicle operations expenditures and total expenditures.

5.3.3 SQ – Respondent Perceptions of Survey Burden

Table 17 displays the distribution of SQ-A, SQ-B, and C respondent answers to the Post-Survey Assessment Questions (PSAQs) which were designed to capture different dimensions of survey burden. We found a strong association between treatment group and perceived burden, with significantly fewer SQ respondents (27.4% SQ-A and 30.6% SQ-B) saying that they found the survey to be “very burdensome” or “somewhat burdensome” than the C respondents (36.5%). Similarly, SQ respondents were more likely to say that the number of pre-interview calls/contact attempts was ‘reasonable’ (76.0% SQ-A, 73.4% SQ-B) compared to C respondents (68.6%).

And, SQ respondents also were less likely to perceive the final interview to be "too long" (10.1% SQ-A, 8.2% SQ-B) compared to C respondents (17.8%). Finally, we examined the actual length of interview in waves 2 and 3 as another proxy measure of perceived burden. As can be seen in Table 18, there were no differences between treatment groups in wave 1, but the SQA and SQB interviews were significantly shorter than the C interviews in waves 2 and 3 (by more than 6 minutes). Together these results lend strong support to hypothesis 1a that the SQ treatment would reduce respondent burden.

Table 17. Distribution of PSAQ Responses for SQ-A, SQ-B, and C

	SQA	SQB	Control	Chi-sq p-value
Sample size	474	562	477	
Interest in survey				0.4556
missing	0.8	0.9	1.1	
Very	21.7	19.2	21.0	
Somewhat	53.8	49.3	51.4	
Not very	13.5	16.7	16.1	
Not at all	10.1	13.9	10.5	
Ease in answering survey questions				0.6503
missing	0.4	0.4	0.8	
Easy	47.9	47.2	44.2	
Some easy	33.5	37.2	38.0	
Some difficult	16.0	14.2	15.5	
Very difficult	2.1	1.1	1.5	
Survey was burdensome				0.0083
missing	3.6	2.7	0.8	
Very	1.9	3.2	4.8	
Somewhat	25.5	27.4	31.7	
Not very	30.6	32.4	27.7	
Not at all	38.4	34.3	35.0	
Number of survey requests for survey panel				0.9103
missing	1.5	1.4	1.5	
Too many	28.9	30.4	31.9	
Reasonable	69.6	68.2	66.7	
Number of pre-interview calls (contact attempts)				0.0482
missing	1.5	2.0	3.6	
Too many	22.6	24.6	27.9	
Reasonable	76.0	73.5	68.6	
Perceived length of final survey				<0.0001
missing	1.1	1.1	0.2	
Too long	10.1	8.2	17.8	
Too short	0.4	0.2	0.2	
About right	88.4	90.6	81.8	

Table 18. Actual Survey Length for the C and SQ Groups by Wave (Minutes)

	Wave 1			Wave 2			Wave 3		
	Mean	**SE**	**p-value for diff**	**Mean**	**SE**	**p-value for diff**	**Mean**	**SE**	**p-value for diff**
SQ-A	28.87	0.42		21.77	0.48		20.58	0.4	
SQ-B	29.51	0.37		20.53	0.38		19.59	0.35	
C	28.57	0.42		28.85	0.56		26.69	0.6	
Estimated difference* SQ-A – CON	0.30	0.58	0.5994	-7.09	0.69	<0.0001	-6.11	0.68	<0.0001
Estimated difference* SQ-B – CON	0.94	0.55	0.0896	-8.32	0.66	<0.0001	-7.10	0.65	<0.0001

* from ANOVA

5.4 Effect of a Shortened Reference Period

In this section we report the results of our examination of the effects of a shortened reference period on data quality, nonresponse error, and respondent burden. Specifically, we compared the findings from the C group to those from the RP group, which employed a 1-month reference period and four consecutive monthly interviews. To do so, we first constructed quarterly estimates from the RP group by aggregating across completed interviews in waves 2 through 4 to compare to the quarterly estimates directly obtained from the C group's wave 2 interview[10]; the reference period months common to both conditions were June 2010 through October 2010. In addition, for the RP group, the "usual weekly expense" variable in section 20 (which includes groceries, alcoholic beverages, and meals away from home) was divided by three after aggregating across the three interviews to account for the "usual weekly" reference period. As before, all analysis variables were zero-filled to compute the sample mean.

5.4.1 RP – Data Quality

Table 19 presents the comparisons between the C and RP groups on each of the key data quality metrics at the aggregate survey (not section) level. The derived (aggregated) RP estimate for overall average expenditure amount ($) exceeded that of the C group ($3107.1 vs. $2752.2, respectively), but the difference was not statistically significant (p=0.2225). The RP group did produce significantly more expenditure reports than the C group (difference of 9.9 reports, se

[10] For the RP-C analyses we dropped RP CUs that had not completed all four interview waves.

0.69, p<0.0001), but it also evinced a greater number of combined reports and "don't know/refused" responses.

Table 19. Comparison of RP and C Aggregate Data Quality Measures

Variable	Recall (R)	Control (C)	Difference (R-C)				
			Mean	95% LCL	95% UCL	SE	p-value for t-test
$\bar{X}$ **Expenditures ($)**	3,107.10	2,752.20	354.90	-255.60	935.45	295.81	0.2225
$\bar{X}$ **# of valid reports**	21.59	11.71	9.88	8.53	11.24	0.69	<.0001
$\bar{X}$ **# of combined reports**	0.67	0.35	0.32	0.20	0.45	0.06	<.0001
$\bar{X}$ **# of DK/refused responses**	4.99	1.67	3.32	3.00	3.64	0.16	<.0001

We also examined the same measures at the section level (data not shown) and found that the RP group obtained more valid expenditure reports (p<0.05) than the C group for all sections except 18B (trips). However, the RP results for expenditures amounts were mixed. They were significantly higher than the C group for sections 6A (major appliances; $120.6 vs. $61.6), 12 (vehicle operation; $437.7 vs. $223.7), and 14 (health insurance; $386.2 vs. $117.9), but lower for sections 17 (subscriptions/entertainment; $192.9 vs. $247.4), 18 (trips; $41.6 vs. $215.6), and 20A (regular expenditure patterns; $173.8 vs. $192.4). These differences are also reflected in the expenditure shares changes between the RP and C groups (see Table 20).

Table 20. Relative Expenditure Shares for the RP Group

Expenditure Category	Aggregates ($)		Relative Share (% distribution)	
	Recall	Control	Recall	Control
Appliances	192,291	199,976	15.3	13.1
Clothing	124,457	96,791	7.4	8.5
Vehicle operations	119,203	183,413	14.0	8.1
Non-Health insurance	260,286	203,932	15.6	17.7
Health insurance	67,282	166,461	12.7	4.6
Education	358,506	284,866	21.8	24.4
Subscriptions & entertainment	131,877	80,808	6.2	9.0
Trips	114,927	17,435	1.3	7.8
Regular weekly expenditures	102,536	72,835	5.6	7.0
Total expenditures	1,471,365	1,306,517	100.0	100.0

RC: Test of Homogeneity of expenditure relative shares;
Rao-Scott Chi-Square Test Statistic = 57.42(df=8, $p < 0.0001$)

The RP group also evidence poorer data quality by producing a greater number of combined reports than the C group in sections 9A (clothing) and 13B (non-health insurance), and more "don't know/refused" responses in sections 13B, 14B, 18A, and 20A, though the incidence of both behaviors was very low overall.

Finally, RP respondents were more likely than CG respondents to use the Information Booklet to prepare for the interview (49.7% vs. 43.4 %; $p < 0.05$). They were also slightly more likely use it to follow along during the interview (38.8% vs. 36.3%; $p < 0.05$), but again this variable may be suspect given its high missingness rate (over 50%). There was no difference between the RP and C groups in their prevalence of records use (36.1% vs. 37.1%, respectively; $p = .9321$).

5.4.2 RP – Nonresponse Properties

Table 21 displays RP response rates by selected characteristics. Treating each wave as independent, the response rate was higher for the C group than RP group in waves 1 – 3 ($p=0.0503$). There was a significant effect for Census region, as well – the RP group achieved a

Table 21. Response Rates for RP and C Groups by Selected Characteristics

Characteristic	CONTROL		RECALL		Chi-sq p-value
	No. Eligible[1]	Response rate %	No. Eligible[1]	Response rate %	
Interview wave					0.0503
1	2,019.8	39.9	2,873.9	38.3	
2	756.1	70.5	1,044.4	58.1	
3	735.1	64.9	1,007.6	60.1	
4	n/a	n/a	1,006.0	59.4	
Interview panel					0.1385
1	1,189.0	51.2	2,108.4	49.0	
2	1,084.4	46.5	1,798.2	46.5	
3	1,237.6	56.7	2,025.3	51.5	
Percent of population in poverty in the area					0.4517
20% or more	396.4	42.4	577.7	43.5	
Less than 20%	3,114.5	52.9	5,354.2	49.7	
Census region					0.0067
North-East	833.9	51.9	1,503.8	48.9	
Mid-West	926.9	55.5	1,628.4	50.8	
South	1,127.9	47.2	1,880.9	48.8	
West	622.3	54.0	918.9	46.9	
Urban area					0.5377
Rural	546.0	54.8	892.0	56.1	
Urban	2,964.9	51.1	5039.9	47.9	

[1] The proportion of eligibility among cases with "unknown" final disposition was assumed to be 0.33

lower response rate in the West and mid-West than the C group (p=0.0067) – but the RP – C groups obtained similar response rates within the other geographic groups. The cumulative response rates (conditional on wave 1 participation) presented in Table 22 reveal that the rate of attrition in the RP group was highest between waves 1 and 2 (-17.3% vs. -13.5% for C) but remained essentially unchanged after that, whereas respondents in the C group continued to attrite between waves 2 and 3.

Table 22. Cumulative Response Rates for RP and C Groups by Wave

Wave	Response Rate Conditional on Eligibility at Wave 1 (%)	
	C	**RP**
1	39.9	38.3
2	26.4	21.1
3	23.6	21.1
4	na	20.8

Analysis of the RP and C groups' sample compositions in waves 1 and 4 revealed no significant differences between treatments in either wave, and no significant changes over the life of the panel, suggesting that the RP group's lower wave response rates and steeper attrition rate at wave 2 may not have increased nonresponse error relative to the C group (assuming that the sample composition variables examined are correlated with expenditure reporting).

Table 23 displays the estimated relative nonresponse bias for each expenditure category in the RP and C final interviews. Thus, we are examining relative nonresponse bias for wave 4 monthly expenditure estimates in the RP group, and nonresponse bias for wave 3 quarterly estimates in the C group. The procedure for computing the total sample expenditure estimate at the final wave was the same as before (i.e., final wave nonrespondent estimates are based on the reported value for that item given in most recent available interview and we used sample base weights in the analysis). Again, negative values suggest that we are underestimating the expenditure due to nonresponse, and positive values suggest we are overestimating it. If zero is included in the 95 percent confidence interval of the estimated relative nonresponse bias, it suggests nonresponse bias is not affecting the estimated expenditure for that item.

Table 23. Estimated Relative Nonresponse Bias in the Final Wave for RP and C (base-weighted)

	RP (monthly expenditure estimates)			**C** (quarterly expenditure estimates)		
Expenditure category	Relative bias (%)	95%CI		Relative bias (%)	95%CI	
Appliances	**17.2**	0.7	33.6	**18.8**	2.4	35.2
Clothing	-15.2	-31.8	1.3	6.4	-9.5	22.3
Education	**46.1**	30.9	61.2	1.4	-32.4	35.2
Health insurance	17.3	-3.5	38.1	**-22.7**	-44.4	-1.0
Non-health insurance	**-25.5**	-37.1	-14.0	-5.3	-11.7	1.0
Regular weekly expenditures	-1.7	-6.0	2.6	**53.9**	50.3	57.5
Subscriptions & entertainment	5.1	-9.4	19.5	-12.7	-25.7	0.3
Trips	-20.6	-59.2	18.0	34.4	-15.3	84.1
Vehicle operations	**15.8**	0.1	31.5	**21.0**	8.6	33.3
Total expenditures	5.6	-3.4	14.6	**9.7**	1.7	17.7

The data indicate that attrition in the RP group may be causing us to overestimate final wave RP expenditures for vehicle operations by 16 percent (95CI: 0.1% to 31.5%), education by 46 percent (95CI: 30.9% to 61.2%), and appliances by 17.2 percent (95CI: 0.7% to 33.6%), but under-

estimate non-health insurance by 26 percent (95CI: -37.1% to -14.0). However, nonresponse bias does not appear to be significantly affecting the RP total monthly expenditure estimate (5.6% estimated relative bias; 95CI: -3.4% to 14.6%). Comparing these results to those obtained for the C group, we find that the RP treatment reduces potential nonresponse bias in the total expenditure estimate as well as for a number of expenditure categories (appliances, health insurance, regular weekly expenses, and vehicle operations), but worsens it for two others (education and non-health insurance).

5.4.3 RP – Respondent Burden

Table 24 displays the distribution of RP and C respondent answers to the Post-Survey Assessment Questions (PSAQs), and there is evidence of significant and strong treatment effects on a number of burden dimensions. For example, 34.4 percent of RP respondents said that the survey was 'not very / not at all interesting' compared to 26.6 percent in the C group. Significantly more RP respondents than C respondents said that the survey questions were "easy" (52.2% vs. 44.2%). RP respondents also were more likely than C respondents to rate the survey as 'very burdensome' or 'somewhat burdensome' (45.3% vs. 36.5%), and to say that there were "too many" MIS survey requests (42.1% vs. 31.9%). Moreover, despite the fact that interviews were significantly shorter in the RP group than the C group (more than 4 minutes shorter in waves 2 and 3; see Table 25), proportionally more respondents in the RP group perceived their final interview to be "too long" though this difference did not reach statistical significance (p= 0.1207).

Table 24. Distribution of PSAQ Responses for RP and C Groups

	RP (wave 4)	C (wave 3)	Chi-sq p-value
Sample size	598	477	
	Column percent %		
Interest in survey			0.0478
missing	0.7	1.1	
Very	16.2	21.0	
Somewhat	48.8	51.4	
Not very	19.4	16.1	
Not at all	14.9	10.5	
Ease in answering survey questions			0.0098
missing	0.0	0.8	
Easy	52.2	44.2	
Some easy	35.5	38.0	
Some difficult	10.9	15.5	
Very difficult	1.5	1.5	
Survey was burdensome			0.0157
missing	0.7	0.8	
Very	7.5	4.8	
Somewhat	37.8	31.7	
Not very	27.4	27.7	
Not at all	26.6	35.0	
Number of survey requests for survey panel			0.0015
missing	2.0	1.5	
Too many	42.1	31.9	
Reasonable	55.9	66.7	
Number of pre-interview calls (contact attempts)			0.0968
missing	1.7	3.6	
Too many	30.9	27.9	
Reasonable	67.4	68.6	
Perceived length of final survey			0.1207
missing	0.5	0.2	
Too long	20.9	17.8	
Too short	1.2	0.2	
About right	77.4	81.8	

Table 25. Actual Survey Length (minutes) for RP and C Groups by Interviewer Wave

	Wave 1			Wave 2			Wave 3			Wave 4		
	Mean	**SE**	**p-value for diff**	**Mean**	**SE**	**p-value for diff**	**Mean**	**SE**	**p-value for diff**	**Mean**	**SE**	**p-value for diff**
RP	28.9	0.3		22.4	0.4		22.3	0.5		21.6	0.4	
C	28.6	0.4		28.9	0.6		26.7	0.6				
Estimated difference* (RP – C)	0.3	0.53	0.6015	-6.5	0.7	<0.0001	-4.4	0.6	<0.0001			

***** from ANOVA

5.4 Effect of a Global Expenditure Questions on Data Quality

The third research objective of this study was to examine whether global questions can solicit data of sufficient quality to replace detailed questions. The use of global questions in the SQ treatment condition provided an opportunity to address this question through an analysis of expenditure amounts and reporting rates of "don't know/refused" responses collected through global questions and detailed questions. In this section, we begin by revisiting in more detail the differences between the detailed-based expenditure estimates obtained in the C group and the global-based estimates observed in the SQ-A and SQ-B groups. We then examine comparisons of estimates from a single SQ group (formed by simply concatenating SQ-A and SQ-B records) and the C group both at the section level and overall. Finally, as an ancillary analysis, we investigate whether the order of the block of global questions and the block of detail questions had an effect on expenditure estimates within each SQ group.

In order to compare global and detail estimates, we created analysis variables that were sourced either from the detailed items in the C and SQ groups, or from the global variables in the two SQ groups. As before, we zero-filled the source variables.

Tables 26 and 27 display expenditure estimates derived from the detailed questions in the C group and those derived from the global questions about the same category in the SQ-A and SQ-B groups, respectively. Global questions elicited significantly higher expenditure reports in five of the ten expenditure categories in wave 2 (six in wave 3), and a significantly lower expenditure

estimate of subscriptions in both waves (see bolded cells). There was no significant effect in either wave of question form (detailed vs. global) on expenditure estimates of appliances, education, and weekly groceries.

Table 26. Detailed-Based (C) and Global-Based (SQ-A) Expenditure Estimates for Waves 2 and 3

	Global (SQ-A)	Detailed (C)	Difference	95LCI diff	95UCI diff	SE diff	p-value for t-test
Wave 2							
Clothing	274.9	233.5	41.4	-12.6	95.4	27.5	0.1378
Vehicle operations	**304.6**	**223.7**	**80.9**	19.4	142.5	31.4	0.0112
Non-health insurance	**839.1**	**488.3**	**350.8**	232.0	469.5	60.5	<.0001
Subscriptions	**37.5**	**121.8**	**-84.3**	-120.0	-48.6	18.2	<.0001
Entertainment	**168.7**	**125.6**	**43.1**	2.6	83.7	20.7	0.0406
Wave 3							
Clothing	**357.8**	**249.2**	**108.6**	8.5	208.7	51.0	0.034
Vehicle operations	**319.1**	**177.3**	**141.8**	59.6	224.1	42.0	0.0008
Non-health insurance	**832.0**	**520.8**	**311.2**	183.6	438.8	65.0	<.0001
Subscriptions	**45.3**	**122.6**	**-77.3**	-107.3	-47.2	15.3	<.0001
Entertainment	**188.1**	**118.4**	**69.7**	13.6	125.6	28.5	0.0152

Table 27. Detail-Based (C) and Global-Based (SQ-B) Expenditure Estimates for Waves 2 and 3

	Global (SQ-B)	Detailed (C)	Difference	95LCI diff	95UCI diff	SE diff	p-value for t-test
Wave 2							
Appliances	271.3	360.8	-89.5	-186.2	7.3	49.3	0.0748
Health insurance	**613.4**	**117.9**	**495.5**	422.0	568.9	37.4	<.0001
Education	839.5	672.6	166.9	-289.7	623.3	232.6	0.4704
Trips	**779.0**	**215.6**	**563.4**	372.0	754.8	97.6	<.0001
Weekly groceries	120.0	114.4	5.6	-4.2	15.3	5.0	0.2649
Wave 3							
Appliances	368.8	375.4	-6.6	-113.7	100.5	54.6	0.9040
Health insurance	**725.7**	**140.4**	**585.3**	447.4	723.2	70.3	<.0001
Education	966.5	559.8	406.7	-139.1	952.6	278.2	0.1289
Trips	**430.2**	**65.0**	**365.2**	256.3	474.1	55.5	<.0001
Weekly groceries	124.7	114.0	10.7	-0.6	22.0	5.8	0.0642

Another way to look at the effects of global questions is to examine their impact on the final survey estimates. In a typical SQ design, data from the various questionnaire subcomponents are combined in some way to produce a single dataset for analysis. Although systematic examination

and evaluation of the methods for combining SQ files was beyond the scope of this project, we were interested to see how estimates from the C group would compare to those from a single, combined SQ dataset. To create this data files, we simply concatenated the SQ-A and SQ-B files and summed across responses from the detailed and global questions in each expenditure category. We then calculated expenditure estimates and counts of 'don't know/refuse' for this combined SQ file, and compared them to those obtained in the C group (see Tables 28 and 29).

Table 28: Comparison of Combined-SQ and C Group Estimates of Quarterly Expenditures ($)

Expenditure category	Wave 2					Wave 3				
	SQ	C	Diff (SQ-C)	SE diff	p-value T-test	SQ	C	Diff (SQ-C)	SE diff	p-value T-test
Appliances	296.2	360.8	-64.6	43.1	0.1765	358.7	375.4	-16.7	45.1	0.7225
Clothing	2 63.1	233.5	29.6	22.2	0.1601	**328.7**	**249.2**	**79.5**	38.0	0.0138
Vehicle operations	**285.7**	**223.7**	**62.1**	26.3	0.0108	**288.8**	**177.3**	**111.5**	33.4	<.0001
Non-health insurance	**679.0**	**488.3**	**190.7**	49.2	<.0001	**704.2**	**520.8**	**183.4**	53.5	<.0001
Health insurance	**405.7**	**117.9**	**287.8**	32.5	<.0001	**468.0**	**140.4**	**327.6**	55.4	<.0001
Education	862.9	672.6	190.3	201.6	0.3281	749.1	559.8	189.3	218.2	0.3363
Subscriptions & entertainment	238.1	247.4	-9.3	25.2	0.7233	251.0	241.0	10.0	28.3	0.6898
Trips	**483.1**	**215.6**	**267.5**	77.2	0.0008	**249.7**	**65.0**	**184.7**	42.4	<.0001
Groceries	**50.5**	**114.4**	**-63.9**	4.4	<.0001	**51.9**	**114.0**	**-62.1**	4.5	<.0001
Total expenditures	**3564.3**	**2674.2**	**890.1**	284.9	0.0001	**3449.9**	**2442.9**	**1007.1**	297.0	0.0001

Significant differences between the combined-SQ and C group estimates are bolded in both tables. Total reported expenditures were higher for the combined-SQ group than the C group in both waves 2 and 3 (by $890 and by $1007, respectively). At the section level, wave 2 and wave 3 estimates from the single SQ data file were significantly higher than those from the C group in the categories of vehicle operations, health insurance, non-health insurance, and trips. In addition, although there was no difference between the combined-SQ and C groups in wave 2 estimates of clothing expenses, the SQ estimate was significantly higher in wave 3.

Each of these results is consistent with what we found earlier when we looked separately at the SQ-A and SQ-B comparisons to the C group. One departure from those findings here is that estimates of weekly grocery spending were significantly lower in the combined-SQ group than the C group. The significant differences we noted in our earlier SQA-C and SQB-C comparisons

of higher entertainment estimates and lower subscriptions estimates in the SQ group were not evident when using the combined-SQ data.

We next examined the incidence of "don't know/refused" responses as an indicator of data quality. However, since the number of questions asked about each expenditure category varied by question format (i.e., there were more opportunities for DK/REF response with detailed questions than global items), we created a section-level flag to indicate their presence or absence. As shown in Table 29, the proportion of "don't know/refused" responses trended lower in the combined-SQ group than the C group, with significant differences observed for non-health insurance and weekly grocery spending in waves 2 and 3, and subscriptions and memberships in wave 2.

Table 29. Comparison of Combined-SQ and C Group: Proportion of "Don't Know/Refused"*

Expenditure category	Wave 2					Wave 3				
	SQ	C	Diff (SQ-C)	SE diff	p-value T-test	SQ	C	Diff (SQ-C)	SE diff	p-value T-test
Appliances	0.0112	0.0206	-0.0090	0.0063	0.1775	0.0068	0.0168	-0.0100	0.0055	0.1189
Clothing	0.0375	0.0525	-0.0150	0.0226	0.5529	0.0125	0.0273	-0.0150	0.0082	0.1338
Vehicle operations	0.0169	0.0356	-0.0190	0.0101	0.1288	0.0087	0.0231	-0.0140	0.0069	0.0936
Non-health insurance	0.1078	0.0844	0.0234	0.0202	0.2156	0.0840	0.0901	-0.0060	0.0184	0.7372
Health insurance	**0.1097**	**0.1895**	**-0.0800**	0.0227	0.0011	**0.1014**	**0.2055**	**-0.1040**	0.0234	<.0001
Education	0.0056	0.0094	-0.0040	0.0048	0.4477	0.0039	0.0042	-0.0003	0.0040	0.9302
Subscriptions & entertainment	**0.0337**	**0.0675**	**-0.0340**	0.0110	0.0057	0.0222	0.0252	-0.0030	0.0083	0.7225
Trips	0.0028	0.0019	0.0009	0.0027	0.7061	0.0029	0.0042	-0.0010	0.0032	0.7029
Groceries	**0.5511**	**1.1614**	**-0.6100**	0.0516	<.0001	**0.4884**	**1.109**	**-0.6210**	0.0525	<.0001

* A "dk/refused" is flagged for a section if there is one or more "dk/refused" response in any subsection.

Finally, as noted in Section 4.1, the MI study design allowed us to examine potential order effects stemming from the placement of the block of global items within each SQ subsample. Half of the respondents in each SQ group received the global expenditure questions prior to the detailed expenditure questions, and half received the global questions after the detailed items (with the order of presentation fixed across interview waves). We investigated the impact of the global-detail ordering on respondents' reported expenditure amounts (for both global and detailed

items), number of expenditure reports, DK/REF reports, and use of combined reports for detailed items. Tables 30 and 31 present the results of the global-detail order analyses for the SQ-A and SQ-B groups, respectively.

Table 30. Effect of Global-Detail Ordering on Key Outcome Measures – SQ-A

	Wave 2			Wave 3		
	Global Q Order		Diff	Global Q Order		Diff
	1st	2nd		1st	2nd	
$\bar{X}$ **Expenditures on Global Items ($)**	1,906.8	1,382.7	**524.1**	2,210.6	1,333.2	**877.4**
$\bar{X}$ **Expenditures on Detailed Items ($)**	1,985.9	1,147.4	**838.5**	1,359.9	804.4	**555.5**
$\bar{X}$ **# of valid reports**	3.59	2.99	**0.60**	3.58	3.26	0.32
$\bar{X}$ **# of DK/refused responses**	0.20	0.17	0.03	0.19	0.16	0.03
$\bar{X}$ **# of combined reports**	0.05	0.01	0.04	0.03	0.02	0.01

Table 31. Effect of Global-Detail Ordering on Key Outcome Measures – SQ-B

	Wave 2			Wave 3		
	Global Q Order		Diff	Global Q Order		Diff
	1st	2nd		1st	2nd	
$\bar{X}$ **Expenditures on Global Items ($)**	2,908.6	2,341.4	567.2	3,008.1	2,212.3	795.8
$\bar{X}$ **Expenditures on Detailed Items ($)**	1,381.7	1,284.1	97.61	1,542.1	1,313.8	228.3
$\bar{X}$ **# of valid reports**	9.23	9.51	-0.28	9.89	10.10	- 0.21
$\bar{X}$ **# of DK/refused responses**	0.21	0.34	**-0.13**	0.17	0.19	-0.02
$\bar{X}$ **# of combined reports**	0.40	0.31	0.09	0.28	0.26	0.02

The bolded cells in these tables indicate significant differences between the outcome measures based on the order of the global item administration ($p<.05$). As shown in Table 30, when the block of global questions came before the block of detailed questions, SQ-A respondents reported higher expenditure estimates for both the global and detailed items than when the block of detailed questions were administered first, and this result was obtained in both waves 2 and 3. In

addition, in wave 2, SQ-A respondents who were asked the global items first reported significantly more expenditure reports than those who were asked the block of detail items first. Table 31 reveals a similar though non-significant trend for the expenditure amounts reported by SQ-B respondents in waves 2 and 3, although SQ-B respondents asked global questions first provided significantly fewer DK/REF responses than those who were asked the detailed items first.

6. DISCUSSION

6.1 Summary

The objectives of this study were to: (1) assess the effects of administering a shorter CEQ instrument on respondent burden, data quality, and nonresponse error; (2) examine the impact of using a one-month (versus the current three-month) reference period on respondent burden, data quality, and nonresponse error; and (3) evaluate the quality of data collected from global, as opposed to, detailed questions on expenditures. To achieve these objectives, the MIS study implemented an experimental design in which respondents were randomly assigned to a control group (C) which received no treatment, a test group that received a shortened questionnaire design (SQ), or a test group that was administered a shortened reference period (RP).

The following is a summary of the study findings as they pertain to the set of hypotheses laid out in Section 3.3 of this report.

Did a shorter interview achieved by splitting the questionnaire reduce respondent burden?
Yes. Respondent burden was significantly lower in the SQ groups than C group. SQ respondents perceived the survey to be less burdensome and of appropriate duration and frequency, compared to the control group respondents. SQ interviews were 6 minutes shorter than C interviews on average.

Did a shorter interview achieved by splitting the questionnaire increase data quality?
Somewhat. Data quality moderately improved under the SQ treatment relative to the control condition. Both SQ subsamples (SQ-A and SQ-B) produced total expenditure estimates that were higher than the control estimates, although only the SQ-B group reached statistical significance. In addition, the SQ-B group reported significantly more expenditure reports than the C group. The SQ treatment did not substantively impact the incidence of negative respondent behaviors (i.e., combined reports, “don’t know/refusals”) or the use of recall aids or records.

Did a shorter interview achieved by splitting the questionnaire reduce nonresponse error?
The effects of the SQ treatment on indicators of nonresponse error were minor, varied, but generally positive. Response rates examined independently by interview wave revealed no treatment effect (i.e., they were comparable for the SQ and C groups at each wave). However, the SQ groups attained significantly lower attrition rates between wave 2 and wave 3 than the C group (0.7% and 0.9% for the SQ groups vs. 2.8% for C). The final wave cumulative response rate (i.e., conditioned on participation in wave 1) also was higher in the SQ groups than the C group. There were no observed differences in sample composition between the SQ and C groups in the final wave. Finally, compared to the C group, the SQ treatment reduced the relative nonresponse bias in total expenditures estimates as well as vehicle operations expenditures estimates, though there was evidence that it also exacerbated the bias existing in a few of the C group expenditure estimates.

Did the 1-month reference period treatment increase respondent burden?
Yes. There were significant and strong RP treatment effects on a number of respondent burden dimensions. Significantly more RP than C respondents said that the survey was 'not very / not at all interesting' and 'very / somewhat burdensome,' and that that there were 'too many' survey requests. In contrast, more RP respondents than C respondents said that the survey questions were 'easy.' Moreover, despite the fact that actual interview durations were significantly shorter in the RP group than the C group (by more than 4 minutes in waves 2 and 3), proportionally more respondents in the RP group perceived their final interview to be "too long."[11]

Did the 1-month reference period treatment improve data quality?
Evidence on the effect of RP treatment on data quality was mixed. There were some indications that RP improved data quality. For example, respondents in the RP group did report significantly more valid expenditure reports, and the total expenditures estimate in this group was higher than the C estimate (but not significantly so). In addition, RP respondents were more likely than C respondents to use the Information Booklet to prepare for the survey in advance. On the other hand, RP respondents were significantly more likely than the C respondents to engage in

[11] The burden in the RP condition stems from both longitudinal burden (the burden associated with being interviewed multiple times in relative quick succession) and the cognitive burden associated with the reference period. The MI study design did not permit an examination of the effects of these two sources separately.

undesirable reporting behaviors (e.g., use of combined item reporting and "don't know" and/or "refused" responses). In particular, the RP group was higher in both of these undesired reporting behaviors for section 9 (clothing), a section that is already problematic in the current instrument using a 3-month recall. The RP group had nearly three times as many "don't know/refusals" as the C group; represented as a percent of the average total number of reports, the RP group's rate of DK/REF was 23% compared to 13% for the C group. There was no difference in use of records between the RP and C groups.

Did the 1-month reference period treatment increase nonresponse error?

The RP treatment had a negative impact on survey participation. Response rates examined independently by wave and conditional on wave 1 participation were lower for the RP group than the C group in waves 2 and 3. The attrition rate between wave 1 and 2 also was substantially higher for the RP group (17.2% vs. 13.5%), possibly due to the RP group's tighter fielding period and/or the saliency of respondents' prior wave (negative) experience.

Overall, it does not appear that RP treatment worsened any potential nonresponse bias that may have existed in the C group. The sample of respondents in RP and C were generally similar in distribution on the selected demographic characteristics. In addition, the RP data showed less relative nonresponse bias in total expenditure estimates and estimates of health insurance spending and regular weekly expenditures compared to the C group. However, the RP group showed worse nonresponse bias for estimates of education and appliances expenditure (which were over-estimated) and non-health insurance expenditures (which was underestimated).

Did global expenditure questions increase data quality?

Global-based spending estimates were significantly higher than detailed-based estimates in six of the ten expenditure categories examined in this study (clothing, vehicle operations, non-health insurance, health insurance, entertainment, and trips), and significantly lower in only one (books/subscriptions). We present evidence that the use of global questions reduced levels of "don't know / refused" responses, as well.

6.2 Limitations

As noted earlier in this report, the prohibitive cost of conducting in-person data collection impelled us to rely on centralized computer-assisted telephone interviews (CATI). As a consequence of this mode change, we also eliminated sections of the survey to shorten its overall length. Changes to mode, length, and question context impact the response process and associated errors, so it is likely that some of our results would have been different under a design closer to that of the CEQ. We also were restricted by the project budget to a relatively small sample size. This reduced our power to detect some treatment effects and prevent us from examining effects at lower levels of analysis (below the section-level).

In addition, as noted elsewhere in this report, there were potential limitations with some of our analytic techniques and outcome measures. For example, we had no direct way to assess the 'more is better' hypothesis of data quality because we did not have true values on expenditures. This limitation is not unique to the MI study, but it deserves underscoring as CE embarks on redesign efforts that will look to measures of data quality improvements. In addition, the nonresponse bias analyses we conducted should only be viewed as suggestive. The method involved carrying forward the last available observation, and this may be tenuous for expenditures that are unlikely to be recurring in two subsequent interview periods. Moreover, our nonresponse bias estimates may provide a worse-case scenario since the data in this study were not subject to the same rigorous nonresponse adjustment procedures as utilized in the CE production environment.

6.3 Recommendations

The results of this study suggest that a SQ design may hold promise in a redesigned CEQ. Additional research is needed to determine the optimal length of a shortened survey, composition of questionnaire splits (in terms of their statistical properties and impact on respondent processes/errors), and dataset construction and analysis methods. We are less sanguine about the adoption of a 1-month reference period, given the concomitant need for conducting monthly interviews, and our findings on the negative effects of this design on response rates and respondent burden. That said, the optimal reference period likely will vary across expenditures, and additional laboratory research is needed in this area. Similarly, we recommend additional research (e.g., cognitive studies, controlled experiments, validation studies) on respondents' use of global questions.

References

Apodaca, R. , Lea, S. and Edwards, B. (1998). “The Effect of Longitudinal Burden on Survey Participation.” Proceedings of the 1998 American Statistical Association, Survey Research Methods Section.

Backor, K., Golde, S. and Norman, N. (2007). “Estimating Survey Fatigue in Time Use Study.” Paper presented at the 2007 International Association for Time Use Research Conference, Washington, DC.

Battistin, E. (2003) Errors in Survey Reports of Consumption Expenditures. Working Paper W03/07, Institute for Fiscal Studies, London.

Belli, R., Schwarz, N., Singer, E. and Talarico, J. (2000). "Decomposition Can Harm the Accuracy of Behavioural Frequency Reports." *Applied Cognitive* Psychology, 14, 295-308.

Bradburn, N. M. (1978). “Respondent Burden.” Proceedings of the Section on Survey Research Methods, American Statistical Association, 35—40.

Bradburn, N.M., Rips, L.J., & Shevell, S.K. (1987). “Answering autobiographical questions: The impact of memory and inference on surveys.” *Science* , 236.

Branden, L., Gritz, R. M., and Pergamit, M. R. (1995). "The Effect of Interview Length on Nonresponse in the National Longitudinal Survey of Youth," Proceeding of the 1995 Census Bureau Annual Research Conference, U.S. Bureau of the Census, pp. 129-154.

Champion, D. and Sear, A. (1969). "Questionnaire Response Rate: A Methodological Analysis," *Social Forces*, 47, 335-339.

Collins, M., Sykes, W., Wilson, P., and Blackshaw, N. (1988). “Nonresponse: The UK experience.” In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls II and J. Waksberg (Eds.), *Telephone Survey Methodology* (pp. 213-231). New York: Wiley.

Conrad, F.G., & Schober, M.F. (2000). “Clarifying question meaning in a household telephone survey.” *Public Opinion Quarterly, 64*, 1-28.

Dashen, M. and Fricker, S. (2001). “Understanding the cognitive processes of open-ended categorical questions and their effect on data quality.” *Journal of Official Statistics, 17*, 457–477.

Dillman, D., Sinclair, M. and Clark, J. (1993), “Effects of Questionnaire Length, Respondent-Friendly Design, and a Difficult Question on Response Rates for Occupant-Addressed Census Mail Surveys," *Public Opinion Quarterly*, 57, 289-304.

Galesic, M. (2006). “Dropouts on the Web: Effects of Interest and Burden Experienced During an Online Survey.” *Journal of Official Statistics*, *22*, 313-328.

Ghosh, D. and Vogt, A. (2000). “Determining an optimal split for a lengthy questionnaire.” Paper presented at the American Statistical Association (ASA) meetings, Indianapolis.

Gonzalez, J. M. and Eltinge, J. L. (2007a). “Multiple Matrix Sampling: A Review.” Proceedings of the Section on Survey Research Methods, American Statistical Association, 3069–75.

Gonzalez, J. M. and Eltinge, J. L. (2007b). “Properties of Alternative Sample Design and Estimation Methods for the Consumer Expenditure Surveys.” Paper presented at the 2007 Research Conference of the Federal Committee on Statistical Methodology, Arlington, VA, November, 2007.

Gonzalez, J. M. and Eltinge, J. L. (2008). “Adaptive Matrix Sampling for the Consumer Expenditure Quarterly Interview Survey.” Proceedings of the Section on Survey Research Methods, American Statistical Association, 2081–8.

Gonzalez, J. M. and Eltinge, J. L. (2009). “Imputation Methods for Adaptive Matrix Sampling.” A poster presented at the 2009 Joint Statistical Meetings, Washington, DC, August.

Groves, R., Singer, E., and Corning, A. (1999). “A Laboratory Approach to Measuring the Effects on Survey Participation of Interview Length, Incentives, Differential Incentives, and Refusal Conversion,” *Journal of Official Statistics*, 15(2), 251-268.

Groves, R., Fowler, F.J., Couper, M., Lepkowski, J., Singer, E., and Tourangeau, R. (2004). *Survey Methodology*. New York: Wiley.

Haraldsen, G. (2002). "Identifying and Reducing the Response Burden in Internet Business Surveys." Paper presented at the International Conference on Questionnaire Development, Evaluation and Testing Methods (QDET). Charleston, South Carolina, November 14-17.

Herzog, A.R. and Bachman, J.G. (1981). "Effects of Questionnaire Length on Response Quality." *Public Opinion Quarterly*, 45, 549-559.

Hubble, D. L. (1995). "The National Crime Victimization Survey Redesign: New Questionnaire and Procedures Development and Phase-in Methodology." Paper presented at the Joint Statistical Meetings, Orlando, FL.

Huttenlocher, J., Hedges , L.V., and Prohaska, V. (1988) . "Hierarchical organization in ordered domains: Estimating the dates of events." *Psychological Review*, *95*, 471 – 484.

Krosnick, J. (1999). "Survey research." *Annual Review of Psychology*, 50, 537-567.

Lee, L. and Carr, J. (2009). "Evaluation of a 12-Month Reference Period in the National Crime Victimization Survey (NCVS)," presented at the Federal Committee on Statistical Methodology Conference, Washington, DC. November.

Menon, G. (1994). "Judgments of Behavioral Frequencies: Memory Search and Retrieval Strategies," in *Autobiographical Memory and the Validity of Retrospective Reports*, Norbert Schwarz and Seymour Sudman, eds., Springer-Verlag, 161-172.

Menon, G. (1997). "Are the Parts Better than the Whole? The Effects of Decompositional Questions on Judgments of Frequency Behaviors." *Journal of Marketing Research*, 34, 335-346.

McCarthy, J. S., Beckler, D., and Qualey, S. (2006). "An Analysis of the Relationship Between Survey Burden and Nonresponse: If We Bother Them More, Are They Less Cooperative?" *Journal of Official Statistics*, *22*(1), 97-112.

Miller, P. V. & R. M. Groves. (1985). "Matching survey responses to official records: An exploration of validity in victimization reporting." *Public Opinion Quarterly*, 49, 366-380.

Navarro, A. and Griffin, R.A. (1993). "Matrix Sampling Designs for the Year 2000 Census." *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 480-5.

Neter, J., & Waksberg , J.. (1964). "A study of response errors in expenditures data from household interviews." *Journal of American Statistical Association, 59,* 18 -5 5.

Peytchev, A. and Tourangeau, R. (2005). "Causes of Context Effects: How Questionnaire Layout Induces Measurement Error." Presented at the American Association for Public Opinion Research (AAPOR) 60th Annual Conference.

Prohaska, V., Brown, N. R., & Belli, R. (1998). Forward telescoping: The question matters. *Memory, 6,* 455-465.

Raghunathan, T.E. and Grizzle, J.E. (1995). "A Split Questionnaire Survey Design." *Journal of the American Statistical Association*, 90, 54-63.

Roberts, C. Eva, G., Allum, N., and Lynn, P. (2010). "Data quality in telephone surveys and the effect of questionnaire length: a cross- national experiment," ISER Working Paper Series 2010-36, Institute for Social and Economic Research.

Sharp, L.M. and Frankel, J. (1983). "Respondent Burden: A Test of Some Common Assumptions." *Public Opinion Quarterly*, 47(1), 36-53.

Shields, J. and N. To (2005). "Learning to Say No: Conditioned Underreporting in an Expenditure Survey." American Association for Public Opinion Research Annual Conference, Miami Beach, American Statistical Association.

Tourangeau, R., Rips, L.J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Wedel, M. and Adiguzel, F. (2008), "Split Questionnaire Design for Massive Surveys," *Journal of Marketing Research*, 25(5), 608-617.

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

Zimmet, P. (2004). "Comparison of Fruit and Vegetable Frequency Data from Two Australian National Surveys." *Nutrition & Dietetics: The Journal of the Dietitians*, 61, 88-97.