The Golden Double Standard: Accounting for Social Desirability in Cross-Modal Comparisons

Steven H. Gittelman, Ph.D. - President, Mktg, Inc.
Victor Lange - Statistical Analyst, Mktg, Inc.
William Cook – CSO, e-Strategic Advantage
Susan Frede – VP, Research Methods & Best Practices, Lightspeed Research
Paul J. Lavrakas, Ph.D. - Independent Consultant, Visiting Scholar Northern Arizona U., and 2013-2014 AAPOR Past President
Christine Pierce – VP, Watch Methods, Nielsen
Randall K. Thomas-- VP, Online Research Methods, GfK

Introduction

Survey research relies heavily on the accurate capture and analysis of the words of others. Practitioners must carefully craft their survey questionnaires to avoid biasing respondents, to avoid confusing them, and to avoid boring them, all in the effort to achieve as accurate and pure a response as possible. Among the problems for survey researchers to address is response inaccuracy resulting not from the imprecision of our measures, but from the individuals we are measuring; i.e., from Respondent-related Measurement Error (cf. Groves, 1989). We often confront issues of data quality, where respondents may have to be removed from the final dataset because they were clearly not engaged enough with a study to properly consider and respond to its questions. Far more difficult to remove or even to detect, however, are those respondents who are diligent in their completion of the questionnaire but at times "distort" their answers when they are uncomfortable providing fully accurate responses to questions, especially regarding topics of a sensitive nature.

Surveys have changed little in light of this consideration, but the phenomenon of this "socially desirable" pattern of responding has not gone unstudied. As early as 1964, Crowne & Marlowe characterized "the need of subjects to respond in culturally sanctioned ways." Later research showed that social desirability was an attribute that could be differentially associated with survey items as well as individuals. That is, whereas certain persons are more likely than others to provide socially desirable responses, certain survey items also are more likely to elicit socially desirable responses than are other items (Philips & Clancy, 1972). Holding what has since proven to be a minority view of this phenomenon, these authors went as far as stating that the problem of social desirability was a threat to the entire concept of survey research, where no result could be taken at face value so long as it relied on an individual's self-report.

Views of social desirability in surveys have become more nuanced since that time. Today it is generally accepted that the tendency to respond in socially desirable ways is not uniform across all studies. According to Baker et al. (2010) "The *social desirability* hypothesis proposes that in the presence of an interviewer, some respondents may be reluctant to admit embarrassing attributes about themselves and/or may be motivated to exaggerate the extent to which they possess admirable attributes." Conversely, these effects do not emerge as strongly in self-administered survey modes that do not have an interviewer present. Topics that have been shown to display significantly less social desirability bias in self-administered modes include health (Baker, Zahs & Popa, 2004), donations to charity (Taylor, Krane & Thomas, 2005), and racially-motivated political views (Chang & Krosnick, 2009).

Though these papers may identify the presence of a social desirability bias and also report the degree of difference between modes of survey data collection, they did not provide an explicit endorsement for the quality of data collected through self-administered mail and online methods. Furthermore, there is a great deal of difference

between the probability samples employed in these studies and the non-probability samples employed in the surveys conducted in market research on a day-to-day basis. For this reason, it is difficult to quantify the degree to which the differences between estimates derived from online panels and those collected by phone and in-person methods is due to social desirability effects versus the share that is due to sample selection.

There have been previous attempts to resolve this issue. Frisina and Thomas (2007) attempted to scale item-specific social desirability across a set of ten behavioral survey questions by asking respondents how 'good' or 'bad'[1] each of the behaviors were on a five point scale. Although the method proved effective in predicting the direction of bias between phone and online data, it had difficulty predicting the magnitude of those differences.

Over the past 15 years, an increasing number of surveys have been conducted online and then compared to interviewer-administered survey results. To enable more apt comparisons between modes, we believe that we must be able to estimate not only the direction of bias but also the degree of bias. While many other differences exist between modes of survey administration (e.g. visual versus oral processing), the social desirability bias is probably one of the stronger distortions to control for in understanding modal differences, and may even outweigh the impact of sample selection bias across modes.

We have accordingly conducted an original study that gathered information about both the direction and magnitude of the social desirability bias that would be anticipated with specific survey items. We anticipate that such responses should be effective in predicting actual differences between data collected through different modes.

Data:

Measures of difference between phone and online survey modes were obtained from the Foundations of Quality 2 (FOQ2) survey study conducted by the Advertising Research Foundation in 2013. FOQ2 administered mostly identical online questionnaires to 70,377 respondents from 17 of the largest online suppliers in the United States, along with a shorter version of the questionnaire asked of 1,008 telephone respondents[2]. This was an omnibus study designed not only to examine responses across a variety of topics, but also across question types, and the survey was custom-built to allow for testing the effects of survey design decisions ranging from post-hoc weighting to quality control. Central to any comparison of data in this ARF study, however, is knowledge of the "correct" value for any question of interest. (Historically, deference has been given to RDD telephone research for greater accuracy, but research in the past decade has begun to question the validity of this assumption.)

To approach the question of socially desirable responding, we followed the example set by Frisina and Thomas by creating a questionnaire aimed at generating item-specific measures of sensitivity. However, we reasoned that the Frisina and Thomas method confounded respondents' moral compass with commonly held views by using a "Good-Bad" evaluation scale. For example, a socially conservative respondent might feel that homosexuality is a negative trait while a more liberal respondent would not, thereby creating a heterogeneous social desirability rating for this attribute. Nevertheless, both conservatives and liberals would likely agree that, on average, homosexuality is an attribute that survey respondents are more likely to falsely deny (i.e., False Negatives) than falsely claim (i.e. False Positives). To attempt to better capture this phenomenon, we asked respondents in our study to predict the tendencies of most survey respondents to distort their answers for a set of specific survey questions.

[1] The phrasing of the item was: "How good or bad would most people consider the actions listed below?"

[2] The phone dataset included a representative proportion of cell phone respondents following best practices for dual-Frame RDD surveys (cf. AAPOR, 2010). Both datasets received post-stratification weights for both the phone and online components to adjust for deviations between the final unweighted samples and the U.S. general population's characteristics for sex, age, region, ethnicity and education.

Our respondents were provided the following instructions:

> *You are going to be presented with a set of standard survey questions. In surveys like these, we often find that some respondents are uncomfortable with the questions asked or wish to make a positive impression on the interviewer. Because of this, they provide answers that are not completely accurate in order to portray themselves in a better light. Please read each question carefully and tell us, how likely it is that the average respondent would misrepresent their opinions and by how much?*

Respondents were presented with a 5-category response format with responses ranging from "Likely to give a much higher answer" to "Likely to give a much lower answer", adjusted as was necessary to conform to the question wording.[3] Our study questionnaire consisted of 28 items, using each of the benchmark-style questions that were contained in the ARF FOQ2 (17 for which it collected both online and phone data and 23 for which only online data were collected, but national U.S. benchmark data from external sources were available). Our study was administered to 998 online panelists provided by Survey Sampling International, balanced using sampling quotas to US proportions of sex, age, income, ethnicity, and region in an online questionnaire with an average duration of approximately ten minutes.

Results:

Table 1 displays the relative rankings of the 17 items tested in our study for which both ARF phone and online data are available. The left column presents the relative percentage difference between the FOQ2 phone and online mean values[4] and the right column displays how respondents rated the item relative to the midpoint in our evaluation survey. We refer to this measure as the item-specific social desirability rating (ISSDR)[5]. Negative ISSDR scores represent an expected under-reporting of the behavior in question (i.e., a tendency to expect others to falsely deny being associated with this behavior). Positive scores represent an expected over-reporting of the behavior in question (i.e., a tendency to expect others to falsely agree to being associated with this behavior). A score of zero represents no expectation of inaccuracy (i.e., a tendency to expect that others would not exhibit any social desirability in their responses about this behavior).

For example, the relative deviation between FOQ2 phone and online estimates for "Smoking Frequency" is -0.134, which constitutes one of the largest negative differences between the items we tested. That value means that respondents in the ARF telephone survey reported smoking significantly less frequently than did respondents in the ARF online survey. This corresponds with the expectations of the respondents in our survey – the fact that "Smoking Frequency" received the second-most negative ISSDR score (-0.46 as shown in the right column of Table 1) of items tested indicates the belief of respondents that it had a high likelihood of underreporting. The strong similarity in order between the two comparisons is compelling evidence that the magnitude of social desirability bias was successfully captured by the method. The correlation between the paired scores in the two columns of Table 1 is R=0.88. Thus there is a very high correspondence between what respondents in our experiment thought would happen and what was actually observed in the ARF surveys.

Table 1

Phone-Online Proportional Deviations and Ratings of FOQ2 Items

Item	***Phone-Online Deviation***	***ISSDR***

[3] A second version of this questionnaire was tested that separated out the concepts of direction and degree. The results were discouraging and this approach did not reach full fielding.

[4] Due to the diverse scaling of the 17 items, deviation was calculated as a relative proportion difference:

$$\left(\bar{x}_{(Phone)} - \bar{x}_{(Online)} \right) / \bar{x}_{(Phone)}$$

[5] Raw responses were on a 1-5 scale; values are reported as differences from the midpoint of 3.

Drinks in Past Year	-0.152	-0.59
Smoking Frequency	-0.134	-0.46
100 Cigarettes in Life	-0.087	-0.36
12 Drinks in Life	-0.029	-0.09
Have Cell Phone	-0.018	0.16
Hours of sleep/night	0.000	0.04
Days under-slept	0.001	0.09
Married	0.019	0.03
Driver's License	0.021	0.20
Proportion of calls on cell	0.052	0.09
Valid Passport	0.083	0.10
Self-reported health	0.090	0.52
Religiosity	0.117	0.44
Church Attendance	0.122	0.57
Hours working/week	0.127	0.26
Strengthening Phys. Activity	0.429	0.70
Vigorous Phys. Activity	0.454	0.63

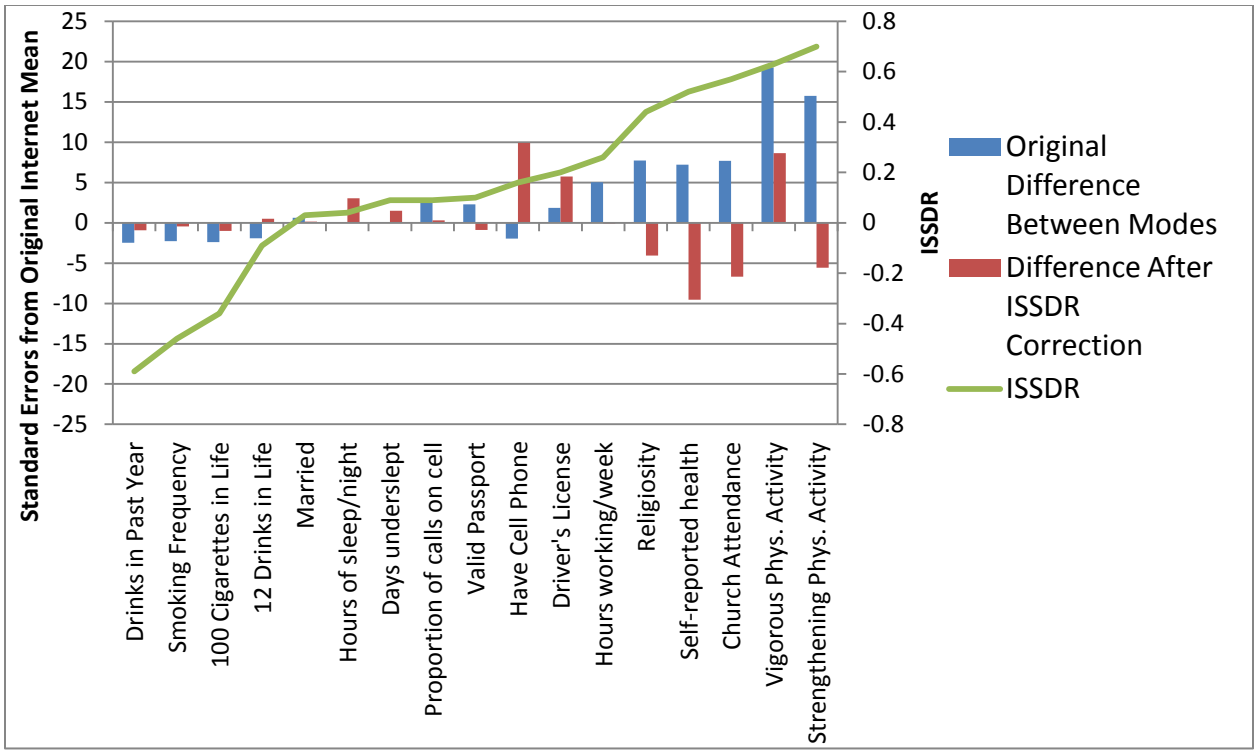
In order to convert this correlation into an actual predicted correction, we used a univariate regression of deviation on ISSDR with $R^2 = 0.71$, such that each of the 17 survey items tested was treated as an observation (results are reported in Table 2). The predicted values generated by this regression can be interpreted as the expected social desirability bias for each item. After removing this social desirability component from our phone estimates, the residual deviations should theoretically represent remaining bias due to other factors (i.e., ones other than social desirability, such as sample selection).

Table 2
Univariate Regression Results

Phone-Online Deviation	**β**	**Std. Err.**	**t**	**P>t**
ISSDR	-0.3780664	0.0590392	-6.4	<0.001
Constant	1.146786	0.1703642	6.73	<0.001
Adjusted R^2=0.71				

The magnitude of this "correction" for each of the 17 questionnaire items we tested is displayed in Figure 1. If social desirability is the dominant driver of difference between the modes, we would expect these "corrected" values to be much closer to the online mode than they were originally. While the model manages to reduce much of the bias from our original estimates in an absolute sense, it has no shortage of problems, largely resulting from overcorrections to measures near the top of the ISSDR scale. In the case of self-reported health, for example, applying our derived social desirability "correction" resulted in a mean value that was further from the original online mean, but in the opposite direction, going from a significantly higher to a significantly lower value.

Figure 1
Bias Corrections using univariate regression



However, this univariate method is only the simplest application of the data at our disposal. It is logical to compensate for respondents' nonlinear usage of the ISSDR scale by applying a quadratic transformation that emphasizes extreme responses. Additionally, by creating an interaction between the standard deviation of each item (obtained within the FOQ2 data) with ISSDR, we can moderate our corrections to conform to the manner in which the various question scales in the FOQ2 items were used by employing each item's standard deviation in an interaction (results are reported in Table 3). This new model achieves an R^2 of 0.96 and manages to remove almost all differences between our phone and online estimates, to a degree where only one measure retains a statistically significant difference between the two modes at $\alpha = 0.05$ (Figure 2). In this case, respondents in our survey indicated that they believed that cell phone ownership was likely to be overestimated by a moderate amount, while actual modal differences in the ARF data were very small. After applying our social desirability correction factor, the resulting prediction shows ownership among ARF online panelists to be significantly higher than among ARF phone respondents. It seems plausible that this may represent a true difference in sample characteristics when we consider that a set of respondents who all own computers are likely to own other technology at higher rates as well.

Table 3

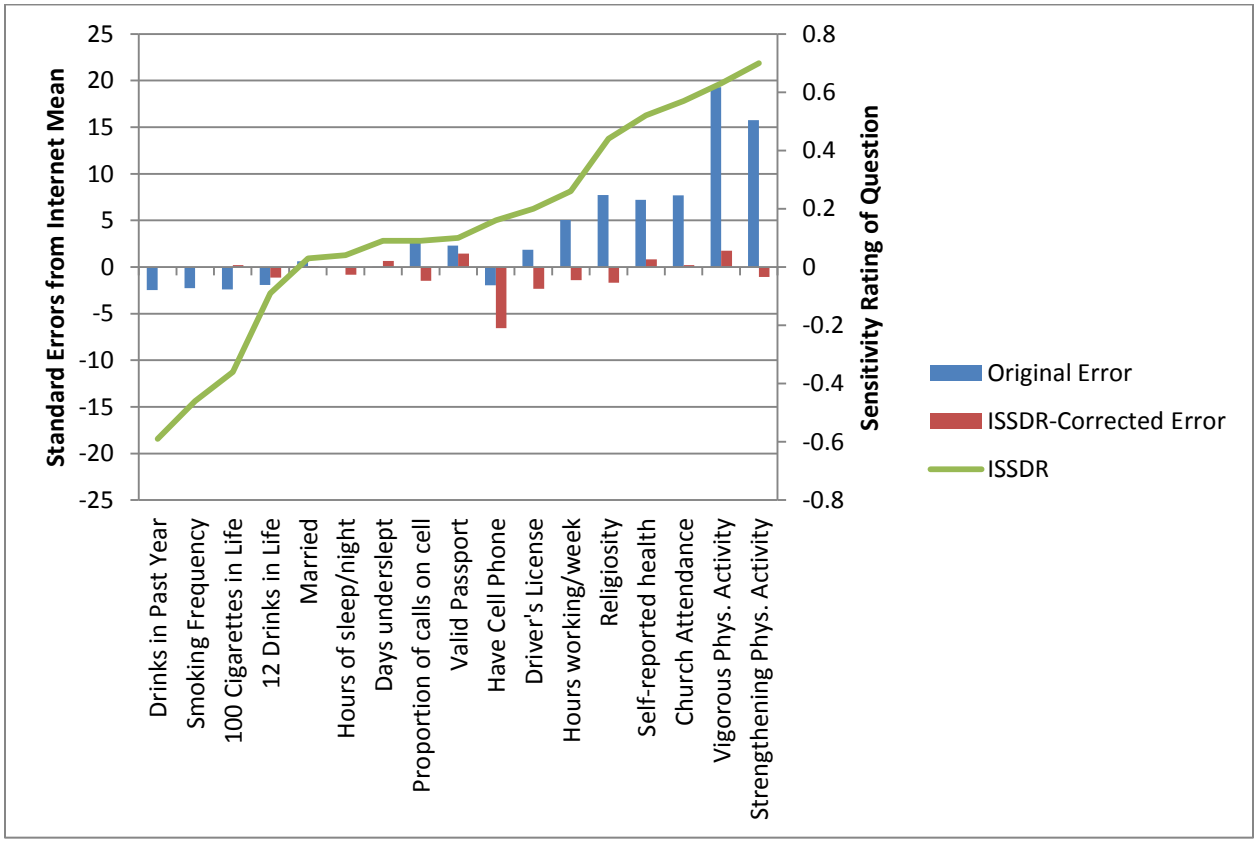
Multivariate Nonlinear Regression

Phone-Online Deviation	**β**	**Std. Err.**	**t**	**P>t**
ISSDR	0.9676106	0.5989531	1.62	0.134
ISSDR[26]	-0.1971139	0.1011118	-1.95	0.077
Standard Deviation	0.7996123	0.1172743	6.82	<0.001
ISSDR X Standard Deviation	-0.5074258	0.0763459	-6.65	<0.001
ISSDR^2 X Standard Deviation	0.0793256	0.012191	6.51	<0.001

[6] Though ISSDR is reported elsewhere in the paper as being zero-centered, it is actually a positive value centered around "3," because it is based on a 1-5 scale with the value "3" being the scale midpoint and meaning no social desirability expected. This avoids any problems of sign inherent in a quadratic transformation.

Constant	-1.1154	0.8804368	-1.27	0.231
Adjusted R^2=0.96				

Figure 2
BiasCorrections using multivariate non-linear regression



The larger implication of the scale and apparent accuracy of these corrections is that the lion's share of the differences we observe between modes seems attributable to social desirability bias and *not* other factors (e.g. sample selection differences).

We cannot expect this finding to hold beyond the set of questions we tested in our experiment. For example, the set of items used include virtually nothing pertaining directly to internet usage, which is the area where we would expect authentic differences between respondents surveyed through the two modes to be most prevalent. Nevertheless, the topics that we tested do represent a fairly broad range of lifestyle measures and our results suggest that there are fewer authentic differences than has been argued in the past.

Benchmarks

In addition to establishing a substantial amount of social desirability bias within the FOQ2 telephone data, these results also have implications for data collected in the context of a more rigorously executed Dual Frame RDD survey. Theoretically, it may be the case that a telephone survey conducted to more exacting standards, with response rates upwards of 40%, may be far more defensible as a national standard than that of a commercially generated random sampling of 1,000 phone respondents with a single-digit response rate. We would not expect a more representative sample frame to impact response error, however.

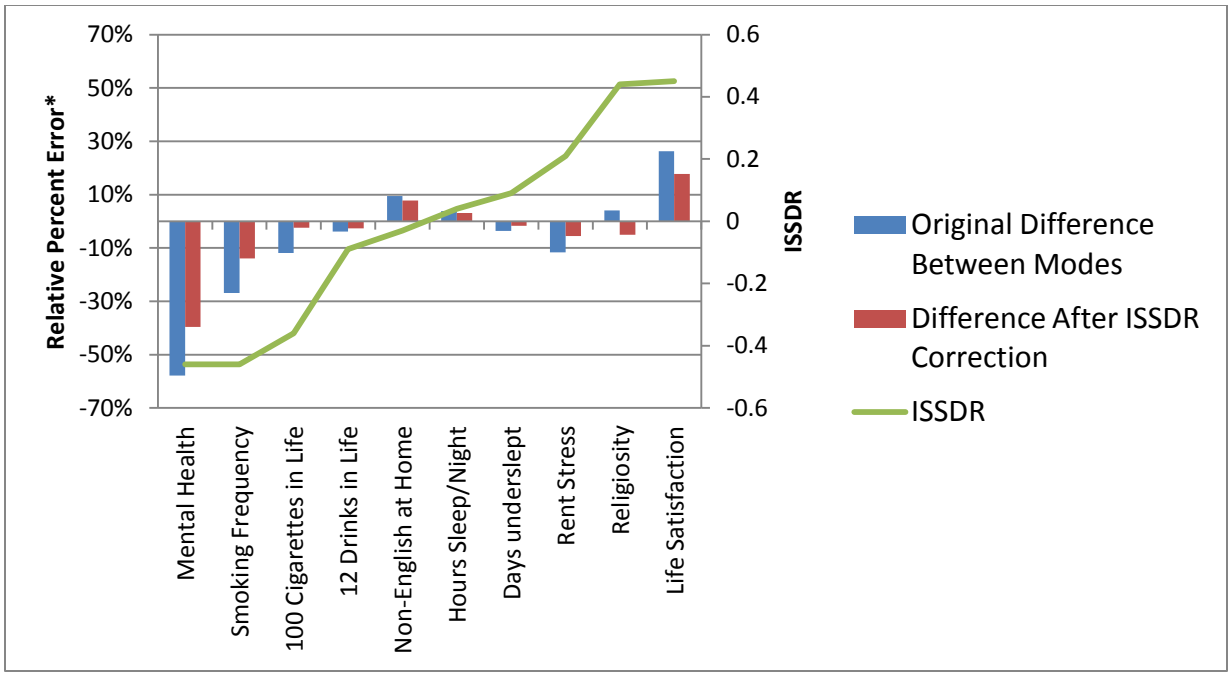
In order to test whether the influence of ISSDR extends to data collected by more rigorous means, we made use of the same set of 28 benchmarks that were contained in the FOQ2 survey. We drew mean estimates from four major surveys: the Behavioral Risk Factor Surveillance System (BRFSS), the National Health Interview Survey (NHIS), the American Community Survey (ACS), and the General Social Survey (GSS)[7]. Though these questions covered many topics contained within these studies, the question wording and response scales used were not all identical. These differences make for a far less ideal testing ground for our theory than the FOQ2 data, which was not only identically worded and scaled[8], but also collected at identical time periods, while some benchmark data was as many as three years old.

In order to evaluate the similarity of the FOQ2's chosen question format to those of available benchmark surveys, six experienced industry professionals from FOQ2's analysis group were asked to rate each item on a 1-9 scale[9]. Using the combination of these six independent evaluations, the total set of 28 eligible questions was broken into two sets. Ten questions received average scores of greater than 7 and were treated as functionally identical. The remaining 18 questions varied from receiving very low comparability scores to being very near this arbitrary cut-off point, but this cut-off was strictly enforced to avoid the problem of justifying our results in the context of any substantial differences in question wording.

Once a dataset of comparable questions was created, the multivariate model from the previous section was applied to this subset of identical questions in an attempt at cross-validation. Among these ten items, the predictions made in the initial phone versus online model continue to explain a significant share of the deviation from FOQ2 estimates, as is shown in Figure 3. The correlation between the predictions of the model and observed deviations is R=0.77, with the most notable over-correction being Religiosity. An additional and more problematic shortcoming was that a few measures, especially those not used in modeling the FOQ 2 internal comparison, were substantially under-corrected by the model.

Figure 3
Predictions of the ISSDR model versus deviations from benchmark values

[7] The BRFSS is a 500,000 interview survey conducted annually by telephone. The NHIS is a 35,000 household survey conducted annually in person. The ACS is an interview of 2 million conducted by mixed mode (see http://www.census.gov/acs/www/Downloads/ACS_Information_Guide.pdf for more details). The GSS is 5,000 interview study conducted in person. For specific breakdowns of the variables and benchmark equivalents used, see appendix Table A1.
[8] Excluding the inclusion of a volunteered 'Refusal' option as is required in an interviewer-assisted context.
[9] Such that an item rated 1 when corresponding items were "So different that results would clearly be biased or simply cannot be compared due to scaling differences" and an item rated 9 was a comparison "Sufficiently similar to produce effectively identical results"



*- Note that the error scale along the y-axis has changed, as the standard errors of benchmark estimates aretoo sensitive to form a manageable scale

A sample of ten observations is sufficient to support the initial finding that observed deviations from interviewer-assisted (in this case phone) modes of research are correlated with expectations of bias. Ideally, a model could be formulated that better matched the scale of deviations across modes and across questions not originally included in its formulation. The seventeen cross-modal items we used are probably too few to build such a broadly applicable model. Additionally, many items with the greatest ISSDR in the survey set were excluded from the FOQ2 phone study to reduce the length of the interview. Consequently, when those items are reintroduced to the model, it provides a much smaller magnitude of correction than observed in the pure phone versus online model.

Further research may also approach the specific question wordings used in the ISSDR survey. Although we believe that the results we obtained are compelling, little experimentation went into identifying the most effective language for prompting the desired results. Increased study of potential metrics could identify a superior question format that would increase respondent comprehension of the task, perhaps leading to less reliance on neutral responses and a more dynamic social desirability scale.

Discussion

It may be a foregone conclusion to many readers that online research is destined to replace traditional phone research in all but the most exceptional circumstances. In large part, this transition has already occurred. Nevertheless, the specter of phone research always looms over any attempt at attaining population or political data online. Practitioners and analysts remember a time when data were rooted in probability theory and when inferences about the population were made from it without caveats. On average, the differences between the modes are negligible, but whenever a large discrepancy emerges, it casts doubt (perhaps unfairly) on online research as a whole.

Our findings should not be interpreted as a claim that online research has somehow overcome its shortcomings. In many ways, the average online study is no more scientific in its execution than it was when the mode was in its

infancy. What may need to change is the reverence often given to research collected via telephone (and possibly face-to-face) by comparison. Whether a small study conducted to the standards of market research or a flagship study conducted by a government agency, the social desirability effects of interviewer-assisted modes may play at least as large a result in modal discrepancies as any form of selection bias. What may be more concerning is that unlike telephone's more recent issues with expanding non-coverage, the effects of social desirability most likely reach far back into past decades of interviewer-assisted survey research. This makes parsing out the extent to which the differences we obtain may be due to social desirability versus sample differences is an important endeavor.

The magnitude to which modal differences are predicted by the very simple ratings of the concept we identify as item-specific social desirability suggests it plays a dominant role in the results of phone surveys. This does not prove that online (or any self-administered mode) will provide an unbiased measurement of any sensitive concept. It does, however, indicate problems with the practice of using standard phone surveys in the pursuit of a representative standard without taking the concept of social desirability into account. In this climate of uncertainty, establishing reliable benchmarks will require the experience of an expert practitioner, capable of applying corrections of the sort suggested either in this research or that which follows it.

The presence of a viable procedure for correcting estimates of social desirability should have far-reaching ramifications. With the organizations behind a number of venerable phone and in-person studies (such as the Behavioral Risk Factor Surveillance System and American National Election Study) examining the consequences of switching to more cost-effective modes (often self-administered and online), it is important that the issue of making proper comparisons does not go overlooked. Understanding the magnitude of bias present in interviewer-assisted modes and the degree of correction required for sensitive questions may improve the accuracy of measurement in all modes of interviewing.

Appendix:

Table A1

Comparison of Survey Items to Benchmark Values

Identical Questions				
FOQ Question	**Scale**	**Reference Source**	**Reference Question**	**Reference Scale**
During the past 30 days, for about how many days have you felt you did not get enough rest or sleep?	0-30 days	BRFSS 2011 - QLREST2	During the past 30 days, for about how many days have you felt you did not get enough rest or sleep?	0-30 days
How often in the past 12 months would you say you were worried or stressed about having enough money to pay your rent/mortgage? Would you say that you were worried or stressed…?	Always->Never (5 point scale)	BRFSS 2011 - SCNTMONY	How often in the past 12 months would you say you were worried or stressed about having enough money to pay your rent/mortgage? Would you say you were worried or stressed---	Always-> Never (5 point scale)
Now thinking about your mental health, which	0-30 days	BRFSS 2011 - MENTHLTH	Now thinking about your mental health, which includes stress,	0-30 days

includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?			depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?	
Next, we'd like to ask you about some of your general feelings and perceptions. <p>In general, how satisfied are you with your life? [TELEPHONE ADD: Are you very satisfied, satisfied, dissatisfied, or very dissatisfied?	Very satisfied-> Very dissatisfied (4 point scale)	BRFSS 2010 - LSATISFY	In general, how satisfied are you with your life?	Very satisfied-> Very dissatisfied (4 point scale)
Have you smoked at least 100 cigarettes in your ENTIRE LIFE?	Yes/No	BRFSS 2011- smokek100	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]	Yes/No
Do you smoke cigarettes every day, some days, or not at all?	Every day-> Not at all (3 point scale)	BRFSS 2011 - SMOKDAY2	Do you now smoke cigarettes every day, some days, or not at all?	Every day-> Not at all (3 point scale)
Have you had at least 12 alcoholic drinks over your ENTIRE LIFE?	Yes/No	NHIS 2011 - ALCLIFE	In your ENTIRE LIFE, have you had at least 12 drinks of any type of alcoholic beverage?	Yes/No
Do you speak a language other than English at home?	Yes or No	ACS Q14	Does this person speak a language other than English at home	Yes/No
On average, how many hours of sleep do you get in a 24-hour period?	Hours/Minutes per 24-hour period	NHIS - ACISLEEP	On average, how many hours of sleep do you get in a 24-hour period?	1-24 hours, round up minutes
To what extent do you consider yourself to be a religious person?	Very-> Not at all (4 point scale)	GSS-RELPERSON	To what extent do you consider yourself a religious person? Are you . .	Very-> Not at all (4 point scale)
Non-Identical Questions				

On how many days during the PAST 30 DAYS, how often did you do the following kinds of physical activity? A VIGOROUS leisure-time physical activities for AT LEAST 10 MINUTES that cause HEAVY sweating or LARGE increases in breathing or heart rate?	0-30 days	NHIS 2011 [Adult File] - vigfreqw	How often do you do VIGOROUS leisure-time physical activities for AT LEAST 10 MINUTES that cause HEAVY SWEATING or LARGE increases in breathing or heart rate?	Respondent chooses time period with which to phrase their response
Physical activities specifically designed to STRENGTHEN your muscles (such as lifting weights, doing calisthenics, etc.)?	0-30 days	NHIS 2011 [Adult File] - strfreqw	How often do you do LEISURE-TIME physical activities specifically designed to STRENGTHEN your muscles such as lifting weights or doing calisthenics? (Include all such activities even if you have mentioned them before.)	Varies: See above
In the PAST 12 MONTHS (365 days), on how many different DAYS would you estimate that you drank any type of alcoholic beverage? Your best guess is fine.	0-365 days	NHIS 2011 - ALC12MYR	In the PAST YEAR, how often did you drink any type of alcoholic beverage?	Varies: See Above
About how many minutes or hours in a Typical Day do you do each of the following? Watch Television	Do not do this at all -> More than 5 hours a day (8 points scale)	NHANES - PAQ.710	Over the past 30 days, on average how many hours per day did {SP} sit and watch TV or videos? Would you say…	less than 1 hour; 1 hour, 2 hours; 3 hours, 4 hours; 5+ hours; none, do not watch TV or Videos
In the past 7 DAYS, about how long in total did you spend working or doing any work-related activities for pay? Please include	Hours and Minutes	GSS - HRS1, HRS 2	How many hours did you work last week, at all jobs?; How many hours a week do you usually work, at all jobs?	0-89 Hours

total time for all the jobs you work for pay.				
Do you own or rent your home?	Own/Rent/Other Arrrangement	GSS-DWELOWN	Do you/Does your family) own your (home/apartment), pay rent, or what?	Own/Buying, Rent, Other
Are you currently married?	Yes or No	ACS - Q20	What is this person's marital status?	Now married, Widowed, Divorced, Seperated, Never Married
Do you, or does anyone else in the household, have any of the following conditions? Blindness or serious difficulty seeing even when wearing glasses	Yes or No	ACS Q17b	Is this person blind or does he/she have serious difficulty seeing even when wearing glasses?	Yes/No
Do you, or does anyone else in the household, have any of the following conditions? Deafness or serious difficulty hearing	Yes, I have this; Yes, another member of the household has this; Yes, both I and another member of the household have this; No, no one in the household has this	ACS Q17a	Is this person deaf or does he/she have serious difficulty hearing?	Yes/No
Do you, or does anyone else in the household, have any of the following conditions? A serious difficulty in walking or climbing stairs	Yes, I have this; Yes, another member of the household has this; Yes, both I and another member of the household have this; No, no one in the household has this	ACS Q18b	Does this person have serious difficulty walking or climbing stairs	Yes/No
Do you, or does anyone else in the household, have	Yes, I have this; Yes, another member of the household has this; Yes,	ACS Q18a	Because of a physical, mental, or emotional condition, does this	Yes/No

any of the following conditions? A serious difficulty in concentrating, remembering, or making decisions because of a physical, mental, or emotional condition	both I and another member of the household have this; No, no one in the household has this		person have serious difficulty concentrating, remembering or making decisions?	
How well can you read a newspaper or book in…? A. English B. Spanish	Not at all -> Very Well (4 Point Scale)	Pew	Would you say you can read a newspaper or book in English:	very well ->not at all (4 Point Scale)
How much do you weigh without shoes on?	Pounds:	NHANES - WHQ.025	How much {do you/does SP} weigh without clothes or shoes? [If {you are/she is} currently pregnant, how much did {you/she} weigh before your pregnancy?	Pounds or Kilograms
Of all the telephone calls that you or other members of your home receive, how many are received on a cell phone or smartphone?	All recieved on cell phones/smartphones; Almost all calls are recieved on cell phones/smartphones; Some recieved on cell phones/smart phones and some on regular phones; Very few are recieved on cell phones/smartphones; None on cell phones/smartphones	NHIS - PHONEUSE	Of all the telephone calls that you or your family receives, are…	All or almost all calls received on cell phones; Some received on cell phones and some on regular phones, Very few or none on cell phones
EXCLUDING YOU, how many other adults (age 18 or over) live in your household?	0-25 Adults	ACS 2011	How many people are living or staying at this address?	Open Ended Response

How many bedrooms are in your house, apartment, or mobile home? That is, how many bedrooms would you list if your house, apartment, or mobile home were on the market for sale or rent?	1 -> 5+ Bedrooms (5 points scale)	ACS 2011	How many of these rooms are bedrooms? Count as bedrooms those rooms you would list if this house, apartment, or mobile home were for sale or rent. If this is an efficiency/studio apartment, print "0"	Open Ended Response
How many automobiles, vans, and trucks of one-ton capacity or less are kept at home for use by members of your household?	0-20	ACS 2011	How many automobiles, vans, and trucks of one-ton capacity or less are kept at home for use by members of this household?	None -> 6 or more (7 point scale)

Works Cited:

Baker, R., Zahs, D., &Popa, G. (2004). 'Health Surveys in the 21st Century: Telephone vs. Web. In *Eighth Conference on Health Survey Research Methods, Peachtree City, GA* (Vol. 34).

Baker, R., Blumberg, S., Brick, J.M., Couper, M., Courtright, M., Dennis, M., Dillman, D., Frankel, M., Garland, P., Groves, R., Kennedy, C., Krosnick., J., Lee, S., Lavrakas, P.J., Link, M., Piekarski, L., Rao, K., Rivers, D., Thomas, R.K. &Zahs, D. (2010). *AAPOR Report on Online Panels.* Retrieved from the American Association of Political Opinion Research website:http://www.aapor.org/AAPOR_Releases_Report_on_Online_Survey_Panels1.htm

Chang, L., &Krosnick, J. A. (2009). National surveys via RDD telephone interviewing versus the internet comparing sample representativeness and response quality. *Public Opinion Quarterly*, *73*(4), 641-678.

Crowne, D.P. & Marlowe, D. (1964).*The Approval Motive: Studies in Evaluative Dependence*. New York: Wiley.

Frisinia, L., Thomas, R.K., Krane, D. & Taylor, H. (2007) Scaling Social Desirability: Establishing Its Influence Across Modes. *American Association of Public Opinion Research Annual Conference,* Anaheim, CA.

Groves, R. M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley and Sons.

Phillips, D.L., and Clancy, K.J. Some effects of"social desirability" in survey studies. *American Journal of Sociology* (1972): 921-940.

Taylor, H., Krane, D., & Thomas, R.K. (2005). "Best Foot Forward: Social Desirability in Telephone vs. Online Surveys." *Public Opinion Pros*, Feb. Retrieved from http://www.publicopinionpros.com/from_field/2005/feb/taylor.asp.