Measurement Error in Panel Data: A Comparison of Face-to-Face and Internet Survey Samples

Ariela Schachter, Stanford University
Department of Sociology
450 Serra Mall
Stanford, CA 94305

arielas1@stanford.edu

DRAFT: 3/24/2015

Abstract

Despite recent research demonstrating that national-probability internet samples are just as accurate as Random-Digit Dialing telephone surveys, many scholars remain skeptical of all internet-based surveys. Using panel data from the General Social Survey (a face-to-face interview survey) and the How Couples Meet and Stay Together Survey (a national-probability internet survey conducted by Knowledge Networks/GfK), I compare measurement error in common demographic variables from wave to wave of both surveys, and find that error rates are lower in the national probability internet survey. I also find that being less-educated is associated with a higher probability of error in the internet survey, but not the face-to-face survey. The results suggest that overall skepticism of national-probability internet samples is unwarranted, but that the accuracy of survey modes may vary depending on respondent characteristics.

1

The quality of internet-based surveys remains a contested issue among social scientists. This research note contributes to the debate by comparing measurement error on key demographic characteristics in national-probability face-to-face and internet panel surveys. Measurement error is a serious problem for researchers (Bound, Brown and Mathiowetz 2001) but is not commonly the focus of survey mode effect studies. I surprisingly find lower error rates in the internet panel, and the results suggest that skepticism towards national-probability internet surveys is unwarranted.

Survey Modes and Measurement Error

Face-to-face interviews of national probability samples remain the gold-standard in survey research (e.g. Holbrook, Green and Krosnick 2003), but their high cost and logistic difficulties have led researchers to seek alternatives. Random Digit Dialing (RDD) telephone surveys are slightly less accurate than face-to-face interviews and more cost-effective (Groves 1990), and today they are widely accepted as a valid method of data collection. Recently scholars have compared RDD telephone surveys to internet surveys, finding that national probability internet samples are more accurate than RDD telephone surveys (Chang and Krosnick 2009; Baker et al. 2010). Similar comparisons of face-to-face and national probability internet survey modes have found mixed results depending on the types of questions being asked (Duffy et al. 2005), but generally conclude that face-to-face surveys remain more accurate (Heerwegh and Loosveldt 2008). While the research is not definitive, it suggests that national probability internet surveys are at least as good as RDD telephone surveys, but not as good as face-to-face interview surveys. In spite of these findings, some scholars remain skeptical of all internet-based surveys.

Survey modes can differ along many dimensions, including respondent coverage, response rate, item nonresponse, and measurement error (Groves 1990), making it difficult to compare across modes and diagnose the source of any differences (Biemer 2001). Panel data provides unique leverage by making it possible to observe measurement error by capturing the extent to which theoretically unchanging characteristics or deterministically changing characteristics (eg, age) vary from survey wave to wave (Biemer 2001). Nevertheless, even panel data confounds measurement error with other issues, including attrition, question wording, response rate and respondent coverage. In sum, while panel data allows for a better comparison of measurement error rates across survey modes, it is difficult to definitely identify why differences by survey mode exist.

Also, survey mode effects may be particularly strong for less-educated respondents with lower cognitive abilities (Chang and Krosnick 2010; Malhotra 2008). Face-to-face interviews do not require respondents to read and comprehend survey questions without the assistance of an interviewer, as is the case in internet surveys. This suggests that less-educated respondents will have less difficulty completing face-to-face interviews and less measurement error relative to internet surveys (Malhotra 2008). On the other hand, in an experimental trial, Chang and Krosnick (2010) find that respondents with lower cognitive skills produce better-quality data in internet surveys relative to face-to-face interviews, and posit that face-to-face interviews have higher cognitive demands because respondents must remember the question and outcomes without looking at a screen for reference. These opposite findings may be driven by different definitions of 'lower cognitive ability', as Chang and Krosnick (2010) use only college students and differentiate by SAT score, while Malhotra (2008) uses nationally-representative data and

differentiates by education-level. Regardless, both studies suggest that survey mode effects on measurement error may vary by respondent education.

Data

Since 1972 the General Social Survey (GSS) has collected information on demographics and attitudes from a nationally representative sample of adults living in the United States using face-to-face interviews. Starting in 2006, the GSS switched from a purely cross-sectional design to include an additional panel data design. Of the 2,000 respondents who participated in 2006 (with a 71.2% response rate; (AAPOR RR5)) [1], 1,536 were re-interviewed in 2008 (which is a 76.8% response rate among those who participated in wave 1), and 1,276 were re-interviewed in 2010 (an 83% response rate among those who participated in the first two waves).[2] Overall, 63.8% of respondents who participated in wave 1 completed waves 2 and 3 (GSS 2011; GSS 2012). Only respondents who provided valid responses on each measure for each wave are examined, yielding a maximum sample size of 1,276 in all analyses of the GSS data.[3]

The 'How Couples Meet and Stay Together" (HCMST) survey (Rosenfeld, Thomas and Falcon 2011) is a nationally-representative, longitudinal, internet survey implemented by GfK (formerly known as Knowledge Networks). The HCMST consists of 4,002 English-literate adults who were first interviewed in 2009, and oversamples gay, lesbian and bisexual respondents. The HCMST is used in order to access the basic demographic characteristics

---

[1] The 2006 GSS was initially fielded as a cross-sectional sample (n=4,510), and the 2,000 panel respondents were randomly selected from these respondents. The response rate is based on the entire cross-sectional sample.

[2] Smith and Son (2010) analyze the first two waves of the GSS panel and find only small differences by attrition status, such that less educated, younger, and non-married respondents—people with fewer social connections—are more likely to drop out of the study. These characteristics commonly predict attrition or non-response and indicate that the attrition in the GSS panel data is common to all panel studies; nevertheless, the analyses presented here do not account for differences in attrition across the two surveys, a potentially important source of any mode effects.

[3] Smith (2009) describes the re-contact procedures used by the GSS to ensure that the same person is re-interviewed at each wave, and concludes that the measurement error present in the GSS panel is not systematic or driven by the wrong person being interviewed.

included in all GfK panel surveys. GfK panel participants are initially recruited through a nationally representative RDD telephone survey in order to create a nationally representative panel. Respondents can either use their own computer with internet access to answer survey questions or are provided with internet access and a WebTV in exchange for their participation.

Follow-up surveys were conducted in 2010 and 2011; [4] while the HCMST only re-interviewed respondents in romantic relationships, core demographic characteristics are potentially available for all 4,002 respondents at each wave because these demographic variables are collected separately from the HCMST-specific questions. [5] The 2010 sample includes valid demographic data for 3,693 respondents, and the 2011 sample includes 2,597 valid respondents, yielding a maximum sample size of 2,597 respondents.

The HCMST response rate must be calculated in several stages. First, the initial RDD phone survey response rate is 32.6%. Of those who responded to the phone survey, 56.8% agreed to become GfK panel members. Next, 71% of GfK panel members agreed to participate in the first wave of the HCMST. Combined, this means that the 2009 HCMST response rate is 13% (.326*.568*.71=.13; AAPOR CUMRR2). Among those who participated in the first HCMST wave, 92.3% responded to the second wave in 2010, and 70.3% of respondents who participated in the first two waves also participated in wave 3. Overall, among those who participated in the initial 2009 survey, 64.9% continued to participate in waves 2 and 3. For more information on the HCMST see http://data.stanford.edu/hcmst. [6]

---

[4] The GSS panel interviews respondents once every other year, while the HCMST interviews respondents yearly. This difference may also contribute to any differential measurement error rates.
[5] Background demographic data is not collected at the same date/time as HCMST-specific questions, but the date of background data collection is available for each wave.
[6] HCMST data is publically available, but age and gender data from waves 2 and 3 are not. Please see http://data.stanford.edu/hcmst for information on restricted-access variables.

Based on initial response rates, there are clear differences between the two surveys. The low initial response rate in the GfK panel (13%) is one of the perceived drawbacks to internet surveys, but response rate is not a good indicator of accuracy or representativeness (Krosnick 1999). Also, while the overall response rates are not directly comparable because of differences in recruitment strategies, I can compare attrition rates. Here the two surveys are more similar, with the GSS at 36.2% of respondents attriting by wave 3, and the HCMST at 35.1%. This suggests that any differences in measurement error are not driven by attrition.

Measures

Four common demographic measures are used in the analysis. All of the measures were re-collected in each wave of each survey, and neither survey firm did any post-data-collection manipulation of these measures to artificially lower the number of errors.[7] Exact wording and response categories for each question are reported in the Appendix.

First, two characteristics that are unlikely to change from wave to wave are examined: race and gender.[8] In both surveys race is measured using a standard fifteen-category Census race question combined with a separate Hispanic ethnicity question. These measures are recoded into *White, Black, Other* and *Latino* in the GSS and recoded into *White, Black, Other, Latino and Two or more races* in the HCMST. Gender is coded as female (=1) in both surveys. An error is coded if the respondent changes race or gender categories at any point throughout the waves of the surveys.

---

[7] Information verified via personal email correspondence with GfK and GSS representatives.
[8] Both race and gender are not inherently fixed characteristics, as demonstrated in work by Saperstein and Penner (2012) and Schilt and Westbrook (2009); however, the proportion of the population experiencing real changes in their racial or gender identities is quite small, such that most changes from wave to wave are likely measurement error.

Next, two characteristics that can only change deterministically are examined. Age is reported in years in both surveys. An error is coded if the respondent reports a decrease (of any magnitude) in age or if the respondent reports increases in age that are impossible based on the timing of each survey wave (3 years for the GSS and 2 years for HCMST). Finally, highest degree is a categorical measure in both surveys; five categories are reported in the GSS based on responses to a set of questions on educational attainment, and a single question with fourteen response categories are recorded in the HCMST. HCMST is recoded into the GSS's five-category measure of: *less than high school, high school graduate, some college, college graduate,* and *graduate degree*. Highest degree can increase from wave to wave; thus, an error is coded only if the respondent reports a decrease.

Method

I first compare error rates on each measure from each survey using z-tests for differences in proportions that take into account the different sample sizes of the two surveys. Second, I compare results from bivariate logistic regression models predicting errors, where each of the four variables is the dependent variable, and years of education (a continuous measure in both surveys) is the independent variable. Here I am interested in whether there is a pattern to the error rates such that less educated respondents are more likely to report errors.

For the models predicting errors in the race, gender, and age variables, I only include respondents who did *not* have an error in their highest degree variable, to ensure that I am capturing the actual relationship between education and error rate. This restriction is impossible for the model predicting errors in highest degree; however, I re-estimated the models for race, gender, and age without this restriction and found no substantive differences, suggesting that the

education-error rate relationship is not driven by errors in the education variable. I also compare

the relationship between education and error across surveys using Wald tests for the equality of

coefficients, which account for the differences in sample size and standard errors when

comparing the two education coefficients for each error variable. [9]

Results

Table 1 reports the frequency and percent of errors for each variable for each survey,

along with tests comparing the error rates across surveys. The error rates for age and gender are

relatively low (below 5%) for both the HCMST and the GSS, but on both measures the GSS

error rate is significantly larger (p<.001). The error rate for highest degree is larger for both

surveys but again significantly larger in the GSS, at almost 5.9% in the HCMST and 11.8% for

the GSS (p<.001), translating to about 150 respondents from each survey reporting a decrease in

highest degree at some point. The results for these three measures demonstrate that measurement

error is surprisingly lower in HCMST, the internet survey, compared to the GSS' face-to-face

interviews.

[Table 1 about here]

However, the race error rate for HCMST is higher than the GSS at 5.43%, compared to

3.24% on the GSS (p<.001). The race error rate likely includes both true error and some small

rate of respondents actually changing their racial identifications (eg, Saperstein and Penner

2012). Thus, it is difficult to judge whether there is more measurement error in the HCMST or if

---

[9] In order to calculate these Wald tests I estimate a single model with cases from both datasets and test for an interaction effect between dataset and years of education. A significant interaction term is equivalent to a Wald test.

racial fluidity is captured at a higher rate via the internet survey mode, where respondents may

feel more comfortable reporting a change in their racial self-identification. [10]

[Table 2 about here]

Next, Table 2 reports results from bivariate logistic regression models using years of

education to predict errors in age, gender, race and highest degree in each survey, as well as

Wald tests comparing the years of education coefficients from each survey. Education has a

consistent, negative effect on the error rates in the HCMST, such that more educated people are

less likely to report errors. This difference is significant at the $p<.010$ level for age, race and

highest degree, but not gender. [11] However, education does not have a consistent effect on the

error rates in the GSS. It has virtually no effect on age and race errors, while there is a

significant, negative effect on gender such that more educated people in the GSS are less likely

to report gender errors ($p<.05$), and a significant, *positive* effect on highest degree, such that

more educated people are *more* likely to report decreases in their highest degree in the GSS

($p<.010$). Finally, given these different patterns it is not surprising to see that the education

effects are significantly different across the two surveys on all of the error variables except

gender, as demonstrated by the Wald tests reported in Table 2.

Alternative Source of Error: The Interviewer?

The lack of consistent explanatory power of respondent education on errors in the GSS

data begs the question: what does account for these errors? One set of likely culprits are the

interviewers. Recent work by Paik and Sanchagrin (2013) finds significant interviewer effects in

GSS data on social network size, suggesting that the errors identified in this analysis may be due

---

[10] There is no way to resolve this issue using the existing HCMST and GSS data, but it is a topic worthy of further research.

[11] The non-significant gender finding is likely due to the small number of gender errors (4) in the HCMST.

to a (small) number of error-prone GSS interviewers. Testing for interviewer effects is particularly important given the lack of an education effect in the GSS data; interviewers may introduce more overall measurement error, but level the playing field between more and less-educated respondents.

In order to examine interviewer effects in the GSS, I calculated the error rate per interviewer (for each panel wave), which I define as the proportion of interviews with at least one error per interviewer. While there are four GSS interviewers with error rates of 100%, all of these interviewers completed only one total interview. Similarly, there are about ten interviewers at each wave with error rates of at least 25%, but none of these interviewers completed more than 20 interviews (out of over 1200 interviews per wave), and the majority of them completed fewer than 10.[12] Put simply, while this pattern suggests that the least-experienced interviewers (based on number of completed interviews) may be more likely to record errors, by virtue of being the least-experienced interviewers, this does not account for the vast majority of errors identified in the GSS data. This preliminary evidence suggests that errors in the GSS data may be random; however, more work must be done to rule out other sources of systematic measurement error in the GSS.

Discussion

Several noteworthy findings emerge from this brief analysis. First, on three out of the four demographic variables (age, gender, and highest degree), there are significantly lower rates of measurement error in the internet survey compared to face-to-face interview survey, a clear contradiction of the existing literature. The only variable with a higher error rate on the internet survey is race; given recent research on racial fluidity (Saperstein and Penner 2012), whether the

---

[12] Full results from this analysis are available from the author upon request.

different error rates captured here reflect differences in measurement error or differences in reports of racial fluidity is unclear and worthy of further study.

Second, it is unclear why we observe higher overall error rates in the face-to-face interview survey. Measurement error in the GSS does not exhibit a consistent pattern with years of education or interviewer. I speculate that relative to internet surveys, face-to-face interviews may be more exposed to random measurement error because they depend on both the interviewer and respondent accurately relaying information to one another, rather than just the respondent. In a face-to-face interview there are more steps between a question being asked and a response being a recorded, thus creating more opportunities for random errors.

 On the other hand, the HCMST errors have a consistent, negative relationship with years of education, such that more educated people are less likely to have measurement error. This pattern suggests that internet survey data may be just as, if not more, reliable than face-to-face interview data overall, but less reliable for certain populations, such as the less educated. More research is needed to understand why internet surveys appear to be more cognitively demanding, and whether question designs, or options to have questions 'read' to respondents via audio-files, might reduce these demands. More generally, exploring whether there are interactions between survey mode and respondent characteristics is important as social scientists continue to work to survey harder-to-reach populations, and increasingly rely on internet surveys to do so.

Previous research has asserted that national-probability internet survey designs meet or exceed established quality levels for survey research (Chang and Krosnick 2009; Baker et al. 2010), a finding corroborated here. The panel data used in this analysis allows me to actually observe and compare measurement error across survey modes, an improvement over previous research. Combined, the results suggest that skepticism towards the accuracy of national-

probability internet surveys is unwarranted. Future research ought to focus on why differences by

survey mode in terms of measurement error and overall survey accuracy exist, and how this may

vary for different respondent populations.

Race Measures
GSS
"Are you Spanish, Hispanic, or latino/latina?" [yes, no]

"What is your race? Indicate one or more races that you consider yourself to be."
Race hand card [White; Black or African American; American Indian or Alaska Native; Asian Indian; Chinese; Filipino; Japanese; Korean; Vietnamese; Other Asian; Native Hawaiian; Guamanian or Chamorro; Samoan; Other Pacific Islander; Some other race]

HCMST/GfK
"This is about Hispanic ethnicity. Are you of Spanish, Hispanic, or Latino descent?" [no, ;yes, Mexican/Mexican American/Chicano; yes, Puerto Rican; yes, Cuban; yes, central American; yes, south American; yes, Caribbean; yes, other Spanish/Hispanic/Latino]

"Please check one or more categories below to indicate what race(s) you consider yourself to be." [White; Black or African American; American Indian or Alaska Native; Asian Indian; Chinese; Filipino; Japanese; Korean; Vietnamese; Other Asian; Native Hawaiian; Guamanian or Chamorro; Samoan; Other Pacific Islander; Some other race]

Gender Measures
GSS
Interviewer selects gender of respondent [male; female]

HCMST/GfK
"Please enter your age on your last birthday and whether you are male or female in the spaces below" [open grid for responses]

Age Measures
GSS
"What is your date of birth?" [month, day, year recorded]

HCMST/GfK
"Please enter your age on your last birthday and whether you are male or female in the spaces below" [open grid for responses]

"Please confirm your date of birth." [open boxes for Month, Day, Year]. If DOB does not match age, then:

"Your date of birth does not match the age you entered earlier. Please enter the correct information here. My correct age is [number box]. I was born in the month of [single choice month]"

Highest degree Measures
GSS
 "Did you ever get a high school diploma or a GED certificate?" [Yes; no]

"Did you ever complete one or more years of college for credit—not including schooling such as business college, technical or vocational school?" [yes; no]

"Do you have any college degrees?"
[yes; no]

"What degree or degrees?" [associate/junior college; bachelor's; graduate]

HCMST/GfK
"What is the highest level of schooling you received?" [no formal education; 1st-4th grade; 5th - 6th grade; 7th - 8th grade; 9th grade; 10th grade; 11th grade; 12th grade no diploma; HS graduate or GED; some college, no degree; associate degree; bachelor's degree; master's degree; professional or doctorate degree]

References

Baker, Reg, Stephen J. Blumberg, et al. 2010. "Research Synthesis: AAPOR Report on Online Panels." *Public Opinion Quarterly* 74:711-781.

Biemer, Paul. 2001. "Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing." *Journal of Official Statistics-Stockholm* 17(2): 295-320.

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. "Measurement error in survey data." *Handbook of Econometrics* 5: 3705-3843.

Chang, Linchiat, and Jon A. Krosnick. 2009. "National Surveys via RDD Telephone Interviewing versus the Internet: Comparing Sample Representativeness and Response Quality.*" Public Opinion Quarterly* 73:661-674.

Chang, Linchiat and Jon A. Krosnick. 2010. "Comparing Oral Interviewing with Self-Administered Computerized Questionnaires: An Experiment." *Public Opinion Quarterly* 74(1): 154-167.

Duffy, Bobby, et al. 2005. "Comparing data from online and face-to-face surveys."*International Journal of Market Research* 47(6): 615.

Groves, Robert M. 1990. "Theories and methods of telephone surveys." *Annual Review of Sociology*: 221-240.

GSS. 2011. "GSS Panel Data Release Notes: July 2011." http://www3.norc.org/GSS+Website/.

GSS. 2012. "Appendix A: Sampling design and Weighting." <http://www3.norc.org/GSS+Website/>.

Heerwegh, Dirk, and Geert Loosveldt. 2008. "Face-to-Face versus Web Surveying in a High-Internet-Coverage Population Differences in Response Quality." *Public Opinion Quarterly* 72(5): 836-846.

Holbrook, Allyson L., Melanie C. Green, and Jon A. Krosnick. 2003."Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias." *Public Opinion Quarterly* 67(1): 79-125.

Krosnick, Jon A. 1999. "Survey research." *Annual review of psychology* 50(1): 537-567.

Malhotra, Neil. 2008. "Completion Time and Response Order Effects in Web Surveys*." Public Opinion Quarterly* 72(5): 914-934.

Paik, Anthony and Kenneth Sanchagrin. 2013. "Social isolation in America: An artifact." *American Sociological Review* 78(3): 339-360.

Rosenfeld, Michael J., Reuben J. Thomas, and Maja Falcon. 2011. How Couples Meet and Stay Together, Waves 1, 2, and 3: Public version 3.04 [Computer file]. Stanford, CA: Stanford University Libraries.

Saperstein, Aliya, and Andrew M. Penner. 2012. "Racial Fluidity and Inequality in the United States." *American Journal of Sociology* 118(3): 676-727.

Schilt, Kristen and Laurel Westbrook. 2009. "Doing gender, doing heteronormativity: 'gender normals,' transgender people, and the social maintenance of heterosexuality." *Gender and Society* 23(4): 440-464.

Smith, Tom W. 2009. "2006-2008 General Social Survey Panel Validation." GSS Methodological Report No. 111.

Smith, Tom W. and Jaesok Son. 2010. "An analysis of panel attrition and panel change on the 2006-2008 General Social Survey Panel." GSS Methodological Report No. 118.

Table 1. Measurement Error in HCMST and GSS Panels

	HCMST	GSS
Age		
Errors	39	47
Valid Responses	2597	1242
Error Rate (%)	1.50	3.78
Z-statistic		4.47***
Gender		
Errors	4	19
Valid Responses	2597	1276
Error Rate (%)	0.15	1.49
Z-statistic		5.08***
Race		
Errors	141	41
Valid Responses	2597	1266
Error Rate (%)	5.43	3.24
Z-statistic		3.02***
Highest Degree		
Errors	152	150
Valid Responses	2597	1275
Error Rate (%)	5.85	11.76
Z-statistic		6.45***

Sources: GfK Core Demographic Profile (from the HCMST) and General Social Survey 2006 3-Wave Panel. ***$p \leq .001$ (two-tailed test).

Note: Valid responses are defined as respondents who answered the question at all three waves.

Table 2. Logistic Regression Models Predicting Measurement Error: HCMST and GSS Panels

	HCMST	GSS
Age Error		
Years of education	-.26***	-.01
Standard Error	[.07]	[.06]
Sample Size	2445	1100
Wald Test [a]		-.25**
Standard Error		[.09]
Gender Error		
Years of education	-.26	-.13*
Standard Error	[.21]	[.07]
Sample Size	2445	1125
Wald Test [a]		-.13
Standard Error		[.22]
Race Error		
Years of education	-.11**	.01
Standard Error	[.04]	[.06]
Sample Size	2445	1117
Wald Test [a]		-.12†
Standard Error		[.07]
Highest Degree Error [b]		
Years of education	-.12**	.09**
Standard Error	[.04]	[.03]
Sample Size	2597	1274
Wald Test [a]		-.21***
Standard Error		[.05]

Notes: Coefficients are log-odds. [a] Wald test for equality of coefficients calculated by estimating a single model with cases from both datasets and testing for an interaction effect between dataset and years of education. [b] Except for highest degree, all models exclude respondents with an error in their highest degree.
Sources: GfK Core Demographic Profile (from the HCMST) and General Social Survey 2006 3-Wave Panel.† $p \leq .10$, *p .05, ** $p \leq .01$, ***p≤ .001 (two-tailed test).