Sociological Methods & Research

Panel Conditioning in the General Social Survey

Andrew Halpern-Manners, John Robert Warren and Florencia Torche

The online version of this article can be found at:

Published by:

$SAGE

Additional services and information for *Sociological Methods & Research* can be found at:

Email Alerts: http://smr.sagepub.com/cgi/alerts

Subscriptions: http://smr.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

>> OnlineFirst Version of Record - May 20, 2014

What is This?

Panel Conditioning in the General Social Survey

Andrew Halpern-Manners[1],
John Robert Warren[2], and Florencia Torche[3]

Abstract

Does participation in one wave of a survey have an effect on respondents' answers to questions in subsequent waves? In this article, we investigate the presence and magnitude of "panel conditioning" effects in one of the most frequently used data sets in the social sciences: the General Social Survey (GSS). Using longitudinal records from the 2006, 2008, and 2010 surveys, we find convincing evidence that at least some GSS items suffer from this form of bias. To rule out the possibility of contamination due to selective attrition and/or unobserved heterogeneity, we strategically exploit a series of between-person comparisons across time-in-survey groups. This methodology, which can be implemented whenever researchers have access to at least three waves of rotating panel data, is described in some detail so as to facilitate future applications in data sets with similar design elements.

Keywords

[1] Indiana University, Bloomington, IN, USA
[2] University of Minnesota, Minneapolis, MN, USA
[3] New York University, New York, NY, USA

Corresponding Author:
Andrew Halpern-Manners, Department of Sociology, Indiana University, 744 Ballantine Hall, 1020 E Kirkwood Avenue, Bloomington, IN 47405, USA.
Email: ahm@indiana.edu

Sociologists have long recognized that longitudinal surveys are uniquely valuable for making causal assertions and for studying change over time. Scholars have also long been aware of the many special challenges that accompany the use of such surveys: they are more expensive to administer, they raise greater data disclosure concerns, and they suffer from additional forms of nonresponse bias. Nevertheless, researchers have generally been content to assume that longitudinal surveys do *not* suffer from the sorts of "testing" or "reactivity" biases that sometimes arise in the context of experimental or intervention-based research. The implicit assumption is that answering questions in one round of a survey in no way alters respondents' reports in later waves. If this assumption is false, scholars risk mischaracterizing the existence, magnitude, and correlates of changes across survey waves in respondents' attitudes and behaviors.

In this article, we investigate the presence and magnitude of "panel conditioning" effects in the General Social Survey (GSS).[1] The GSS is a foundational data resource in the social sciences, surpassed by only the U.S. Census and the Current Population Survey in terms of overall use (Smith 2008). In 2006, the survey made the transition from a replicating cross-sectional design to a design that uses rotating panels. Respondents are now asked to participate in up to three waves of survey interviews, with an identical set of core items appearing in each wave. The core GSS questionnaire touches on a variety of social and political issues, including abortion, intergroup tolerance, crime and punishment, government spending, social mobility, civil liberties, religion, and women's rights (to name just a few). Basic sociodemographic information is also collected from each respondent at the time of their interview and then recollected in subsequent rounds.

Our primary objective is to determine whether panel conditioning influences the overall quality of these data.[2] Along the way, we provide a useful methodological framework that can be used to identify panel conditioning effects in other commonly used data sets. Simply comparing response patterns across individuals who have and have not participated in previous waves of a survey is a good first step, but more sophisticated techniques are needed to convincingly differentiate between panel conditioning and biases introduced by panel attrition (Das, Toepoel, and van Soest 2011; Warren and Halpern-Manners 2012). As we describe in more detail below, our approach (which can be implemented in any longitudinal data set that contains at least three waves of overlapping panel data) resolves this issue by strategically exploiting between-person comparisons across rotation groups.

We believe that this is an important contribution to the emerging literature on panel conditioning effects in social science surveys. Most prior research

on this subject, including our own, has focused on the incidence and magnitude of panel conditioning using a narrow subset of attitudinal or behavioral measures (e.g., employment status *or* life satisfaction). These analyses have tended to use weaker methods to measure panel conditioning effects and have rarely considered the prevalence of the problem across topical domains. In this article, we offer a *general assessment* of panel conditioning in an omnibus survey that is heavily used by social scientists for a wide variety of research purposes. Our results should be valuable to users of the GSS and to researchers who are interested in identifying panel conditioning effects in other data sets that also include an overlapping panel component.

The remainder of this article is organized into four main sections. In the section that follows, we summarize the literature on panel conditioning and provide a theoretical rationale for examining the issue within the context of the GSS. Next, we describe the methodology we use to identify panel conditioning effects. This discussion is meant to be nontechnical, so as to facilitate future applications in data sets with similar design elements. In the third section, we present our main findings and then subject these findings to a falsification test. Finally, we conclude by discussing the implications of our research for scholars who work with the GSS, as well as other sources of longitudinal social science data.

Panel Conditioning and the GSS

When does survey participation change respondents' actual attitudes and behaviors? When does survey participation change merely the quality of their reports about those attitudes and behaviors? Elsewhere, we have developed seven theoretically motivated hypotheses about the circumstances in which panel conditioning effects are most likely to occur (Warren and Halpern-Manners 2012).[3] These hypotheses are grounded in theoretical perspectives on the cognitive processes that underlie attitude formation and change, decision making, and the relationship between attitudes and behaviors (see, e.g., Feldman and Lynch 1988). In short, responding to a survey question is a cognitively and socially complex process that may or may not leave the respondent unchanged and/or equally able to provide accurate information when reinterviewed in subsequent waves. Five of these hypotheses suggest that panel conditioning effects could potentially arise within the context of the GSS.

First, respondents' attributes may at least appear to change across waves when items (like many of those featured on the GSS) require them to provide socially nonnormative or undesirable responses (Torche et al. 2012). The

experience of answering survey questions can force respondents to confront the fact that their attitudes, behaviors, or statuses conflict with what mainstream society regards as normative or appropriate (Schaeffer 2000; Toh, Lee, and Hu 2006; Tourangeau, Rips, and Rasinski 2000).[4] Some respondents may react by bringing their *actual* attitudes or behaviors into closer conformity with social norms. Others may simply avoid cognitive dissonance and the embarrassment associated with offering nonnormative responses by bringing their *answers* into closer conformity with what they perceive as socially desirable.[5] In both cases, the end result would be the same: researchers would observe changes over time in respondents' attributes that would not have occurred had the initial interview not taken place.

Second, respondents' attributes may appear to change across waves as they attempt to manipulate the survey instrument in order to minimize their burden (see, e.g., Bailar 1989). Respondents sometimes find surveys to be tedious, cognitively demanding, and/or undesirably lengthy (Krosnick 1991; Krosnick et al. 2002; Tourangeau et al. 2000). To get around these hassles, respondents in longitudinal studies may learn how to direct or manipulate the survey experience in such a way that minimizes the overall amount of time or energy that they have to devote to it (Duan et al. 2007; Wang, Cantor, and Safir 2000).[6] In the GSS, for example, a respondent may learn during their first interview that they are asked to provide many additional details about their job characteristics and work life. In order to reduce the duration of follow-up surveys, some respondents may subsequently report that they are out of the labor force or unemployed.[7] The result would be the appearance of change across waves when no change has actually occurred.

Third, as hypothesized by Waterton and Lievesley (1989:324), it is possible that some respondents change their answers to survey questions as they gain an "improved understanding of the rules that govern the interview process." When first interviewed, participants in the GSS may not have had full access to the information requested from them, may not have known how to make use of various response options, or may not have known how or when to ask clarifying questions. Upon reinterview, these individuals may be better prepared and more cognizant of "how surveys work." While this may translate into undesirable manipulation of the survey instrument, as posited previously, it may also lead to more accurate and complete responses over time. This would again result in the appearance of change over time when respondents' underlying attributes remain entirely the same (see, e.g., Mathiowetz and Lair 1994; Sturgis, Allum, and Brunton-Smith 2009).

Fourth, respondents may become more comfortable with and trusting of the survey experience after being exposed to the survey process and interviewers

(van der Zouwen and van Tilburg 2001). Survey methodologists have found that respondents' judgments about the relative benefits and risks associated with answering survey questions are significantly related to the chances that they provide complete and accurate answers (Dillman 2000; Krumpar 2013; Rasinski et al. 1999; Willis, Sirken, and Nathan 1994). As respondents become more familiar with and trusting of the survey process and with interviewers and interviewing organizations, they may become less suspicious and their confidence in the confidentiality of their responses may grow. Participating in the GSS may provide evidence about the survey's harmless nature, reduce suspicion, or increase respondents' comfort level. Any of these effects could lead to changes in respondents' reported attitudes or behaviors.

Finally, respondents' answers to factual questions may change over time as they acquire more and better information about the topic at hand (Toepoel, Das, and van Soest 2009). After an initial interview, respondents may ''follow-up'' on unfamiliar items by consulting external sources and/or people who are knowledgeable in the area. In this scenario, prior questions serve as stimuli for obtaining the type of information that is needed to give correct responses in later waves. In many cases, it may not even be necessary that respondents remember that they encountered the item during a previous interview. As Cantor (2008:136) points out, all that matters is that ''the process of answering the question the first time changes what is eventually accessible in memory the next time the question is asked.'' The GSS includes a number of ''knowledge tests'' that may be especially prone to this form of panel conditioning.

Unfortunately, these hypotheses have not been well validated using the sorts of data sets social scientists typically rely on. One consequence of this is that we know very little about the nature and magnitude of panel conditioning in important data resources like the GSS.[8] Whereas most large-scale surveys provide users with methodological documentation about issues like sampling, attrition, and missing data, we know of none that routinely provides information about panel conditioning based on strong methods for understanding such biases. In the short run, we hope that our empirical estimates of panel conditioning in the GSS will improve the scholarship that is based on analyses of these data. In the longer run, we intend for our research design to serve as a methodological model for assessing panel conditioning in surveys like the GSS that employ rotating panel designs.[9]

Data and Research Design

The GSS is a large, full-probability survey of noninstitutionalized adults in the United States. It has been administered annually (1972–1993) or

biennially (1994 onward) since 1972 by NORC at the University of Chicago. In 2006, the GSS switched from a cross-sectional design to a rotating panel format. Under the new setup, subsets of about 2,000 respondents are randomly selected in each wave for reinterview two and four years later. The longitudinal panel that began the GSS in 2006 was reinterviewed in 2008 and 2010; the panel that began in 2008 was reinterviewed in 2010 and 2012. As described below, our focus is on responses to the 2008 survey by two groups of individuals: those who were interviewed for the first time in 2006 (or cohort A) and those who were interviewed for the first time in 2008 (or cohort B).

At first glance, it might seem that the easiest way to identify panel conditioning effects in these data would be to compare the responses given by individuals who were new to the survey in 2008 (cohort B) to those given by individuals who first participated in 2006 (cohort A). The problem with this approach is its inability to distinguish the effects of panel *conditioning* from the effects of panel *attrition*. Whereas the new rotation group may be representative of the original target population (i.e., noninstitutionalized adults living in the United States at the time of the 2008 survey), the 2006 cohort may have suffered from nonrandom attrition between the 2006 and 2008 waves. Unless credible steps are taken to adjust for the resulting panel selectivity, differences in responses between cohorts cannot be clearly attributed to panel conditioning (Halpern-Manners and Warren 2012).

Various methodologies have been proposed to deal with this issue (see, e.g., Das et al. 2011; Warren and Halpern-Manners 2012). One of the most common involves the use poststratification weights (Clinton 2001; Nukulkij et al. 2007). Under this approach, attrition is assumed to be random conditional on a predetermined set of observable characteristics, which are then used to generate weights that correct for discrepancies between different cohorts of respondents. As others have pointed out, the overall effectiveness of this technique depends entirely on whether or not assumptions concerning "ignorability" are met (Das et al. 2011; Sturgis et al. 2009; Warren and Halpern-Manners 2012). If the two cohorts under consideration (i.e., the 2006 and 2008 cohorts) differ in ways that are not easily captured by the variables used to construct the weights, contamination due to panel attrition cannot be ruled out.

One way around this problem is to "preselect" individuals who have the same underlying propensity to persist in the sample. Consider, for example, cohorts A and B as defined previously. These groups of respondents began the GSS in 2006 and 2008, respectively. If we systematically select individuals from both cohorts who participated in *at least the first two waves of*

survey interviews, and then examine their responses in 2008, we can accurately identify the effects of panel conditioning in that year. Both sets of respondents experienced the same social and economic conditions at the time of their interview in 2008, and both exhibited the same propensity to persist in (or attrite from) the GSS panel (because both participated in the same number of waves).[10] The key difference between the groups is that members of the 2006 cohort were experienced GSS respondents in 2008 and members of 2008 cohort were not.[11]

This is the approach that we use in our analysis. Using panel data from the 2006, 2008, and 2010 waves of the GSS, we were able to identify 3,117 respondents who completed at least the first two rounds of survey interviews.[12] Of these respondents, 1,536 entered the sample in 2006 (the 2006 cohort) and 1,581 entered the sample in 2008 (the 2008 cohort). If the responses given by individuals in the first group are significantly different than the responses given (in the same year) by individuals in the second, we can infer that these differences came about from panel conditioning. No adjustments for panel attrition are necessary and person weights are not needed to correct for subsampling and/or nonresponse.[13] By design, the 2006 and 2008 cohorts have already been equated on both observed *and* unobserved characteristics.[14]

As we noted previously, items on the GSS span a wide variety of substantive topics (Smith et al. 2007). Although theory suggests that some of these topics may be more or less prone to panel conditioning effects, we feel it is important (for the sake of completeness) to examine every instance in which such biases could possibly occur. For this reason, we considered *all* 2008 GSS variables that met two very basic requirements: (1) the item had to be answered by the respondent and not the survey interviewer and (2) the variable in question had to be empirically distinct from other measures in our analysis. The first rule meant that items like "date of interview" and "sex of interviewer" were excluded from the study. The second rule meant that we considered variables like "age" and "year of birth," but not both.[15]

After eliminating items that did not satisfy these criteria, we were left with a total of 310 variables. To analyze panel conditioning effects in each of these measures, we carried out hypothesis tests comparing the response patterns in 2008 across cohorts. For continuous measures, we used *t*-tests to compare group means; for categorical measures, we used chi-square tests (if all cell sizes were in excess of 5) and Fisher's exact tests (if they were not).[16] Because the GSS employs a split-ballot design, where certain items are only asked of certain individuals in a given year, there is no guarantee that experienced respondents had prior exposure to all of the variables in our
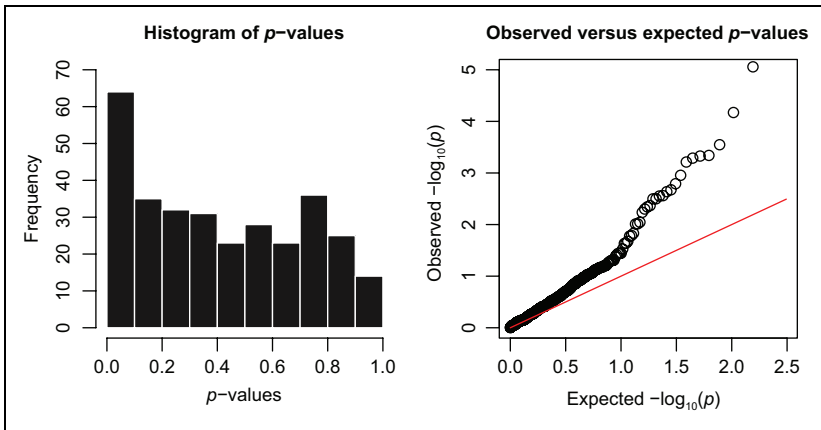
sample.[17] Such cases were removed from the analysis using pairwise deletion. See online Appendix Table A1 for complete information on all measures, including sample sizes disaggregated by cohort.

Results

Our analysis includes significance tests for 310 different items; this makes it extremely susceptible to multiple comparison problems. Even if the null hypothesis (of no panel conditioning) is true for every item in our data set, the probability of finding at least one statistically significant effect just by chance is $1 - (1 - 0.05)^{310} \approx 1$, assuming a standard α level of .05. To address this issue, we examined the *distribution* of test statistics across all items in our sample. Under the null, the p values obtained from our tests should be uniformly distributed between 0 and 1 (Casella and Berger 2001). Approximately 5 percent of the test statistics should be below 0.05, another 5 percent should fall between 0.05 and 0.09, and so on throughout the entire [0, 1] interval. Depending on where they occur in the distribution, departures from this pattern could indicate an overabundance of significant results.

Figure 1 gives a visual summary of the main findings. In the panel on the left, we provide a simple histogram of the p values we obtained from our comparisons of the 2006 and 2008 cohorts. In the panel on the right, we provide a quantile–quantile (Q-Q) plot comparing the empirical distribution of these values (as indicated by the black circles) to a theoretical null distribution (as indicated by the diagonal line).[18] In both instances, there is clear clustering of estimates in the extreme low end of the distribution.[19] Overall, 63 of the 310 tests that we conducted were significant at a .10 level (whereas 31 would be expected by chance); 37 were significant at a .05 level (whereas 16 would be expected by chance); and 22 were significant at a .01 level (whereas 3 would be expected by chance). We take this as good evidence that panel conditioning exists in the GSS among certain subsets of items.

In order to confirm this interpretation, we calculated p values that have been adjusted for the false discovery rate (FDR) using the algorithm of Benjamini and Hochberg (1995). Many techniques exist for dealing with multiple comparison problems and there is some debate over which is the most appropriate (Gelman, Hill, and Yajmia 2012). The FDR is generally thought to be more powerful than Bonferroni-style procedures and is frequently used when the volume of tests is high. Instead of controlling for the chances of making even a single type 1 error, the FDR controls for the expected *proportion* of type 1 errors among all significant results. In total, the FDR-adjusted estimates include eight significant results at the $p < .05$

Histogram of *p*–values

Observed versus expected *p*–values

Figure 1. Histogram and Q-Q plot of observed *p* values. The panel on the left shows the observed distribution of *p* values for all items in our sample (*n* = 310). Under the null, the values should be uniformly distributed between 0 and 1. The panel on the right compares the observed distribution to a theoretical (null) distribution. If *p* values are *more* significant than expected, points will move up and away from the diagonal line. If *p* values are uniformly distributed, the circles will track closely with the line throughout the entire range. See text for further details.

level and 19 significant results at the $p < .10$ level (see online Appendix Table A1). If we set the FDR threshold to 5 percent, we can say with confidence that only one of these "discoveries" occurred by chance.

The Direction and Magnitude of Panel Conditioning Effects

These results suggest that some people may respond differently to GSS questions based on whether or not they have previously participated in the survey. Although this is an important finding in its own right, users of these data should also be interested in knowing which variables are subject to panel conditioning, in what direction the observed effects operate, and how big they are from a substantive standpoint. In this section, we describe the direction and magnitude of panel conditioning biases in the 2008 survey and provide some preliminary thoughts about possible mechanisms. To be appropriately conservative when interpreting the results for individual variables, we focus on items that (with a few exceptions) produced FDR-adjusted *p* values $< .10$. The exceptions to this rule are noted in the text below.

First, members of the 2006 and 2008 cohorts sometimes differed in their responses to attitudinal questions about "hot-button" issues. Examples include items dealing with premarital sex, first amendment rights and racism, and governmental aid to minorities. As indicated in Table 1, members of the 2006 cohort were 14 percent more likely to say that sex before marriage is always or almost always wrong, 10 percent more likely to say that people have a right to make hateful speeches in public, and 23 percent more likely to say that current levels of assistance for African Americans are neither too high nor too low. These effect sizes are generally in line with estimates that have been produced in past panel conditioning research (see, e.g., Torche et al. 2012).

Second, panel conditioning effects emerged in several questions related to household composition. These include items dealing with the respondent's relationship to the household head (members of the 2006 cohort were more likely to be the head or spouse), the number of adults present (members of the 2006 cohort reported more adults), the number of visitors present (members of the 2006 cohort reported more visitors), and the number of family generations that live with the respondent (members of the 2006 cohort reported more generations). The fact that experienced GSS respondents reported higher numbers in all of these cases may be related to our hypothesis concerning survey skill and/or trust. After completing the survey for the first time, respondents may have become more willing to open up, to report on more people, or to ask follow-up questions about who qualifies as living in their household.[20]

Third, members of the 2006 and 2008 cohorts frequently differed in their responses to questions about demographic and economic attributes. Respondents with prior survey experience were 20 percent more likely to be divorced or widowed, 11 percent more likely to be upwardly mobile relative to their parents, and 31 percent less likely to refuse to answer questions about their personal income. Although we cannot provide definitive tests, these patterns could also be attributable to differences in respondents' trust. As we discussed earlier, being interviewed repeatedly may make the interview process seem less threatening to the respondent, which could decrease their need to give guarded and/or socially desirable responses in the follow-up wave (van der Zouwen and van Tilburg 2001). That this would occur for potentially sensitive items like those listed above makes good theoretical sense.[21]

Finally, we found large and consistent differences between groups with respect to their knowledge about science. Although these differences were typically not below the FDR-adjusted $p < .10$ threshold, the frequency with which they occurred is at the very least suggestive of a "true" effect. As

Table 1. Differences Between Cohorts in 2008, Illustrative Results.

Variable	Description of Response Options/Measure	Estimate (Percentage or Mean) 2006 Cohort	2008 Cohort
phone	Respondent refuses to give information about their phone	1.17	7.93
visitors	Average number of visitors in the household	0.05	0.01
parsol	Respondent's standard of living is higher than their parents' standard of living	66.21	59.50
rplace	The respondent is the householder or their spouse	91.70	88.21
adults	Average number of adults in the household	1.97	1.87
natracey	Respondent thinks current levels of public assistance for blacks are about right	53.51	43.60
marital	Respondent is divorced or widowed	25.88	21.56
spkrac	Respondent agrees that people have a right to make hateful speeches in public	67.08	60.81
rincom06	Respondent refuses to report income	4.27	6.05
famgen	Reports that there is only one generation in household	53.26	57.12
premarsx	Respondent reports that sex before marriage is always or almost always wrong	34.94	30.75
radioact	Correctly answers question about the source of radioactivity	84.79	79.40
viruses	Correctly answers question about efficacy of antibiotics	65.64	59.35
condrift	Correctly answers question about plate tectonics	91.34	87.21
electron	Correctly answers question about sizes of electrons/atoms	75.77	70.45

Note: The 2006 cohort is restricted to respondents who were interviewed in 2006 and 2008; the 2008 cohort is restricted to respondents who entered the panel in 2008 and were also interviewed in 2010. Comparisons *between* cohorts are made in 2008, the year that they overlap in the sample. All of the variables presented in this table produced false discovery rate (FDR)-adjusted *p* values below .10, except for the science knowledge items. We included these items because of the consistency across measures (all four were significant by conventional standards and all four effects were in the same, theoretically sensible, direction). See text for more details, and online Appendix Table A1 for the full set of results.

shown in Table 1, respondents in the treatment group were markedly more likely to answer correctly questions about the source of radioactivity, the efficacy of antibiotics in killing viruses, the ongoing process of plate tectonics (*condrift*), and the relative sizes of electrons and atoms. One possible explanation for these results is the "learning hypothesis" that we proposed

earlier: if respondents who previously participated in the GSS seek out information about questions that have one objectively correct answer, we would expect to see differences between cohorts on precisely these sorts of items.

A Note on Exceptions

Although the empirical patterns that we present in Table 1 are generally consistent with theoretical expectations, there are also plenty of counterexamples where the two cohorts of respondents did not differ in predictable or meaningful ways. We did not *always* find differences between cohorts when examining questions about socially charged issues, nor did we observe significant effects for *all* items that required factual knowledge or increased levels of respondent trust (for the complete set of results, see online Appendix Table A1). These interitem inconsistencies do not invalidate our findings, but they do suggest the need for more finely grained analyses that are capable of isolating and carefully testing the various hypotheses that we outlined out earlier. We will return to this idea later on in the Discussion section.

Falsification Test

In the final part of our analysis, we carry out a falsification test to confirm the adequacy of our empirical approach. As a part of its mission to provide up-to-date information about a wide variety of topics, the GSS frequently introduces new survey content through the use of special topical modules. This allows us to perform an important methodological check. Using the same analytic setup as before, we can test for differences between cohorts on items that have *not* previously been answered by anyone in the sample, *regardless of which cohort they belong to*. In the absence of any contaminating influences, we would expect to see a similar distribution of responses across groups for these measures. Any other result (e.g., nonzero differences between the two cohorts on items that should not, in theory, differ) would call into question the internal validity of our empirical estimates.

We present results from these comparisons in Table 2. In total, there are 19 variables that (1) were *not* asked in 2006, (2) *were* asked of both cohorts in 2008, and (3) meet the selection criteria that we defined earlier. Among these items, only one (*autonojb*) shows any evidence of variation between cohorts, and that evidence disappears when corrections are made for multiple comparisons.[22] None of the estimated tests are significant at a .01 level and only two reach significance at the .10 level (with 19 comparisons, we would expect to see ∼1 significant result by chance, assuming a type 1 error rate

Table 2. Results From Falsification Tests.

Variable Description	Name	Tests for Differences Between Cohorts	
		p	FDR-adjusted p
Trying to start a business	startbiz	.50	.75
Number of full-time jobs since 2005	work3yrs	.67	.78
Number of years worked for current employer	curempyr	.54	.75
Amount of pay change since started job	paychnge	.40	.75
Was pay higher/lower/the same in previous job?	pastpay	.28	.68
Why did the respondent leave their previous job?	whyleave	.35	.75
Does more trade lead to fewer jobs in the United States?	moretrde	.27	.68
Computer use at work	wkcomptr	.12	.66
Can job be done without a computer?	wocomptr	.82	.82
Have any coworkers been replaced by computers?	autonojb	.02	.22
Frequency of meetings with customers, clients, or patients	meetf2f1	.15	.66
Frequency of meetings with coworkers	meetf2f2	.28	.68
Frequency of communication with coworkers outside the United States	intlcowk	.22	.68
Does the respondent receive health insurance from their employer?	emphlth	.82	.82
Is there another name for the respondent's insurance or HMO policy?	othplan	.58	.75
Gender of sex partners	sexsex18	.09	.66
Ever been the target of sexual advances by a coworker/supervisor?	harsexjb	.55	.75
Has respondent been the target of a sexual advance by a religious leader?	harsexcl	.66	.78
Do they know others who have been the target of sexual advances?	knwclsex	.47	.75

Note: HMO, Health Maintenance Organization; These items were not asked of the 2006 cohort in 2006, but were asked of both cohorts in 2008. Variable names are given in the "name" column. The FDR-adjusted p is the p value adjusted for the false discovery rate. Adjustments were made using the procedures of Benjamini and Hochberg (1995). See text for more details.

of 0.05).[23] This is a reassuring finding for our purposes, as it minimizes the possibility that the two cohorts differ in ways that could spuriously produce some or all of what we previously deemed to be panel conditioning effects.

Discussion

Sociologists who work with longitudinal data typically assume that the changes they observe across waves are real and would have occurred even in the absence of the survey. Whether or not this assumption is justified is an important empirical question, one that should be of concern to methodologists and nonmethodologists alike. In this article, we provided an analytic framework for detecting "panel conditioning effects" in longitudinal surveys that include a rotating panel component. To demonstrate the utility of our approach, we analyzed data from recent waves of the GSS. Results from these analyses suggest that panel conditioning influences the quality of a small but nontrivial subset of core survey items. This inference was robust to a falsification test and cannot be explained by statistical artifacts stemming from panel attrition and/or differential nonresponse.

What should applied researchers make of these findings? Our analysis suggests that panel conditioning exists in the GSS on a broad scale, but it is much less clear about the specific content domains that are most affected by this form of bias. As we mentioned at the outset, panel conditioning is a complex interactive phenomenon that involves a range of cognitive processes and subjective individual assessments. Predicting when and where it will occur is a difficult theoretical exercise. We have attempted to provide some guidance to users of the GSS by listing the variables that show the most evidence of possible effects. We would advise researchers to weigh this information carefully when conducting studies with these data. Although panel conditioning does not always present itself in an intuitive or internally consistent manner, it would be wrong to dismiss it as an unimportant methodological issue.

There is obviously still work to be done in this area. The analytic techniques described herein can be usefully applied in any longitudinal data set that contains overlapping panels. An interesting future application would be to examine heterogeneity in panel conditioning among different subgroups of respondents. In our analysis, we sought to identify the *average treatment effect* taken over all members of the sample. In reality, these effects may vary considerably across individuals, across social contexts, and across topical domains (see, e.g., Zwane et al. 2011). A treatment effect of zero in the population may nevertheless be nonzero for certain subgroups with particular experiences and/or predispositions. Identifying who these individuals are, and how they differ from others, would go a long way toward refining our theoretical understanding of why panel conditioning occurs.

Another worthwhile extension would be to conduct stand-alone experiments that allow for a closer examination of possible mechanisms. These

experiments would not need to be complicated; it would probably be enough to assign individuals at random to receive alternate forms of a baseline questionnaire and then to ask all questions of all individuals in a follow-up. To speak to the issue in a way that is broadly useful to sociologists, the questions would need to be similar or identical to those that routinely appear in other widely used surveys, like the GSS, *and* would need to be carefully selected in order to isolate the various social and psychological processes that we described earlier. This would obviously require considerable effort and careful planning, but we believe it is the best way to produce a general and theoretically informed understanding of panel conditioning in longitudinal social science research.

Supplementary Material

The online data supplements are available at http://smr.sagepub.com/supplemental.

Notes

1. We use the term "panel conditioning" synonymously with what has been called, among other things, "time-in-survey effects" (Corder and Horvitz 1989), "mere measurement effects" (Godin et al. 2008), "question-behavior effects" (Spangenberg, Greenwald, and Sprott 2008), and "self-erasing errors of prediction" (Sherman 1980).
2. Panel conditioning cannot affect respondents the first time that they are interviewed, so analyses that use only the replicating cross sections are not at risk.
3. Similar hypotheses can be found in reviews by Cantor (2008), Sturgis et al. (2009), and Waterton and Lievesley (1989).
4. Examples from the GSS include questions that deal with respondents' racial attitudes, their history of substance use, their sexuality, their past criminal behavior, and their fidelity to their spouse or partner.

5. It is important to distinguish these sorts of changes from social desirability bias. In some cases, the *mere thought* of providing a nonnormative answer may cause respondents to alter the way that they characterize themselves on a baseline survey and in all subsequent interviews (Tourangeau and Yan 2007). In other cases, the *experience* of admitting to something that is socially undesirable may change the way respondents describe themselves in later waves—because of the feelings of embarrassment or shame that the initial interview provoked. Although both of these things could be happening at the same time within the same survey, our focus in this article is only on the latter problem. For more information about the former problem, the interested reader should see Schaeffer (2000) and Tourangeau and Yan (2007).

6. This question-answering strategy can be thought of as a strong form of *satisficing.* Not only are respondents seeking to provide ''merely satisfactory answers'' (Krosnick 1991), they are also deliberately seeking to avoid additional follow-up questions.

7. This sort of "burden avoidance" behavior can also occur within the context of a cross-sectional survey if respondents learn, through repetition, that certain types of answers lead to additional items (see, e.g., Kessler et al. 1998; Kreuter et al. 2011). We thank an anonymous reviewer for pointing this out.

8. We know of two previous analyses that have examined panel conditioning effects in the GSS (Smith and Son 2010; Warren and Halpern-Manners 2012). Both focused on a fairly narrow subset of survey items ($n < 25$), and neither ruled out alternative explanations for the observed results (including selective attrition, random measurement error, and social desirability bias). The present article represents an improvement on both fronts.

9. Other widely used, nationally representative surveys that employ a rotating panel design include the Current Population Survey and the Survey of Income and Program Participation. Panel conditioning effects have been assessed in both of these surveys (Bailar 1975; Halpern-Manners and Warren 2012; McCormick, Butler, and Singh 1992; Solon 1986), but only for a very select subset of items.

10. Both sets of respondents are probably also subject to similar levels of nonresponse bias, although this is not something that we can verify using available data.

11. The age distribution of respondents will vary slightly between groups because members of the 2006 cohort have aged two years since their initial interview (and thus cannot be 18 or 19 years old), whereas the members of the 2008 cohort have not. In supplementary analyses, we truncated the age distribution so that *all* respondents were above the age of 20 in 2008 and then recalculated our estimates. The results were substantively identical and are available upon request.

12. We only use the 2010 data for the purposes of sample selection; we do not actually analyze respondents' answers from that wave of the survey.

13. Random sampling in two different years (e.g., 2006 and 2008) does not guarantee the same population characteristics when the composition of the population changes gradually over time. To confirm that the differences we attribute to panel conditioning are not due to slight compositional changes that occurred between 2006 and 2008, we fit a series of auxiliary models that included controls for various sociodemographic characteristics (i.e., age, gender, householder status, and race/ethnicity). Our conclusions with respect to panel conditioning were robust to the inclusion of these variables.

14. This approach would provide invalid results if there is an important attrition-by-cohort interaction. Even if members of the 2006 and 2008 cohorts were equally likely to leave the sample, it may still be the case that attriters from these cohorts differ with respect to socioeconomic, demographic, or other attributes that might predict responses to the survey items we consider. To explore this possibility, we pooled our data files and ran a regression model predicting attrition. For independent variables, we included indicators of the respondent's age, gender, socioeconomic status, race/ethnicity, region of residence, marital status, party affiliation, household size, happiness, and health. We then created interactions between these measures and the respondent's cohort. None of these interactions were significant at the $p < .05$ level. This provides reassurance that the process generating attrition was similar across groups.

15. A third stipulation is that the variables under consideration had to appear on the 2006 *and* 2008 waves of the survey. For the most part, this limits our analysis to items that belong to the GSS's replicating core.

16. In very rare instances ($n = 7$), results for a Fisher's exact test could not be obtained for computational reasons. In these cases, we consolidated response categories to reduce data sparseness and then carried out chi-square tests instead.

17. Core GSS items appeared in the same order for both cohorts of respondents in 2008.

18. Q-Q plots are widely used in genetics research to visualize results from large numbers of hypothesis tests (see, e.g., Pearson and Manolio 2008). To draw the plot, we rank ordered the p values ($n = 1, \ldots, 310$) from smallest to largest and then graphed them against the values that would have been expected had they been sampled from a uniform distribution. As noted previously, the diagonal line indicates the expectation under the null and the black circles represent the actual results. Following convention, we show the relevant test statistics as the $-\log_{10}$ of the p value, so that an observed $p = .01$ is plotted as "2" on the y-axis and $p = 10^{-5}$ as "5."

19. The null hypothesis that the observed values are uniformly distributed was easily rejected using a Kolmogorov–Smirnov test ($D = 0.18$, $p < .0001$).

20. These variables are not good candidates for "burden" effects because respondents receive very few additional questions for each household member that they report.
21. We also found that members of the 2006 cohort were much more likely to give out information about their home phone. This is, again, consistent with a "trust" effect.
22. We excluded three employment-related variables (*ownbiz*, *findnwjb*, and *losejb12*) from these analyses because they closely resemble items that appeared on the 2006 survey. One of these variables produced a significant difference between cohorts; the other two did not.
23. None of the comparisons were significant after adjusting the *p* values for the false discovery rate (FDR), and a Kolmogorov–Smirnov test could not reject the null hypothesis that the distribution of results was uniform ($D = .2105$, $p = .81$).

References

Bailar, Barbara A. 1975. "Effects of Rotation Group Bias on Estimates from Panel Surveys." *Journal of the American Statistical Association* 70:23-30.

Bailar, Barbara A. 1989. "Information Needs, Surveys, and Measurement Errors." Pp. 1-24 in *Panel Surveys*, edited by D. Kasprzyk, G. J. Duncan, G. Kalton, and M. P. Singh. New York: John Wiley.

Benjamini, Yoav and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society, Series B (Methodological)* 57:289-300.

Cantor, David. 2008. "A Review and Summary of Studies on Panel Conditioning." Pp. 123-38 in *Handbook of Longitudinal Research: Design, Measurement, and Analysis*, edited by S. Menard. Burlington, MA: Academic Press.

Casella, George and Roger L. Berger. 2001. *Statistical Inference*. New York: Duxbury Press.

Clinton, Joshua D. 2001. "Panel Bias from Attrition and Conditioning: A Case Study of the Knowledge Networks Panel." Unpublished manuscript. Department of Political Science, Stanford University. Retrieved April 5, 2006 (http://www.princeton.edu/~clinton/WorkingPapers/C_WP2001.pdf).

Corder, Larry S. and Daniel G. Horvitz. 1989. "Panel Effects in the National Medical Care Utilization and Expenditure Survey." Pp. 304-18 in *Panel Surveys*, edited by D. Kasprzyk, G. J. Duncan, G. Kalton, and M. P. Singh. New York: John Wiley.

Das, Marcel, Vera Toepoel, and Arthur van Soest. 2011. "Nonparametric Tests of Panel Conditioning and Attrition Bias in Panel Surveys." *Sociological Methods & Research* 40:32-56.

Dillman, Don A. 2000. *Mail and Internet Surveys: The Tailored Design Method*. New York: John Wiley.

Duan, Naiuhua, Margarita Alegria, Glorisa Canino, Thomas G. McGuire, and David Takeuchi. 2007. ''Survey Conditioning in Self-reported Mental Health Service Use: Randomized Comparison of Alternative Instrument Formats.'' *Health Services Research* 42:890-907.

Feldman, Jack M. and John G. Lynch. 1988. "Self-generated Validity and Other Effects of Measurement on Belief, Attitude, Intention, and Behavior." *Journal of Applied Psychology* 73:421-35.

Gelman, Andrew, Jennifer Hill, and Masanao Yajmia. 2012. "Why We (Usually) Don't Have to Worry about Multiple Comparisons." *Journal of Research on Educational Effectiveness* 5:189-211.

Godin, Gaston, Paschal Sheeran, Mark Conner, and Marc Germain. 2008. "Asking Questions Changes Behavior: Mere Measurement Effects on Frequency of Blood Donation." *Health Psychology* 27:179-84.

Halpern-Manners, Andrew and John Robert Warren. 2012. "Panel Conditioning in the Current Population Survey: Implications for Labor Force Statistics." *Demography* 49:1499-519.

Kessler, Ronald C., Hans-Ulrich Wittchen, Jamie A. Abelson, Katharine McGonagle, Norbert Schwarz, Kenneth S. Kendler, Bärbel Knäuper, and Shanyang Zhao. 1998. "Methodological Studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey (NCS)." *International Journal of Methods in Psychiatric Research* 7:33-55.

Kreuter, Frauke, Susan McCulloch, Stanley Presser, and Roger Tourangeau. 2011. "The Effects of Asking Filter Questions in Interleafed Versus Grouped Format." *Sociological Methods & Research* 40:88-104.

Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5:213-36.

Krosnick, Jon A., Allyson L. Holbrook, Matthew K. Berent, Richard T. Carson, W. Michael Hanemann, Raymond J. Kopp, Robert Cameron Mitchell, Stanley Presser, Paul A. Ruud, V. Kerry Smith, Wendy R. Moody, Melanie C. Green, and Michael Conaway. 2002. "The Impact of 'No Opinion' Response Options on Data Quality: Non-attitude Reduction or an Invitation to Satisfice?" *Public Opinion Quarterly* 66:371-403.

Krumpar, Ivar. 2013. "Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review." *Quality & Quantity* 47:2025-47.

Mathiowetz, Nancy A. and Tamra J. Lair. 1994. "Getting Better? Changes or Errors in the Measurement of Functional Limitations." *Journal of Economic & Social Measurement* 20:237-62.

McCormick, Maryann K., Derrick M. Butler, and Rajendra P. Singh. 1992. "Investigating Time in Sample Effect for the Survey of Income and Program Participation.

" Pp. 554-59 in *Proceedings of the Survey Research Methods Section of the American Statistical Association*.

Nukulkij, Poom, Jadfield Hadfield, Subias Subias, and Evan Lewis. 2007. "An Investigation of Panel Conditioning with Attitudes toward U.S. Foreign Policy." Retrieved August 25, 2013 (http://www.knowledgenetworks.com/ganp/docs/jsm2009/Panel%20Conditioning%20and%20Attrition_JSM_2009_submitted.pdf).

Pearson, Thomas A. and Teri A. Manolio. 2008. "How to Interpret a Genome-wide Association Study." *JAMA* 299:1775-877.

Rasinski, Kenneth A., Gordon B. Willis, Alison K. Baldwin, Wenchi Yeh, and Lisa Lee. 1999. "Methods of Data Collection, Perceptions of Risks and Losses, and Motivation to Give Truthful Answers to Sensitive Survey Questions." *Applied Cognitive Psychology* 13:465-84.

Schaeffer, Nora Cate. 2000. "Asking Questions about Threatening Topics: A Selective Overview." Pp. 105-22 in *The Science of Self-report: Implications for Research and Practice*, edited by A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, and V. Cain. Hillsdale, NJ: Lawrence Erlbaum.

Sherman, Steven J. 1980. "On the Self-erasing Nature of Errors of Prediction." *Journal of Personality and Social Psychology* 39:211-21.

Smith, Tom W. 2008. "Repeated Cross-sectional Research: The General Social Surveys." Pp. 33-48 in *Handbook of Longitudinal Research: Design, Measurement, and Analysis*, edited by S. Menard. London, England: Elsevier.

Smith, Tom W., Jibum Kim, Achim Koch, and Alison Park. 2007. *The GSS Model of Social-science Research*. GSS Project Report No. 27. Chicago, IL: National Opinion Research Center, University of Chicago.

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. Chicago, IL: National Opinion Research Center, University of Chicago.

Solon, Gary. 1986. "Effects of Rotation Group Bias on Estimation of Unemployment." *Journal of Business & Economic Statistics* 4:105-09.

Spangenberg, Eric R., Anthony G. Greenwald, and David E. Sprott. 2008. "Will You Read This Article's Abstract? Theories of the Question-behavior Effect." *Journal of Consumer Psychology* 18:102-06.

Sturgis, Patrick, Nick Allum, and Ian Brunton-Smith. 2009. "Attitudes over Time: The Psychology of Panel Conditioning." Pp. 113-26 in *Methodology of Longitudinal Surveys*, edited by P. Lynn. New York: John Wiley.

Toepoel, Vera, Marcel Das, and Arthur van Soest. 2009. "Relating Question Type to Panel Conditioning: Comparing Trained and Fresh Respondents." *Survey Research Methods* 3:73-80.

Toh, Rex S., Eunkyu Lee, and Michael Y. Hu. 2006. "Social Desirability Bias in Diary Panels is Evident in Panelists' Behavioral Frequency." *Psychological Reports* 99:322-34.

Torche, Florencia, John Robert Warren, Andrew Halpern-Manners, and Eduardo Valenzuela. 2012. "Panel Conditioning in a Longitudinal Study of Chilean Adolescents' Substance Use: Evidence from an Experiment." *Social Forces* 90:891-918.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. New York: Cambridge University Press.

Tourangeau, Roger and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133:859-83.

van der Zouwen, Johannes and Theo van Tilburg. 2001. "Reactivity in Panel Studies and its Consequences for Testing Causal Hypotheses." *Sociological Methods & Research* 30:35-56.

Wang, Kevin, David Cantor, and Adam Safir. 2000. "Panel Conditioning in a Random Digit Dial Survey." Pp. 822-27 in *Proceedings of the Survey Research Methods Section of the American Statistical Association*.

Warren, John Robert and Andrew Halpern-Manners. 2012. "Panel Conditioning Effects in Longitudinal Social Science Surveys." *Sociological Methods & Research* 41:491-534.

Waterton, Jennifer and Denise Lievesley. 1989. "Evidence of Conditioning Effects in the British Social Attitudes Panel Survey." Pp. 319-39 in *Panel Surveys*, edited by D. Kasprzyk, G. J. Duncan, G. Kalton, and M. P. Singh. New York: John Wiley.

Willis, Gordon B., Monreau G. Sirken, and Gad Nathan. 1994. "The Cognitive Aspects of Responses to Sensitive Survey Questions." Working Paper Series, No. 9, National Center for Health Statistics, Hyattsville, MD.

Zwane, Alix Peterson, Jonathan Zinman, Eric Van Dusen, William Pariente, Clair Null, Edward Miguel, Michael Kremer, Dean S. Karlan, Richard Hornbeck, Xavier Giné, Esther Duflo, Florencia Devoto, Bruno Crepon, and Abhijit Banerjee. 2011. "Being Surveyed Can Change Later Behavior and Related Parameter Estimates." *Proceedings of the National Academy of Sciences* 108:1821-26.

Author Biographies

Andrew Halpern-Manners is an assistant professor in the Department of Sociology at Indiana University. His research interests include social demography, the sociology of education, aging and the life course, and survey methods. His recent publications have appeared in Demography, Social Forces, and Sociological Methods & Research.

John Robert Warren is Professor of Sociology at the University of Minnesota and Training Director of the Minnesota Population Center. He is co-leading efforts to harmonize and link all available basic and supplemental Current Population Survey

records; to re-interview in 2013-14 the 1980 High School and to link records from the Health and Retirement Survey, the Panel Study of Income Dynamics, the Wisconsin Longitudinal Study, and other aging surveys to the 1940 U.S. Census. He is also the current Editor of Sociology of Education and a member of the National Research Council's Board on Testing and Assessment.

Florencia Torche is associate professor of sociology at New York University. Her research interests include stratification and mobility, sociology of education and the long-term impact of early exposures over the individual life course.