Panel Conditioning in Longitudinal Social Science Surveys

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Andrew Halpern-Manners

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

John Robert Warren, Adviser

July 2013

ACKNOWLEDGEMENTS

This dissertation took a while to write. It began in a very different form while I was living in Minnesota; it traveled with me, cross-country, to Texas; and then it morphed, after a few false starts, into what it is now. Singling out everyone who helped me along the way would require a feat of memory that I am unfortunately incapable of, but some people left such a mark that even I cannot forget them. Before getting down to business, I would like to pause here to acknowledge these people's contributions and to offer them my gratitude for all of their help and support over the years.

From a professional standpoint, no one deserves more thanks than my mentor and friend, Rob Warren. Rob convinced me to come to Minnesota for graduate school, and then proceeded to teach me *everything I know about doing good social science research* (he also taught me a thing or two about coaching a championship-caliber Little League team). As my adviser, he has read drafts of every paper that I have ever written, always offering thoughtful comments no matter how half-baked my ideas were. Rob, thank you so much for everything you have done; words cannot express how incredibly grateful I am.

Thanks are also due to the other three members of my committee. Eric Grodsky, Phyllis Moen, and Michael Davern each braved preliminary drafts of this dissertation and—in a testament to their strength and fortitude as scholars—survived to give me useful and timely feedback. Eric, in particular, deserves thanks for reminding me to not lose track of what matters most: theory, substance, and a strong argument. I can say without a hesitation that I am a better (and more versatile) sociologist for having worked with him.

Other people who have contributed to my work include Carolyn Liebler (a great quilter and an even better friend), Chris Uggen, Scott Eliason, Penny Edgell, Teresa Swartz, Deborah Levison, Jennifer Lee (I never did win that paper award…), Wes Longhofer (master mixologist), Sara Wakefield, Reiping Huan, Liying Liu, Florencia Torche, Jim Raymo, Bill James, Steve Ruggles, Matt Sobek, Monty Hindman, Chandra Muller, and Eve Pattison. I consider myself tremendously lucky to have spent my graduate school years surrounded by such a gifted and giving group of scholars.

Last but not least, I must acknowledge my family and friends. The ins-and-outs of longitudinal data are *not* something that translates well to polite conversation, but the people who are closest to me never batted an eye. To Dan Frey, Ben Narvaez, J. Scott Svien, Cameron McLaughlin, Paul Pender, Reed Fischer, and Matt Jordan—the next round is on me. To my mom, my dad, and my brother Nick—you guys mean the world to me; it's a true blessing to have you as my family. And finally, to my beautiful, loving, and *incredibly* patient wife, Elaine—it's done, can you believe it? We should go celebrate!

To my parents

ABSTRACT

Researchers who utilize data from longitudinal surveys nearly always assume that respondents' attributes are not changed as a result of being measured. Yet research in cognitive psychology, political science, and elsewhere suggests that the experience of being interviewed can spark important changes in the way respondents behave, in the attitudes that they possess, and in their willingness or ability to answer questions accurately when they are re-interviewed in subsequent waves. In this dissertation, I evaluate the severity of this problem in longitudinal social science surveys. Using a combination of observational and experimental data, I show that "panel conditioning" has the potential to affect a wide range of attitudinal and behavioral measures, including many items that are commonly used in sociological and demographic research. The causal mechanisms that give rise to these effects are discussed and a large-scale follow-up project is proposed.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS I

ABSTRACT IV

LIST OF TABLES VII

LIST OF FIGURES VIII

CHAPTER 1: INTRODUCTION 1

CHAPTER 2: PANEL CONDITIONING IN THE GENERAL SOCIAL SURVEY 4

- Background 6
- Data and research design 10
- Results 15
  - The direction and magnitude of panel conditioning effects 18
  - A note on exceptions 22
  - Falsification test 22
- Discussion 25
- Appendix 28

CHAPTER 3: ASSESSING PANEL CONDITIONING USING EXPERIMENTS 48

- Background 49
- Three hypotheses concerning the nature of panel conditioning 52
- The present study 56
  - Study design 57
  - Estimation strategy 63
- Results 64
  - Robustness check 68
  - Heterogenous treatment effects 68

Discussion 74
Appendix 78
CHAPTER 4: A PROPOSAL FOR FUTURE RESEARCH 81
Specific aims 81
Significance 84
Prior research on panel conditioning and behaviors 84
Prior research on panel conditioning and attitudes 86
Critique of existing panel conditioning research 87
Theoretical background and hypotheses 90
Summary 100
Innovation 101
Approach 102
Basic research design 102
Addressing my specific aims 107
Timeline to completion and dissemination plan 114
REFERENCES 116

LIST OF TABLES

Table 2.1: Size and direction of estimated effects, illustrative results 20

Table 2.2: Results from falsification tests 24

Table A1: Linear probability models predicting attrition 29

Table A2: Sample sizes by variable and cohort 30

Table A3: Test statistics for all survey items 39

Table 3.1: Descriptive statistics, by survey ballot at baseline 59

Table 3.2: Linear probability models predicting deviant behavior, by follow-up group 65

Table 3.3: Interactions between treatment condition and survey spacing 67

Table 3.4: Robustness check using “less seasoned” respondents 69

Table B1: Models predicting the probability of attrition between interview occasions 80

LIST OF FIGURES

Figure 2.1: A three-wave identification strategy using overlapping panels in the GSS............13

Figure 2.2: Histogram and Q-Q plot of observed *p*-values 17

Figure A1: *P*-values using identical age distributions...28

Figure 3.1: Identifying panel conditioning effects using an experimental design......................60

Figure 3.2: Treatment effect interactions with age..71

Figure 3.3: Treatment effect interactions with income...72

Figure 3.4: Treatment effect interactions with gender..73

CHAPTER 1: INTRODUCTION

When analyzing longitudinal data, social scientists typically assume that the experience of being interviewed had no effect on the participants in the study. We take it as a given that respondents' attitudes, behaviors, and statuses were unchanged by the questions that they were asked and by the ideas and feelings that those questions provoked. What if this assumption is false? What if the experience of being interviewed changes people—or the way that they describe themselves—in subtle but important ways? In this dissertation, I seek to engage these questions using strong methods and a theoretically-grounded approach. My basic research question is this: does participating in one wave of a longitudinal survey in any way alter the kinds of answers that respondents provide when we re-interview them in subsequent waves? In other words, is there "panel conditioning" in the sorts of longitudinal data sets that sociologists and others frequently rely on?

To answer this question, I have conducted two pieces of original research. In the first piece of research, I analyze data from one of the most frequently used surveys in the social sciences: the General Social Survey (GSS). In 2006, the GSS made the transition from a replicating cross-sectional survey to a survey that uses a rotating panel design. Respondents are now asked to participate in up to three waves of survey interviews, with interviews occurring at equally spaced intervals every two years. To determine whether this change has had any effect on the quality of the data, I divided panel members into two carefully selected groups (i.e., respondents who first participated in the survey in 2006 versus respondents who first participated in in 2008) and then made a series of strategic comparisons across groups within waves. Results from these analyses, which I report in Chapter 2, suggest that panel conditioning has the potential to affect a wide array of attitudinal and behavioral measures—including many that are commonly used in sociological research.

What are the processes driving these results? In the second piece of research, I examine some of the social and psychological mechanisms that are thought to give rise to panel conditioning. Drawing on theories from cognitive psychology, survey methodology, and elsewhere, I present a series of hypotheses concerning the nature of panel conditioning and the circumstances that are most likely to produce it. In order to evaluate these hypotheses, I fielded a small-scale experiment using respondents from a well-known internet panel. As a part of the experiment, respondents were randomly assigned to one of several treatment groups, with each treatment group receiving a slightly different set of questions during their baseline interview. Respondents were then re-surveyed one month or one year later using the full battery of items. Comparisons across groups suggest that panel conditioning tends to be most pronounced when surveys are spaced closely together in time, and when the questions that are asked deal with past events that may be difficult to recall "on the fly."

These studies offer new insight into the prevalence, causes, and magnitude of panel conditioning effects in the social sciences—*but they are really only the beginning of a longer-term research agenda*. In the concluding chapter of this dissertation, I describe a larger and more ambitious data collection effort that would allow researchers to *definitively test* theoretically-motivated hypotheses concerning the nature and scope of panel conditioning in longitudinal social science surveys. Building on the analyses described above, I show how a specially-tailored survey instrument could be used to (1) pinpoint specific substantive areas where panel conditioning effects will and will not occur; and (2) test for heterogeneity in panel conditioning effects across different subgroups in the population. I argue that this type of work is necessary if we are to develop an appropriate methodological response to panel conditioning in future sociological research.

CHAPTER 2: PANEL CONDITIONING IN THE GENERAL SOCIAL SURVEY

Sociologists have long recognized that longitudinal surveys are uniquely valuable for making causal assertions and for studying change over time. Sociologists 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 non-response bias (see, e.g., Lazarsfeld 1940). 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 within the context of experimental, ethnographic, and intervention-based research. The basic 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 waves in respondents' attitudes and behaviors.

In this chapter, I 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

[1] In this and other chapters, I 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), and "question-behavior effects" (Spangenberg, Greenwald, and Sprott 2008).

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 socio-demographic information is also collected from each respondent at the time of their interview and then re-collected in subsequent rounds.

My primary objective is to determine whether panel conditioning influences the overall quality of these data. Along the way, I 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 useful first step, but more sophisticated techniques are needed to convincingly differentiate between panel conditioning and biases introduced by panel attrition (Das, Toepoel, and Soest 2011; Holt 1989; Sturgis, Allum, and Brunton-Smith 2009; Warren and Halpern-Manners 2012). As I describe in more detail below, my approach resolves this issue by strategically exploiting between-person comparisons across rotation groups. This methodology can be easily implemented in any longitudinal data set that contains at least three waves of rotating panel data.

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

Background

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, I have developed seven theoretically-motivated hypotheses about the circumstances in which panel conditioning effects are most likely to occur (Warren and Halpern-Manners 2012).[2] 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 re-interviewed in subsequent waves. Five of these hypotheses suggest that panel conditioning may 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 non-normative or

[2] Similar hypotheses can be found in reviews by Cantor (2008), Sturgis et al. (2009), and Waterton and Lievesley (1989).

undesirable responses (Torche et al. 2012). Survey questions can force respondents to confront the reality that their attitudes, behaviors, or statuses conflict with what their peers regard as normative or appropriate (Fitzsimons and Moore 2008; Toh, Lee, and Hu 2006).[3] 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 non-normative responses by bringing their *answers* into closer conformity with what they perceive as socially desirable. In both cases, panel conditioning would be responsible for changes in respondents' subsequent reports.

Second, respondents' attributes may appear to change across survey waves as they attempt to manipulate a 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. In order to minimize the amount of time and energy that they have to devote to answering questions, some respondents may seek to direct or manipulate the survey experience (Wang, Cantor, and Safir 2000). For example, in the GSS, a respondent may learn during their first interview that they are asked to provide many details about their job characteristics and work life. In an attempt to reduce the duration of follow-up surveys, some respondents may subsequently report that they are out of the labor force or unemployed. The result would be the appearance of change when no change has actually occurred.

[3] Relevant 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.

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 re-interview, 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 above, it may also lead to more accurate and complete responses over time (and thus panel conditioning) (see, e.g., Mathiowetz and Lair 1994).

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; Tourangeau, Rips, and Rasinski 2000). 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 reported attitudes or behaviors across waves.

Finally, respondents' answers to factual questions may change over time as they acquire more and better information about the topic at hand. 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.[4] Whereas most large-scale surveys provide users with methodological documentation about issues like sampling, non-response, panel attrition, and missing data, I know of none that routinely provides information about panel conditioning based on strong methods for

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

understanding such biases. In the short run, I hope that my empirical estimates of panel conditioning in the GSS will improve the scholarship that is based on analyses of these data. In the longer run, I intend for my research design to serve as a methodological model for assessing panel conditioning in surveys like the GSS that employ rotating panel designs.

Data and research design

The GSS is a large, full-probability survey of non-institutionalized adults in the United States. It has been administered annually (1972-1993) or biennially (1994 onward) since 1972 by the National Opinion Research Center 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 re-interview two and four years later. The longitudinal panel that began the GSS in 2006 was re-interviewed in 2008 and 2010; the panel that began in 2008 was re-interviewed in 2010 and 2012. As described below, my 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 (the "control group" or Cohort B) to those given by individuals who first participated in 2006 (the "treatment group" or 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 target population (i.e., non-institutionalized adults living in the United States at the time of the 2008 survey), the "treatment group" may have suffered from non-random 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 (Das et al. 2011; Warren and Halpern-Manners 2012). One of the most common involves the use post-stratification weights (Clinton 2001; Nukulkij et al. 2007). Under this approach, attrition is assumed to be random conditional on a pre-determined set of observable characteristics, which are then used to generate weights that correct for discrepancies between the treatment and control group. 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 treatment and control 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 "pre-select" individuals that have the same underlying propensity to persist in the sample (as shown in Figure 2.1). Consider, for example, Cohorts A and B as defined above. These groups of respondents began the GSS in 2006 and 2008, respectively. If I 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, I can accurately identify the effects of panel conditioning in that year. Both sets of respondents were sampled from the same population using the same procedures, both experienced the same social and economic conditions at the time of their interview, and both exhibited the same propensity to persist in (or attrite from) the panel. The only difference between the groups is that members of Cohort A were experienced GSS respondents in 2008 and members of Cohort B were not.[5]

This is the approach that I use in my analysis. Using panel data from the 2006, 2008, and 2010 waves of the GSS, I was able to identify 3,117 respondents who completed at least the first two rounds of survey interviews. Of these respondents, 1,536 entered the sample in 2006 (the treatment group, or Cohort A) and 1,581 entered the sample in 2008 (the control group, or Cohort B). 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, I can conclude with confidence that these differences came about from panel conditioning. No additional adjustments for panel attrition are necessary and person weights are not needed to correct for sub-sampling and/or non-response. By design, the treatment and control groups have already

[5] The age distribution of respondents will vary slightly between cohorts because the treatment group has aged two years since their initial interview (and thus cannot be 18 or 19 years old), whereas the control group has not. In supplementary analyses, I truncated the age distribution so that *all* respondents were above the age of 20 in 2008 and then recalculated my estimates. The results were substantively identical and can be found in Appendix Figure A1.

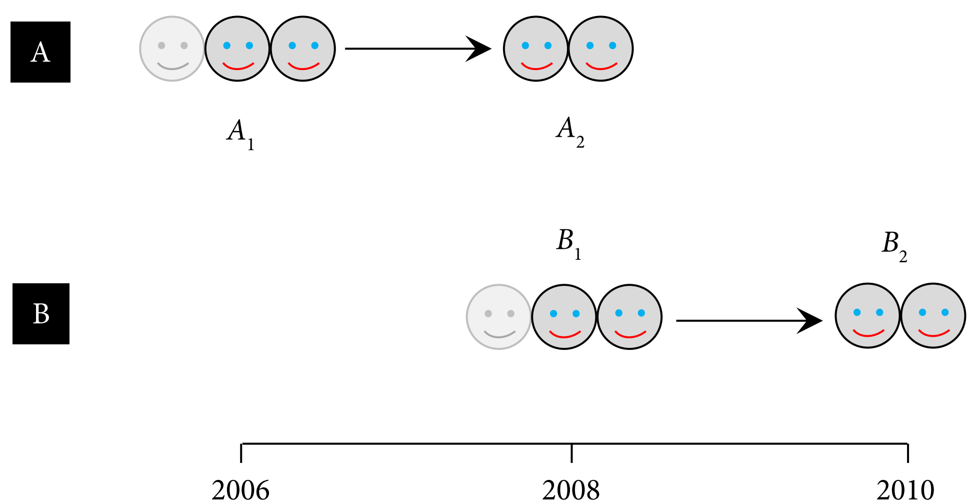

Figure 2.1: A three-wave identification strategy using overlapping panels in the GSS. In this figure, there are two cohorts of respondents: the cohort that entered the sample in 2006 (Cohort A) and the cohort that entered the sample in 2008 (Cohort B). In order to identify panel conditioning effects, I make comparisons between cohorts *in 2008*. In particular, I compare the responses given by members of Cohort A (A_2) to the responses given by members of Cohort B (B_1), *provided that members of Cohort B also participated in the 2010 wave*. This approach ensures that both groups have the same propensity to persist in the study (because they were both in-sample for the same amount of time), and thus eliminates concerns surrounding panel attrition. In the example given above, transparent faces signify respondents who left the sample after their initial interview.

been equated on both observed *and* unobserved characteristics.[6]

As noted above, 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, I feel it is important (for the sake of completeness) to examine every instance in which such biases could possibly occur. For this reason, I 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 conceptually and empirically distinct from other measures in my 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 I considered questions like "usual number of hours worked in a week" or "hours worked last week," but not both.[7]

After eliminating items that did not satisfy these criteria, I was left with a total of 297 variables. To analyze panel conditioning effects in each of these measures, I carried out hypothesis tests comparing the response patterns in 2008 for members of each cohort. For

---

[6] 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 that I consider. To explore this possibility, I pooled my data files and ran a regression model predicting attrition. For independent variables, I included indicators of the respondent's age, gender, socioeconomic status, race/ethnicity, region of residence, marital status, party affiliation, household size, happiness, and health. I then created interactions between these measures and the respondent's cohort. None of these interactions were significant at the $p < 0.05$ level (see Appendix Table A1). This provides reassurance that the process generating attrition was similar across groups.

[7] 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 my analysis to items that belong to the GSS's replicating core.

continuous measures I used *t*-tests to compare group means; for categorical measures I used chi-square tests (if all cell sizes were in excess of 5) and Fisher's exact tests (if they were not).[8] Because the GSS employs a split-ballot design, where certain items are only asked of certain individuals in a given year, members of the treatment group did not necessarily receive the "treatment" for all variables in my sample. Such cases were removed from the analysis using pairwise deletion. See Appendix Table A2 for complete information on all measures, including sample sizes disaggregated by treatment status.

Results

My analysis includes significance tests for 297 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 my data set, the probability of finding at least one statistically significant effect just by chance is $1 - (1 - 0.05)^{297} \approx 1$, assuming a standard α-level of 0.05. To address this issue, I examined the *distribution* of test statistics across all items in my sample. Under the null, the *p*-values obtained from my tests should be uniformly distributed between 0 and 1 (Casella and Berger 2001). Approximately 5% of the test statistics should be below 0.05, another 5% 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

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

this pattern could indicate an over-abundance of significant results.

Figure 2.2 gives a visual summary of the main findings. In the panel on the left, I provide a simple histogram of the *p*-values I obtained from my comparisons of treatment versus control. In the panel on the right, I 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 red line).[9] In both instances, there is clear clustering of estimates in the extreme low end of the distribution.[10] Overall, 61 of the 297 tests that I conducted were significant at a 0.10 level (whereas 30 would be expected by chance); 35 were significant at a 0.05 level (whereas 15 would be expected by chance); and 21 were significant at a 0.01 level (whereas 3 would be expected by chance). I take this as good evidence that panel conditioning exists in the GSS among certain subsets of items.

In order to confirm this interpretation, I 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 active debate over which is the most appropriate (Gelman, Hill, and Yajmia 2012). The FDR is

[9] 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, I rank-ordered the *p*-values ($n = 1, \dots, 297$) 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 above, the red line indicates the expectation under the null and the black circles represent the actual results. Following convention, I 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."

[10] The null hypothesis that the observed values are uniformly distributed was easily rejected using a Kolmogorov-Smirnov test ($D = 0.1796$, $p < .0001$).

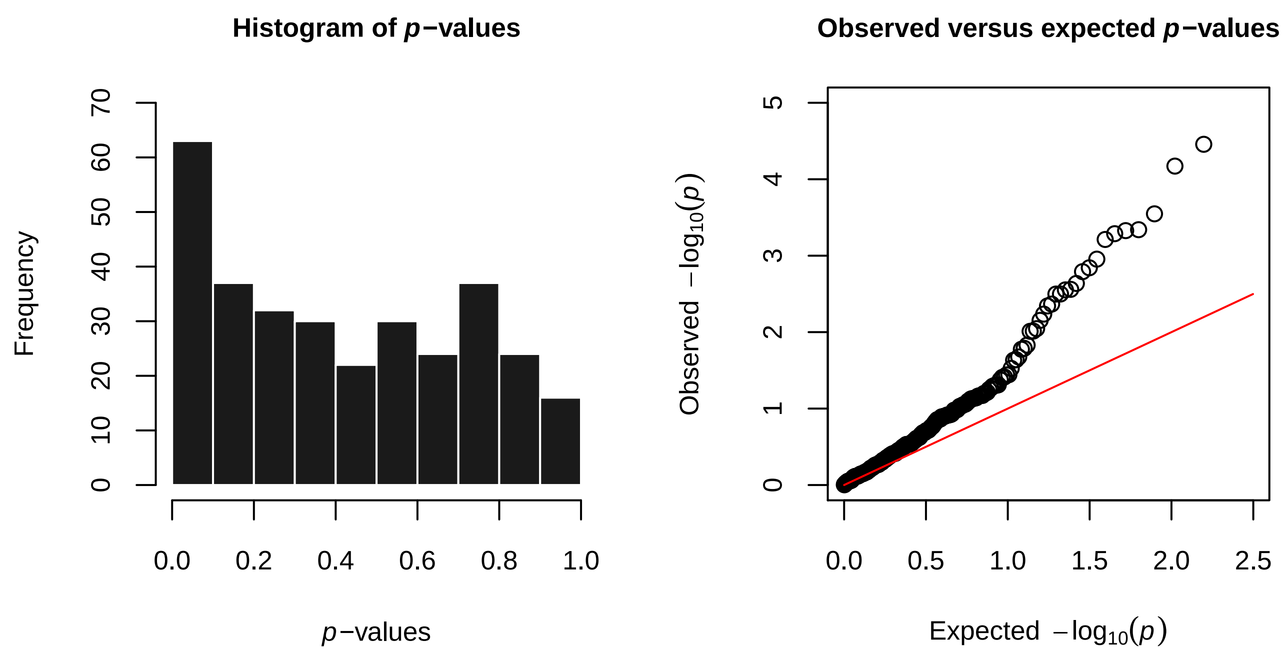

Figure 2.2: Histogram and Q-Q plot of observed p-values. The panel on the left shows the observed distribution of p-values for all items in my sample ($n = 297$). Under the null, the values should be uniformly distributed between 0 and 1. The panel on the right compares the observed distribution to a theoretical (null) distribution. If p-values are *more* significant than expected, points will move up and away from the red line. If p-values are uniformly distributed, the circles will track closely with the red line throughout the entire [0,1] interval. See text for further details.

generally thought to be more powerful than Bonferroni-style procedures, and is frequently used when the volume of tests is high (Benjamini, Krieger, and Yekutieli 2006). 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 10 significant results at the $p < 0.05$ level and 18 significant results at the $p < 0.10$ level (see Appendix Table A3). If I set the FDR threshold to 5%, I can say with confidence that only [18 * .05 = 0.9] ~1 of these "discoveries" occurred by chance.

The direction and magnitude of panel conditioning effects

The results above provide compelling evidence that people respond differently to many 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, I 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, my discussion focuses on variables that (with a few exceptions) produced FDR-adjusted p-values < 0.10. Throughout, I refer to Cohort A (which entered the sample in 2006) as the treatment group and Cohort B (which entered the sample in 2008) as the control group.

First, members of the treatment and control groups sometimes differed in their responses to attitudinal questions about "hot-button" issues. Examples include items dealing with premarital sex (*premarsx*), first amendment rights and racism (*spkrac*), and governmental aid to minorities (*natracey*). While this may dismay researchers who use the GSS for one of its signature purposes (i.e., to study public opinion about important social and political topics), I should note that in most cases the size of these differences was not overwhelmingly large. As indicated in Table 2.1, members of the treatment group were 14% more likely to say that sex before marriage is always or almost always wrong; 10% more likely to say that people have a right to make hateful speeches in public; and 23% more likely to say that current levels of assistance for African Americans are neither too high nor too low.

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 (treated respondents were more likely to be the head or spouse), the number of adults present (treated respondents reported more adults), the number of visitors present (treated respondents reported more visitors), and the number of family generations that live with the respondent (treated respondents reported more generations). The fact that members of the treatment group reported higher numbers in all of these cases may be related to my hypothesis concerning survey skill and trust. After taking the survey for the first time, respondents may become more willing to open up, to report more people, or to ask follow-up questions about who qualifies as living in their household and who does not.[11]

[11] These variables are not good candidates for "burden" effects because respondents receive very few additional questions for each household member that they report.

Table 2.1 Size and direction of estimated effects, illustrative results

		Estimate (% or mean)		Tests for differences	
Variable	Description of response options/measure	Treatment	Control	p	Adjusted p
phone	Respondent refuses to give information about their phone	1.17	7.93	0.00	0.00
visitors	Average number of visitors in the household	0.05	0.01	0.00	0.00
parsol	Respondent's standard of living is higher than their parents' standard of living	66.21	59.50	0.00	0.02
rplace	The respondent is the householder or their spouse	91.70	88.21	0.00	0.02
adults	Average number of adults in the household	1.97	1.87	0.00	0.02
natracey	Respondent thinks current levels of public assistance for blacks are about right	53.51	43.60	0.00	0.02
lngthinv	Average length of the interview in minutes	109.70	114.95	0.00	0.02
marital	Respondent is divorced or widowed	25.88	21.56	0.00	0.04
spkrac	Respondent agrees that people have a right to make hateful speeches in public	67.08	60.81	0.00	0.06
rincom06	Respondent refuses to report income	4.27	6.05	0.00	0.08
famgen	Reports that there is only one generation in household	53.26	57.12	0.01	0.09
premarsx	Respondent reports that sex before marriage is always or almost always wrong	34.94	30.75	0.01	0.10
radioact	Correctly answers question about the source of radioactivity	84.79	79.40	0.02	0.21
viruses	Correctly answers question about efficacy of antibiotics	65.64	59.35	0.02	0.25
condrift	Correctly answers question about plate tectonics	91.34	87.21	0.02	0.26
electron	Correctly answers question about sizes of electrons/atoms	75.77	70.45	0.05	0.42

Note : The adjusted p has been corrected for the False Discovery Rate using the procedures of Benjamini and Hochberg (1995).

Third, members of the treatment and control groups frequently differed in their responses to questions about demographic and economic attributes. Respondents in the treatment group were more likely to be divorced or widowed (25.9% versus 21.6%), more likely to be upwardly mobile relative to their parents (66.2% versus 59.5%), and less likely to refuse to answer questions about their personal income (4.3% versus 6.1%). Although I cannot provide definitive tests, these patterns could also be attributable to differences in trust. As I 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 (van der Zouwen and van Tilburg 2001). That this would occur for potentially sensitive items like those listed above makes good logical sense.[12]

Finally, I found large and consistent differences between groups with respect to their knowledge about science. Although these differences were typically not below the $p < 0.01$ threshold, the frequency with which they occurred is at the very least suggestive of a "true" effect. As shown in Table 1, respondents in the treatment group were markedly more likely to answer correctly questions about the source of radioactivity (*radioact*), the efficacy of antibiotics in killing viruses (*viruses*), the ongoing process of plate tectonics (*condrift*), and the relative sizes of electrons and atoms (*electron*). One possible explanation for these results is the "learning hypothesis" that I proposed earlier: if respondents to the GSS seek out

[12] I also found that respondents in the treatment group were more likely to give out information about their home phone. This is, again, consistent with a "trust" effect.

information about questions that have one objectively correct answer, one would expect to see differences between treatment groups on precisely these sorts of items.

A note on exceptions

Although the empirical patterns that I present in Table 2.1 are generally consistent with theoretical expectations, there are also plenty of counter-examples where the treatment and control groups did not differ in predictable or meaningful ways. I did not *always* find differences between cohorts when examining questions about socially charged issues, nor did I observe significant effects for *all* items that required factual knowledge or increased levels of respondent trust (for the complete set of results, see Appendix Table A3).[13] These inter-item inconsistencies do not invalidate my findings, but they do suggest the need for more finely-grained analyses that are capable of isolating and carefully testing the various hypotheses that I laid out earlier. I will return to this idea in subsequent chapters.

Falsification test

In the final part of my analysis, I carry out a simple falsification test to confirm the adequacy of my 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

---

[13] Beyond these inconsistencies, I found very little evidence to support my hypothesis concerning respondent burden. One possible exception was a variable (*lngthinv*) indicating how long the interview lasted. According to that measure, respondents in the treatment group took 4.7% less time to complete the survey. This amounts to a difference of about 5 minutes.

special topical modules. This allows me to perform an important methodological check. Using the same analytic setup as before, I can test for differences between cohorts on items that have *not* previously been answered by anyone in the sample, *regardless of treatment status*. In the absence of any contaminating influences, I would expect to see a similar distribution of responses across groups for these measures. Any other result (e.g., non-zero differences between the treatment and control groups on items that should not, in theory, differ) would call into question the internal validity of my empirical estimates.

I present results from these comparisons in Table 2.2. In total, there are 19 variables that (1) were *not* asked of the treatment group in 2006; (2) *were* asked of both treatment and control in 2008; and (3) meet the selection criteria that I 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.[14] None of the estimated tests are significant at a 0.01 level, and only two reach significance at the 0.10 level (with 19 comparisons I would expect to see ~1 significant result by chance, assuming a Type 1 error rate of 0.05).[15] This is a reassuring finding for my purposes, as it minimizes the possibility that the treatment and control groups differ in ways that could spuriously produce some or

[14] I 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 significant differences; the other two did not.

[15] None of the comparisons were significant after adjusting the p-values for the FDR, and a Kolmogorov-Smirnov test did not reject the null hypothesis that the distribution of results was uniform ($D = .2105$, $p = 0.81$).

Table 2.2 Results from falsification tests

		Tests for differences between cohorts	
Variable description	Name	*p*	FDR-adjusted *p*
Trying to start a business	startbiz	0.50	0.75
Number of full-time jobs since 2005	work3yrs	0.67	0.78
Number of years worked for current employer	curempyr	0.54	0.75
Amount of pay change since started job	paychnge	0.40	0.75
Was pay higher/lower/the same in previous job?	pastpay	0.28	0.68
Why did the respondent leave their previous job?	whyleave	0.35	0.75
Does more trade lead to fewer jobs in the U.S.?	moretrde	0.27	0.68
Computer use at work	wkcomptr	0.12	0.66
Can job be done without a computer?	wocomptr	0.82	0.82
Have any co-workers been replaced by computers?	autonojb	0.02	0.22
Frequency of meetings with customers, clients, or patients	meetf2f1	0.15	0.66
Frequency of meetings with co-workers	meetf2f2	0.28	0.68
Frequency of communication with co-workers outside the U.S.	intlcowk	0.22	0.68
Does the respondent receive health insurance from their employer?	emphlth	0.82	0.82
Is there another name for the respondent's insurance or HMO policy?	othplan	0.58	0.75
Gender of sex partners	sexsex18	0.09	0.66
Ever been the target of sexual advances by a co-worker/supervisor?	harsexjb	0.55	0.75
Has respondent been the target of a sexual advance by a religious leader?	harsexcl	0.66	0.78
Do they know others who have been the target of sexual advances?	knwclsex	0.47	0.75

Note: These items were not asked of the treatment or control group in 2006, but were asked of both groups in 2008. The FDR-adjusted *p* -value is the p-value adjusted for the False Discovery Rate. See text for more information.

what I 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 methodologists and non-methodologists alike. In this chapter, I provided an analytic framework for detecting panel conditioning effects in longitudinal surveys that include a rotating panel component. To demonstrate the utility of my approach, I analyzed data from the 2006, 2008, and 2010 waves of the GSS. Results from these analyses suggest that panel conditioning influences the quality of a small but non-trivial subset of survey items. This inference was robust to a falsification test, and cannot be explained by statistical artifacts stemming from panel attrition and/or differential non-response.

What should applied researchers make of these findings? My analysis shows quite convincingly 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 I 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. I have attempted to provide some guidance to users of the GSS by listing the variables that show the most evidence of possible effects. I 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 a mistake to dismiss it as an unimportant methodological problem.

There is obviously much more work still to be done in this area. The analytic techniques that I described above 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 sub-groups of respondents. In my analysis, I sought to identify the *average treatment effect* taken over all members of the sample. In reality, these effects may vary considerably across individuals, social contexts, and topical domains (see, e.g., Zwane et al. 2011). A treatment effect of zero in the population may nevertheless be non-zero in certain groups 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 methodological 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 survey. 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 data sets, like the GSS, and would need to be carefully selected in order to isolate the various social and psychological processes that I outlined earlier. This would obviously require considerable effort and careful planning, but I believe it is the best way to produce a general and theoretically-informed understanding of panel conditioning in longitudinal social science research.

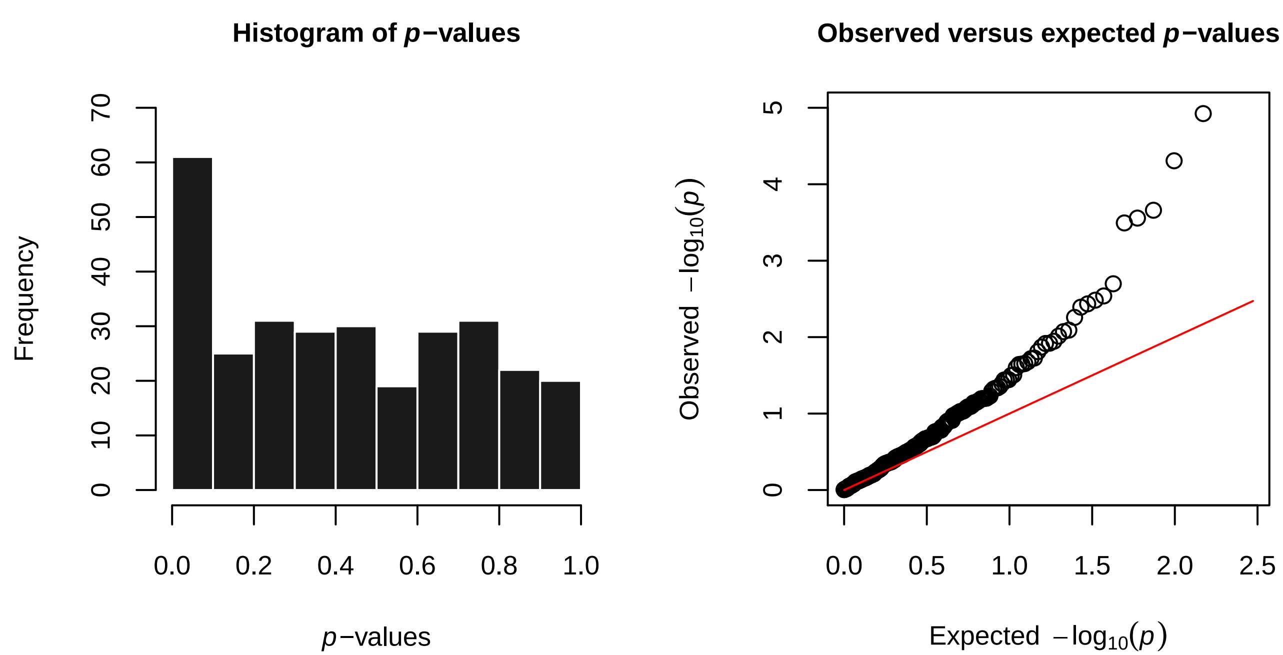

Figure A1: *P*-values using identical age distributions. The panel on the left shows the observed distribution of *p*-values for all items in my sample ($n = 297$). Under the null, the values should be uniformly distributed between 0 and 1. The panel on the right compares the observed distribution to a theoretical (null) distribution. If *p*-values are *more* significant than expected, points will move up and away from the red line. If *p*-values are uniformly distributed, the circles will track closely with the red line throughout the entire [0,1] interval.

Table A1. Linear probability models predicting attrition

	β	(SE)
Main effect		
Cohort (2006 = reference category)	0.08	(0.14)
Interactions with cohort		
Age	0.00	(0.00)
Female	0.02	(0.03)
SEI	0.00	(0.00)
Black (White = reference category)	0.03	(0.05)
Other race	0.05	(0.06)
Mid-Atlantic (New England = reference category)	-0.04	(0.09)
East-North Central	0.06	(0.09)
West-North Central	0.06	(0.11)
South Atlantic	0.02	(0.09)
East-South Central	0.13	(0.11)
West-South Central	0.12	(0.10)
Mountain	-0.06	(0.10)
Pacific	0.11	(0.09)
Widowed (Married = reference category)	0.00	(0.07)
Divorced	0.02	(0.05)
Separated	0.13	(0.10)
Never married	0.04	(0.05)
Not strong democrat (Strong Democrat = reference category)	0.04	(0.06)
Independent, near Democrat	0.02	(0.06)
Independent, near Democrat	-0.03	(0.06)
Independent, near Republican	-0.05	(0.07)
Not strong Republican	0.01	(0.06)
Strong Republican	-0.07	(0.06)
Other party affiliation	-0.02	(0.12)
Size of household	-0.02	(0.01)
Pretty happy (Very happy = reference category)	0.02	(0.04)
Not too happy	0.06	(0.06)
Good health (Excellent health = reference category)	0.03	(0.04)
Fair health	-0.09 *	(0.05)
Poor health	0.09	(0.08)

Note: Main effects were estimated for each of the covariates listed above.
* $p < .10$; ** $p < .05$; *** $p < .01$

Table A2. Sample sizes by variable and cohort

Variable description	Name	Sample size	
		Treatment	Control
Abortion is acceptable for any reason	abany	1,024	1,034
Is abortion okay if chance of serious defect?	abdefect	1,009	1,020
Is abortion okay if woman's health is in jeopardy?	abhlth	1,021	1,025
Is abortion okay if woman does not want more kids?	abnomore	1,022	1,034
Abortion is acceptable for financial reasons	abpoor	1,024	1,036
Abortion is acceptable in event of rape	abrape	1,002	1,024
Abortion is okay if the woman is not married	absingle	1,023	1,041
Number of household members ages 18+	adults	1,515	1,578
Sci. research should be supported by public dollars	advfront	475	1,119
Opinion of affirmative action	affrmact	959	974
Age	age	1,514	1,571
Age when first child was born	agekdbrn	1,106	1,146
Ever read horoscope or astrology report	astrolgy	492	1,163
Believes astrology is scientific	astrosci	474	1,123
Frequency of attendance at religious services	attend	1,533	1,574
Number of household members ages 0-5	babies	1,515	1,560
Feelings about the bible	bible	1,520	1,553
Believes the universe began with a huge explosion	bigbang	358	846
Agrees right and wrong is not black and white	blkwhite	1,511	1,535
Nativity	born	1,535	1,581
Believes father's gene determines sex of child	boyorgrl	219	850
Favor or oppose the death penalty for murder	cappun	1,452	1,496
Number of children	childs	1,535	1,580
Ideal number of children	chldidel	979	998
Subjective class identification	class	1,531	1,567
How close the respondent feels to African Americans	closeblk	1,038	1,060
How close the respondent feels to whites	closewht	1,040	1,060
Believes anti-religionists should be allowed to teach	colath	1,010	1,050
Believes communist teachers should be fired	colcom	993	1,024
Highest college degree	coldeg1	178	423
Believes homosexuals should be allowed to teach	colhomo	1,021	1,055
Believes militarists should be allowed to teach	colmil	1,007	1,038
Believes racists should be allowed to teach	colrac	1,010	1,045

(Continued)

Table A2 (Continued)

Variable description	Name	Sample size	
		Treatment	Control
Ever taken any college-level science course	colsci	488	1,160
Number of college-level science courses	colscinm	185	429
Respondents understanding of questions	comprend	1,533	1,581
Confidence in military	conarmy	1,015	1,044
Confidence in major companies	conbus	1,002	1,039
Confidence in organized religion	conclerg	1,003	1,039
Believes the continents have been moving	condrift	439	1,024
Confidence in education	coneduc	1,023	1,050
Confidence in executive branch	confed	1,006	1,034
Confidence in financial institutions	confinan	1,014	1,047
Confidence in supreme court	conjudge	1,006	1,031
Confidence in organized labor	conlabor	974	1,012
Confidence in congress	conlegis	1,005	1,039
Confidence in medicine	conmedic	1,017	1,054
Confidence in press	conpress	1,015	1,046
Confidence in scientific community	consci	974	1,014
Participation/recording consent	consent	1,536	1,579
Confidence in television	contv	1,016	1,047
Feelings about courts' treatment of criminals	courts	1,430	1,466
Highest degree	degree	1,535	1,581
Specific denomination	denom	761	873
Believes whites are hurt by affirmative action	discaff	1,004	1,038
Equally/less qualified women get jobs instead of men?	discaffm	494	481
Equally/less qualified men get jobs instead of women?	discaffw	483	529
Believes divorce should be easier or more difficult	divlaw	955	986
Ever been divorced or separated	divorce	831	883
Own or rent dwelling	dwelown	1,003	1,024
How many in family earned income	earnrs	1,532	1,580
Believes the earth goes around the sun	earthsun	460	1,081
Years of education	educ	1,533	1,579
Believes electrons are smaller than atoms	electron	388	890
Believes govt. should reduce income inequality	eqwlth	1,011	1,048
Believes human beings developed from animals	evolved	438	1,028

(Continued)

Table A2 (Continued)

		Sample size	
Variable description	Name	Treatment	Control
Ever worked as long as one year	evwork	452	518
Familiar with experimental design	expdesgn	466	1,095
Knows why experimental design is preferred	exptext	459	1,068
Believes people are fair or take advantage of others	fair	1,019	1,056
Reason not living with parents when 16	famdif16	403	502
Number of family generations in household	famgen	1,536	1,581
Living with parents when 16	family16	1,534	1,581
Afraid to walk at night in neighborhood	fear	1,038	1,075
Believes mother working does/does not hurt children	fechld	994	1,019
Better for men to work and women to tend home?	fefam	995	1,011
Make special effort to hire/promote women?	fehire	494	532
For or against preferential hiring of women	fejobaff	491	464
Are women suited for politics?	fepol	946	974
Do preschool kids suffer if mother works?	fepresch	989	1,010
Change in financial situation in last few years	finalter	1,532	1,576
Opinion of family income	finrela	1,517	1,566
How fundamentalist is the respondent?	fund	1,402	1,515
Opinion of how people get ahead	getahead	1,042	1,069
Confidence in the existence of god	god	1,529	1,567
Standard of living will improve	goodlife	1,021	1,053
Number of grandparents born outside the U.S.	granborn	1,418	1,498
Opinion of marijuana legalization	grass	944	969
Opinion of gun permits	gunlaw	1,034	1,063
Happiness of marriage	hapmar	692	760
General happiness	happy	1,525	1,576
Condition of health	health	1,043	1,076
Opinion of government aid for African Americans	helpblk	989	1,016
Are people helpful or selfish?	helpful	1,019	1,059
Should government do more or less?	helpnot	983	1,018
How important is it for kids to learn to help others?	helpoth	517	1,051
Should government improve standard of living?	helppoor	1,001	1,035
Should government help pay for medical care?	helpsick	1,001	1,031
Is the respondent Hispanic?	hispanic	1,533	1,578

(Continued)

Table A2 (Continued)

Variable description	Name	Sample size	
		Treatment	Control
Attitude toward homosexual relations	homosex	998	1,011
Believes the center of the earth is very hot	hotcore	434	1,028
Number of hours worked last week	hrs1	749	952
Ever took a high school biology course	hsbio	467	1,111
Ever took a high school chemistry course	hschem	469	1,111
Highest level of math completed in high school	hsmath	454	1,092
Ever took a high school physics course	hsphys	465	1,109
Does the respondent or their spouse hunt?	hunt	1,043	1,075
Family income when age 16	incom16	1,513	1,561
Total family income	income06	1,488	1,529
Rating of African Americans' intelligence	intlblks	980	993
Rating of whites' intelligence	intlwhts	980	993
Internet access at home	intrhome	488	1,158
Could respondent find equally good job?	jobfind	507	651
Likelihood of losing job	joblose	510	658
Standard of living compared to children's	kidssol	994	1,030
Believes lasers work by focusing sound waves	lasers	362	857
Allow incurable patients to die	letdie1	971	467
Beliefs about immigration	letin1	975	1,007
Allow anti-religionist's book in the library?	libath	1,025	1,051
Allow communist's book in library?	libcom	1,015	1,045
Allow homosexual's book in library?	libhomo	1,022	1,058
Allow militarist's book in library?	libmil	1,018	1,048
Allow racist's book in library?	librac	1,023	1,051
Is life exciting or dull?	life	1,035	1,065
Would live in area where half of neighbors are black?	liveblks	995	1,018
Would live in area where half of neighbors are white?	livewhts	994	1,015
Length of interview	lngthinv	1,533	1,579
Number of employees at work site	localnum	780	979
Mother's years of schooling	maeduc	1,340	1,398
College major	majorcol	193	422
Feelings about relative marrying an Asian	marasian	995	1,021
Feelings about relative marrying an African American	marblk	998	1,021

(Continued)

Table A2 (Continued)

Variable description	Name	Sample size	
		Treatment	Control
Feelings about relative marrying a Hispanic	marhisp	996	1,021
Should homosexuals have the right to marry?	marhomo	1,033	1,065
Marital status	marital	1,534	1,577
Feelings about relative marrying a white person	marwht	998	1,023
Mother's socioeconomic index	masei	817	1,017
Mother's employment when respondent was 16	mawrkgrw	1,428	1,501
Mother self-employed or worked for somebody else	mawrkslf	872	1,033
Men hurt family when they focus too much on work	meovrwrk	999	1,019
Geographic mobility since age 16	mobile16	1,536	1,581
Believes nanotechnology manipulates small objects	nanoknw1	156	411
Believes nanoscale materials are different	nanoknw2	120	326
Familiarity with nanotechnology	nanotech	474	1,132
Costs and benefits of nanotechnology	nanowill	305	701
Amount spent on foreign aid	nataid	746	720
Amount spent on foreign aid (version y)	nataidy	750	782
Amount spent on national defense	natarms	758	742
Amount spent on national defense (version y)	natarmsy	741	779
Amount spent on assistance for childcare	natchld	1,445	1,464
Amount spent on assistance to big cities	natcity	699	687
Amount spent on assistance to big cities (version y)	natcityy	681	697
Amount spent on drug rehabilitation	natdrug	754	735
Amount spent on drug rehabilitation (version y)	natdrugy	711	759
Amount spent on education	nateduc	767	761
Amount spent on education (version y)	nateducy	754	800
Amount spent on environmental protection	natenvir	755	754
Amount spent on environmental protection (version y)	natenviy	746	786
Amount spent on welfare	natfare	747	734
Amount spent on welfare (version y)	natfarey	750	795
Amount spent on health	natheal	769	749
Amount spent on health (version y)	nathealy	754	795
Amount spent on transportation	natmass	1,465	1,477
Amount spent on parks and recreation	natpark	1,504	1,554
Amount spent on assistance to blacks	natrace	695	676

(Continued)

Table A2 (Continued)

Variable description	Name	Sample size	
		Treatment	Control
Amount spent on assistance to blacks (version y)	natracey	669	711
Amount spent on highways and bridges	natroad	1,506	1,537
Amount spent on scientific research	natsci	1,444	1,475
Amount spent on social security	natsoc	1,483	1,511
Amount spent on space exploration	natspac	730	724
How often does the respondent read the newspaper?	news	1,003	1,026
Main source of information about events in the news	newsfrom	492	1,162
Does science give opportunities to future generations?	nextgen	480	1,134
How important is it for children to obey parents?	obey	517	1,051
Test of knowledge about probability	odds1	457	1,085
Test of knowledge about probability	odds2	457	1,089
Gun in home	owngun	1,038	1,074
Father's years of education	paeduc	1,116	1,165
Were parents born in U.S.?	parborn	1,530	1,577
Standard of living compared to parents	parsol	1,012	1,042
Part- or full-time work	partfull	969	1,159
Party identification	partyid	1,531	1,571
Father's socioeconomic index	pasei	1,112	1,223
Father self-employed?	pawrkslf	1,205	1,283
Agrees that morality is a personal matter	permoral	1,494	1,538
Telephone in household	phone	1,536	1,577
Is birth control okay for teenagers between 14-16?	pillok	983	482
Pistol or revolver in home	pistol	1,036	932
Okay for police to hit citizen who said vulgar things?	polabuse	1,003	1,022
Okay for police to hit citizen who is attacking them?	polattak	1,017	1,047
Okay for police to hit citizen if trying to escape?	polescap	979	1,021
Ever approve of police striking citizen	polhitok	982	484
Okay for police to hit murder suspect?	polmurdr	990	1,033
Think of self as liberal or conservative?	polviews	1,493	1,520
Is pope infallible on maters of faith or morals?	popespks	335	167
How important is it for a child to be popular?	popular	517	1,051
Feelings about pornography laws	pornlaw	1,022	1,055
Belief in life after death	postlife	1,377	1,395

(Continued)

Table A2 (Continued)

Variable description	Name	Sample size	
		Treatment	Control
How often does the respondent pray?	pray	1,527	1,564
Should bible prayer be allowed in public schools?	prayer	959	989
Feelings about sex before marriage	premarsx	973	1,005
Which candidate did the respondent vote for in 2004?	pres04	954	976
Number of household members ages 6-12	preteen	1,515	1,560
Agrees that sinners must be punished?	punsin	1,441	1,474
Thinks blacks' disadvant. are due to discrimination?	racdif1	959	990
Thinks blacks' disadvant. are due to disabilities?	racdif2	980	1,008
Thinks blacks' disadvant. are due to lack of education?	racdif3	979	998
Thinks blacks' disadvant. are due to lack of will?	racdif4	955	987
Race	race	1,536	1,581
Any African Americans living in neighborhood?	raclive	1,470	1,528
Feelings about open housing laws	racopen	1,038	1,066
Racial makeup of workplace	racwork	534	637
Believes all radioactivity is man-made	radioact	434	1,034
Ever had a "born again" experience?	reborn	1,517	1,557
Region of residence at age 16	reg16	1,536	1,581
Participates frequently in religious activities?	relactiv	1,530	1,568
Has a religious experience changed life?	relexp	1,526	1,567
Religious preference	relig	1,534	1,574
Strength of religious affiliation	reliten	1,421	1,465
Try to carry religious beliefs into other dealings	rellife	1,511	1,555
Any turning point when less committed to religion?	relneg	754	1,570
Type of community when 16	res16	1,534	1,580
Continue to work if became rich?	richwork	582	704
Respondent's income	rincom06	831	1,009
Agrees that immoral people corrupt society	rotapple	1,509	1,537
Does the gun in the house belong to the respondent?	rowngun	333	382
Relationship to head of household	rplace	1,535	1,578
Is the respondent a visitor in the household?	rvisitor	1,536	1,581
Satisfaction with financial situation	satfin	1,532	1,576
Job satisfaction	satjob	1,054	1,207
Ever tried to convince others to accept Jesus?	savesoul	1,530	1,573

(Continued)

Table A2 (Continued)

		Sample size	
Variable description	Name	Treatment	Control
Do the benefits of scientific research outweigh the costs?	scibnfts	452	1,084
Main source of info about science and technology	scifrom	488	1,152
Has a clear understanding of scientific study?	scistudy	485	1,148
What it means to study something scientifically	scitext	334	844
Likely source of information about scientific issues	seeksci	489	1,142
Socioeconomic index	sei	1,386	1,497
Feelings about sex education in public schools	sexeduc	990	1,010
Number of siblings	sibs	1,534	1,579
Frequently spend an evening at a bar?	socbar	1,002	1,026
Frequently spend evenings with friends?	socfrend	1,001	1,024
Frequently spend evenings with neighbors?	socommun	999	1,025
Frequently spend evenings with relatives?	socrel	1,003	1,026
How long does it take earth to travel around the sun?	solarrev	308	803
Favor spanking to discipline children?	spanking	996	1,012
Spouse's religious denomination	spden	342	431
Spouse's years of education	speduc	692	757
Has spouse ever worked as long as a year?	spevwork	166	226
How fundamentalist is spouse currently?	spfund	660	727
Hours spouse worked last week	sphrs1	413	502
Should anti-religionists be allowed to speak publicly?	spkath	1,039	1,068
Should communists be allowed to speak publicly?	spkcom	1,020	1,051
Should homosexuals be allowed to speak publicly?	spkhomo	1,028	1,059
Should militarists be allowed to speak publicly?	spkmil	1,022	1,057
Should racists be allowed to speak publicly?	spkrac	1,036	1,059
Spouse's socioeconomic index	spsei	609	710
Is spouse self-employed?	spwrkslf	641	732
Spouse's labor force status?	spwrksta	693	759
Is suicide okay if disease is incurable?	suicide1	967	994
Is suicide okay if person is bankrupt?	suicide2	994	1,010
Is suicide okay if person dishonored their family?	suicide3	989	1,012
Is suicide okay if person is tired of living?	suicide4	980	997
Federal income tax	tax	1,016	1,044
Number of household members ages 13-17	teens	1,515	1,559

(Continued)

Table A2 (Continued)

Variable description	Name	Sample size	
		Treatment	Control
Is sex before marriage okay for people ages of 14-16?	teensex	996	1,019
How important is it for kids to think for themselves?	thnkself	517	1,051
Does science make our way of life change too fast?	toofast	478	1,134
Can people be trusted?	trust	1,021	1,060
Hours per day watching television	tvhours	1,001	1,023
Ever unemployed in last ten years?	unemp	1,024	1,060
Union membership	union	1,022	1,055
Number in household who are unrelated	unrelat	946	1,083
Expect another world war in the next 10 years?	uswary	944	970
Believes antibiotics kill viruses as well as bacteria	viruses	454	1,075
Number of visitors in household	visitors	1,536	1,581
Did the respondent vote in the 2004 election?	vote04	1,501	1,541
Weeks worked last year	weekswrk	1,525	1,568
Presence of children under six	whoelse1	1,533	1,581
Presence of older children	whoelse2	1,533	1,581
Presence of other relatives	whoelse3	1,533	1,581
Presence of other relatives	whoelse4	1,533	1,581
Presence of other adults	whoelse5	1,533	1,581
No one else present	whoelse6	1,533	1,581
Ever been widowed	widowed	1,012	1,033
Rating of blacks on wealth scale	wlthblks	985	1,002
Rating of whites on wealth scale	wlthwhts	985	1,004
Wordsum score	wordsum	744	900
Rating of blacks on laziness scale	workblks	985	989
How important is it for a child to work hard?	workhard	517	1,051
Rating of whites on laziness scale	workwhts	988	993
Government or private employee	wrkgovt	1,430	1,503
Self-employed or works for somebody else	wrkslf	1,446	1,534
Labor force status	wrkstat	1,535	1,580
Believes blacks can overcome prejudice without help	wrkwayup	989	1,014
Had sex with person other than spouse	xmarsex	1,037	1,056
Seen x-rated movie in last year	xmovie	1,024	1,054
Astrological sign	zodiac	1,510	1,536

Note: Sample includes respondents who participated in at least the first two waves.

Table A3. Test statistics for all items

		Difference between cohorts	
Variable name	Chi-square/*t*	*p*	FDR-adjusted *p*
abany	1.75	0.19	0.62
abdefect	0.09	0.77	0.91
abhlth	0.21	0.65	0.91
abnomore	0.68	0.41	0.78
abpoor	0.14	0.71	0.91
abrape	0.37	0.54	0.86
absingle	0.30	0.58	0.88
adults	3.50	**0.00**	**0.02**
advfront	1.07	0.77	0.91
affrmact	1.64	0.65	0.91
age	3.08	**0.00**	0.06
agekdbrn	-0.31	0.76	0.91
astrolgy	0.33	0.57	0.87
astrosci	2.88	0.24	0.69
attend	9.59	0.29	0.72
babies	0.92	0.36	0.76
bible	4.73	0.19	0.62
bigbang	3.17	0.07	0.45
blkwhite	3.59	0.31	0.73
born	0.02	0.90	0.94
boyorgrl	3.45	0.06	0.44
cappun	0.51	0.48	0.82
childs	0.96	0.34	0.76
chldidel	-1.32	0.19	0.62
class	2.01	0.57	0.87
closeblk	9.87	0.27	0.71
closewht	11.56	0.17	0.59
colath	0.01	0.91	0.95
colcom	2.03	0.15	0.56
coldeg1	14.77	**0.04**	0.35
colhomo	0.61	0.44	0.79
colmil	4.39	**0.04**	0.35
colrac	3.33	0.07	0.44

(Continued)

Table A3 (Continued)

Variable name	Chi-square/t	Difference between cohorts	
		p	FDR-adjusted p
colsci	8.71	**0.00**	0.06
colscinm	1.11	0.27	0.71
comprend	2.94	0.23	0.68
conarmy	5.21	0.07	0.45
conbus	1.54	0.46	0.81
conclerg	3.41	0.18	0.62
condrift	5.15	**0.02**	0.26
coneduc	1.89	0.39	0.77
confed	0.03	0.98	0.99
confinan	2.51	0.28	0.72
conjudge	2.47	0.29	0.72
conlabor	2.29	0.32	0.73
conlegis	1.91	0.39	0.77
conmedic	0.80	0.67	0.91
conpress	0.59	0.75	0.91
consci	1.99	0.37	0.77
consent	0.14	0.71	0.91
contv	1.23	0.54	0.86
courts	0.71	0.70	0.91
degree	3.32	0.51	0.84
denom	6.08	0.41	0.78
discaff	0.01	1.00	1.00
discaffm	4.20	0.24	0.69
discaffw	2.09	0.55	0.86
divlaw	0.65	0.72	0.91
divorce	0.54	0.46	0.81
dwelown	0.23	0.89	0.94
earnrs	-0.15	0.88	0.94
earthsun	0.02	0.89	0.94
educ	1.56	0.12	0.51
electron	3.80	0.05	0.42
eqwlth	7.26	0.30	0.72
evolved	1.29	0.26	0.70

(Continued)

Table A3 (Continued)

Variable name	Chi-square/*t*	Difference between cohorts	
		p	FDR-adjusted *p*
evwork	1.12	0.29	0.72
expdesgn	0.14	0.71	0.91
exptext	0.13	0.72	0.91
fair	3.16	0.21	0.63
famdif16	4.93	0.25	0.69
famgen	17.70	**0.01**	0.09
family16	5.36	0.72	0.91
fear	0.44	0.51	0.84
fechld	14.49	**0.00**	0.06
fefam	0.62	0.89	0.94
fehire	1.79	0.78	0.91
fejobaff	2.33	0.51	0.84
fepol	0.06	0.81	0.92
fepresch	1.04	0.79	0.91
finalter	1.50	0.47	0.82
finrela	6.96	0.14	0.53
fund	3.95	0.14	0.53
getahead	0.95	0.62	0.90
god	3.79	0.58	0.88
goodlife	8.60	0.07	0.45
granborn	3.16	0.53	0.86
grass	0.72	0.40	0.77
gunlaw	0.08	0.78	0.91
hapmar	3.16	0.21	0.63
happy	4.42	0.11	0.49
health	0.67	0.88	0.94
helpblk	1.07	0.90	0.94
helpful	3.69	0.16	0.57
helpnot	6.81	0.15	0.54
helpoth	1.84	0.78	0.91
helppoor	2.74	0.60	0.89
helpsick	4.76	0.31	0.73
hispanic	21.83	0.26	0.70

(Continued)

Table A3 (Continued)

		Difference between cohorts	
Variable name	Chi-square/*t*	*p*	FDR-adjusted *p*
homosex	4.72	0.19	0.62
hotcore	0.17	0.68	0.91
hrs1	2.59	**0.01**	0.14
hsbio	0.09	0.77	0.91
hschem	2.64	0.10	0.48
hsmath	13.38	0.15	0.54
hsphys	0.04	0.83	0.94
hunt	3.05	0.38	0.77
incom16	3.12	0.54	0.86
income06	35.57	0.08	0.46
intlblks	5.44	0.49	0.83
intlwhts	3.31	0.77	0.91
intrhome	1.34	0.25	0.69
jobfind	5.40	0.07	0.44
joblose	0.39	0.94	0.96
kidssol	5.30	0.38	0.77
lasers	0.12	0.72	0.91
letdie1	0.02	0.89	0.94
letin1	1.53	0.82	0.93
libath	2.83	0.09	0.46
libcom	2.18	0.14	0.53
libhomo	0.04	0.84	0.94
libmil	3.07	0.08	0.46
librac	2.09	0.15	0.54
life	0.87	0.65	0.91
liveblks	4.77	0.31	0.73
livewhts	2.09	0.72	0.91
lngthinv	-3.43	**0.00**	**0.02**
localnum	3.90	0.69	0.91
maeduc	-0.83	0.40	0.78
majorcol	9.41	0.09	0.46
marasian	5.83	0.21	0.64
marblk	3.59	0.46	0.81

(Continued)

Table A3 (Continued)

Variable name	Chi-square/*t*	Difference between cohorts	
		p	FDR-adjusted *p*
marhisp	4.43	0.35	0.76
marhomo	4.33	0.36	0.77
marital	18.24	**0.00**	**0.04**
marwht	2.28	0.68	0.91
masei	1.23	0.22	0.66
mawrkgrw	0.07	0.79	0.91
mawrkslf	0.15	0.70	0.91
meovrwrk	7.15	0.13	0.51
mobile16	3.22	0.20	0.63
nanoknw1	1.49	0.22	0.66
nanoknw2	0.11	0.74	0.91
nanotech	4.72	0.19	0.62
nanowill	4.78	0.09	0.46
nataid	1.37	0.51	0.84
nataidy	0.84	0.66	0.91
natarms	2.17	0.34	0.76
natarmsy	2.44	0.29	0.72
natchld	3.97	0.14	0.53
natcity	5.84	0.05	0.43
natcityy	2.24	0.33	0.75
natdrug	9.42	**0.01**	0.14
natdrugy	2.15	0.34	0.76
nateduc	2.73	0.26	0.70
nateducy	0.90	0.64	0.91
natenvir	0.69	0.71	0.91
natenviy	0.25	0.88	0.94
natfare	2.07	0.36	0.76
natfarey	1.03	0.60	0.89
natheal	0.52	0.77	0.91
nathealy	0.16	0.92	0.95
natmass	4.22	0.12	0.51
natpark	1.74	0.42	0.78
natrace	1.69	0.43	0.79

(Continued)

Table A3 (Continued)

Variable name	Chi-square/*t*	Difference between cohorts	
		p	FDR-adjusted *p*
natracey	15.14	**0.00**	**0.02**
natroad	1.79	0.41	0.78
natsci	0.66	0.72	0.91
natsoc	1.77	0.41	0.78
natspac	1.21	0.55	0.86
news	1.93	0.75	0.91
newsfrom	3.38	0.97	0.98
nextgen	1.49	0.71	0.91
obey	3.10	0.54	0.86
odds1	1.24	0.26	0.70
odds2	1.12	0.29	0.72
owngun	0.79	0.67	0.91
paeduc	-1.00	0.32	0.73
parborn	3.30	0.86	0.94
parsol	21.25	**0.00**	**0.02**
partfull	9.94	**0.00**	**0.05**
partyid	13.49	0.06	0.44
pasei	-0.22	0.83	0.94
pawrkslf	1.88	0.17	0.59
permoral	7.10	0.07	0.44
phone	157.30	**0.00**	**0.00**
pillok	6.29	0.10	0.48
pistol	1.23	0.54	0.86
polabuse	0.04	0.83	0.94
polattak	0.91	0.34	0.76
polescap	0.02	0.88	0.94
polhitok	0.10	0.75	0.91
polmurdr	0.03	0.86	0.94
polviews	4.96	0.55	0.86
popespks	4.46	0.35	0.76
popular	8.54	0.06	0.44
pornlaw	4.74	0.09	0.46
postlife	0.17	0.68	0.91

(Continued)

Table A3 (Continued)

Variable name	Chi-square/t	Difference between cohorts	
		p	FDR-adjusted p
pray	9.58	0.09	0.46
prayer	0.71	0.40	0.78
premarsx	12.52	**0.01**	0.10
pres04	11.42	**0.01**	0.14
preteen	2.44	**0.01**	0.20
punsin	0.32	0.96	0.97
racdif1	2.37	0.12	0.51
racdif2	0.73	0.39	0.77
racdif3	0.61	0.44	0.79
racdif4	0.12	0.73	0.91
race	0.53	0.77	0.91
raclive	0.02	0.88	0.94
racopen	1.89	0.39	0.77
racwork	7.29	0.12	0.51
radioact	5.77	**0.02**	0.21
reborn	0.01	0.93	0.95
reg16	4.33	0.89	0.94
relactiv	16.99	**0.05**	0.42
relexp	0.08	0.77	0.91
relig	3.60	0.46	0.81
reliten	1.12	0.77	0.91
rellife	1.28	0.73	0.91
relneg	0.17	0.68	0.91
res16	1.80	0.88	0.94
richwork	5.73	**0.02**	0.21
rincom06	47.49	**0.00**	0.08
rotapple	3.99	0.26	0.70
rowngun	0.78	0.68	0.91
rplace	26.24	**0.00**	**0.02**
rvisitor	0.23	0.63	0.91
satfin	4.84	0.09	0.46
satjob	3.72	0.29	0.72
savesoul	0.27	0.60	0.89

(Continued)

Table A3 (Continued)

		Difference between cohorts	
Variable name	Chi-square/*t*	*p*	FDR-adjusted *p*
scibnfts	4.51	0.10	0.48
scifrom	7.30	0.61	0.89
scistudy	4.95	0.08	0.46
scitext	17.83	**0.00**	0.06
seeksci	4.72	0.69	0.91
sei	2.28	**0.02**	0.26
sexeduc	0.28	0.60	0.89
sibs	-0.24	0.81	0.92
socbar	4.94	0.55	0.86
socfrend	11.55	0.07	0.45
socommun	7.37	0.29	0.72
socrel	2.54	0.86	0.94
solarrev	1.24	0.54	0.86
spanking	0.23	0.97	0.98
spden	26.34	0.39	0.77
speduc	0.66	0.51	0.84
spevwork	0.03	0.87	0.94
spfund	4.51	0.10	0.48
sphrs1	1.53	0.13	0.51
spkath	0.53	0.47	0.81
spkcom	4.14	**0.04**	0.39
spkhomo	0.75	0.39	0.77
spkmil	2.93	0.09	0.46
spkrac	8.93	**0.00**	0.06
spsei	0.86	0.39	0.77
spwrkslf	0.61	0.43	0.79
spwrksta	2.40	0.93	0.96
suicide1	0.07	0.79	0.91
suicide2	3.86	**0.05**	0.42
suicide3	1.39	0.24	0.69
suicide4	0.28	0.60	0.89
tax	1.61	0.45	0.80
teens	0.87	0.38	0.77

(Continued)

Table A3 (Continued)

		Difference between cohorts	
Variable name	Chi-square/*t*	*p*	FDR-adjusted *p*
teensex	1.28	0.73	0.91
thnkself	8.76	0.07	0.44
toofast	7.82	**0.05**	0.42
trust	4.30	0.12	0.51
tvhours	-1.39	0.17	0.58
unemp	0.02	0.88	0.94
union	0.23	0.97	0.98
unrelat	-0.76	0.44	0.80
uswary	0.02	0.90	0.94
viruses	5.32	**0.02**	0.25
visitors	4.45	**0.00**	**0.00**
vote04	19.21	**0.00**	**0.01**
weekswrk	-2.10	**0.04**	0.35
whoelse1	0.50	0.48	0.82
whoelse2	0.52	0.47	0.82
whoelse3	3.65	0.06	0.44
whoelse4	0.35	0.56	0.86
whoelse5	0.13	0.72	0.91
whoelse6	4.72	**0.03**	0.32
widowed	0.01	0.94	0.96
wlthblks	9.93	0.11	0.50
wlthwhts	5.15	0.53	0.86
wordsum	1.63	0.10	0.48
workblks	6.70	0.35	0.76
workhard	6.11	0.19	0.62
workwhts	7.94	0.24	0.69
wrkgovt	1.61	0.20	0.63
wrkslf	0.17	0.68	0.91
wrkstat	13.41	0.06	0.44
wrkwayup	0.99	0.91	0.95
xmarsex	1.81	0.61	0.89
xmovie	1.06	0.30	0.73
zodiac	7.60	0.75	0.91

Note: The FDR-adjusted *p* is the *p* -value adjusted for the False Discovery Rate. Bolded *p* values are less than .05.

CHAPTER 3: ASSESSING PANEL CONDITIONING USING EXPERIMENTS

Researchers who collect and analyze longitudinal data nearly always assume that respondents' attributes are unaffected by the survey experience. When researchers observe changes in an individual's attributes across waves of a longitudinal survey, they usually assume that the changes are real and would have occurred even if the individual had not been interviewed. This assumption runs counter to research in cognitive psychology, political science, and elsewhere, which suggests that the experience of answering survey questions can sometimes affect (1) people's actual attributes and/or (2) their willingness or ability to provide accurate and valid information about those attributes during subsequent interviews. In the first case, users of longitudinal data would observe within-person changes in some focal attribute that only occurred because that attribute was being measured. In the second case, users would observe within-person changes even when the underlying attribute remained stable.

In this chapter, I present results from an experiment on "panel conditioning," or bias that arises from survey respondents' repeated exposure to survey questions.[16] Previous investigations into this issue have typically relied on small, non-representative samples and/or research designs that conflate panel conditioning with panel attrition. The procedure that I outline below resolves these problems by randomly assigning members of a nationally-representative sample to various treatment conditions. Through this assignment process, I

[16] The data that I use in this chapter came from a survey-based experiment conducted as a part of the Time-sharing Experiments in the Social Sciences (TESS) program, which receives support from the Social, Behavioral, and Economic Sciences Directorate of the National Science Foundation.

am able to (1) obtain information about the incidence and magnitude of panel conditioning effects; and (2) assess theoretically-informed hypotheses about *when* panel conditioning is most likely to arise and *why* it does so.

Background

Panel conditioning has received attention in the social sciences since at least the 1940s, when Lazarfeld (1940:128) identified it as one of the "big problem[s] yet unsolved" in quantitative research. Since that time, scholars in several disciplines have sought to identify panel conditioning effects across a wide range of survey settings (see, e.g., Bailar 1975; Barber, Gatny, and Kusunoki 2012; Bartels 1999; Crespi 1948; Das et al. 2011; Halpern-Manners and Warren 2012; Sherman 1980; Spangenberg et al. 2008; Torche et al. 2012; Veroff, Hatchett, and Douvan 1992; Zwane et al. 2011). In general, the available evidence suggests that panel conditioning *can* occur within the context of a longitudinal data collection effort (Warren and Halpern-Manners 2012), potentially influencing the quality of both attitudinal and behavioral items (Cantor 2008; Sturgis, Allum, and Brunton-Smith 2009).

Progress towards understanding what drives these effects has unfortunately been limited. Although it makes sense to think that the magnitude of panel conditioning might be dependent on the nature of the survey, the characteristics of the respondents, and/or the way the survey is administered, these possibilities have not been explored except in relatively isolated domains (see, e.g., Spanenberg 2012). Some researchers have proposed hypotheses about the kinds of questions that could, in theory, lead to panel conditioning (see, e.g., Sturgis et al. 2009; Warren and Halpern-Manners 2012; Waterton and Lievesley 1988), but these hypotheses have not been tested using strong techniques and representative data. As far as I know, no study has produced a unified (and well-validated) theoretical model that is capable of predicting when panel conditioning effects will and will not occur.

Perhaps the biggest impediment to developing such a model has been our inability to *measure* panel conditioning accurately. In most cases, scholars have proceeded by comparing the responses given by members of a longitudinal panel to the responses given by members of an independent cross-sectional sample drawn from the same population. The problems with this approach have been discussed at length elsewhere (Cantor 2008; Das, Toepoel, and Soest 2007; Holt 1989; Warren and Halpern-Manners 2012). Most importantly, such a design risks confounding biases from panel *conditioning* with biases from panel *attrition*. Whereas the fresh cross-sectional sample may be representative of the original target population, the panel sample may have suffered from non-random attrition over time. Efforts to adjust for the resulting selectivity (via post-stratification weighting and/or related techniques) may or may not resolve this problem.[17]

As I pointed out in the previous chapter, one way to overcome this issue is to conduct targeted methodological experiments (see, e.g., Barber et al. 2012; Borle et al. 2007; Godin et al. 2008; Torche et al. 2012; Veroff et al. 1992; Williams, Block, and Fitzsimons 2006; Zwane

[17] Post-stratification weights account for observable differences between groups with different levels of prior survey experience, but groups may still differ on unobserved characteristics.

et al. 2011). Although these experiments can take a variety of forms, the most common approach is to assign different sets of survey questions to randomly selected groups of respondents and then to re-interview all respondents at a later date using the complete battery of items (see, e.g., Torche et al. 2012). If the experimental design is properly implemented, any differences that emerge between groups after the first interview must be the result of panel conditioning. Internal validity is maximized through randomization, allowing for a more careful assessment of panel conditioning effects *(as well as their underlying causes and potential scope conditions)*.

Although this approach provides a much stronger warrant for identifying panel conditioning effects, most of the experimental studies that I am aware of have focused on a fairly narrow range of substantive topics (e.g., purchasing particular consumer goods, voting in local elections, or cheating on tests) and have usually been conducted using small convenience samples (e.g., customers at certain businesses, voters in specific precincts, or students in college classrooms). These studies provide information about the incidence and magnitude of panel conditioning effects under specific empirical conditions, but the broader generalizability of their results is often unclear. To my knowledge, no study has ever attempted to assess panel conditioning effects by intentionally embedding a well-designed methodological experiment within a nationally-representative social science survey.[18]

[18] Over the years, there have been a few cases where panel conditioning experiments have been built into longitudinal surveys *by accident* (see, e.g., Warren and Halpern-Manners 2012).

In this chapter, I present findings from exactly this sort of study. My main objective—as I stated earlier—is to identify *particular instances* (e.g., types of surveys and types of survey questions) in which panel conditioning effects may be more or less likely to arise. As a first step toward accomplishing this objective, I present a series of hypotheses concerning the nature and scope of panel conditioning in longitudinal social science data. It is to these hypotheses that I turn to next.

Three hypotheses concerning the nature of panel conditioning

Responding to a social science survey is a cognitively and socially complex process that may or may not leave respondents unchanged and/or equally willing (or able) to provide accurate information when re-interviewed in subsequent waves. While there is no well-validated theoretical model of when these changes will occur, the research literature does provide a number of interesting hypotheses (see, e.g., Cantor 2008; Das et al. 2011; Fitzsimons and Moore 2008; French and Sutton 2010; Sherman 1980; Spangenberg et al. 2008; Spangenberg et al. 2012; Sturgis et al. 2009; Torche et al. 2012; Warren and Halpern-Manners 2012; Waterton and Lievesley 1989; Zwane et al. 2011). In this chapter, I focus my attention on three of these hypotheses—two that deal with the content of the survey and one that relates to the way surveys are administered.

> ***Hypothesis 1:*** Respondents' behaviors will appear to change over time when the experience of being surveyed activates previously forgotten or faintly held memories.

Surveys often ask respondents to provide information about their past experiences. To answer these questions accurately, respondents must be able to retrieve the relevant information from their memory in a relatively short amount of time (Krosnick, Narayan, and Smith 1996). Whether or not they are able to do so depends on the nature of the question and its salience to the respondent (Schaeffer and Presser 2003). If the question deals with an unusual or dramatic event (like an arrest), or an event that occurred with some regularity (like habitual drug use), recall is likely to be high (Pearson, Ross, and Dawes 1994). Events of this sort tend to grab people's attention and hold it long enough to ensure that a rich representation is created and stored in their long-term memory (Tourangeau 2000). When asked about the experience on a survey, all the respondent has to do is tap into that memory and provide the appropriate response.

This process becomes more complicated when surveys ask about less salient events or events that occurred infrequently in the distant past. In these cases, respondents may not have the time or energy to engage in the kind memory work that would be required to retrieve an accurate answer "on the fly" (Tourangeau 2000). As an alternative, some respondents may choose to base their answers on what they perceive to be probable ("I'm generally a cautious person and I rarely drink, so I've probably never driven drunk") or appropriate ("Driving drunk is dangerous and against the law, so it's not something that I would do"). In some cases, this experience can trigger additional contemplation *after the survey is over* (Cantor 2008). If this contemplation reveals previously forgotten details ("Actually, there was one time…”), it could lead to altered reports from one survey wave to the next. I call this the *memory activation hypothesis*.

> ***Hypothesis 2:*** Respondents’ behaviors will (appear to) change over time when surveys ask about socially non-normative or stigmatized topics.

Survey questions often force respondents to confront the reality that their behaviors conflict with what mainstream society regards as normative or appropriate. Reflecting on this conflict can create “cognitive dissonance” in respondents’ minds, as well as feelings of embarrassment and/or psychological discomfort (Perkins et al. 2008). In the context of research on the effects of answering questions about behavioral intentions or expectations on respondents’ actual future behaviors, cognitive psychologists note that one way for respondents to mitigate cognitive dissonance is to modify their subsequent behaviors (Dholakia 2010; Fitzsimons and Moore 2008). I contend that this process, which is well documented for questions regarding predicted or intended behaviors, may also operate in the context of questions regarding *previous* or *ongoing* behaviors.[19]

Some respondents may react to questions about sensitive topics by bringing their *actual* behaviors into closer conformity with social norms. Others may simply avoid cognitive dissonance and the embarrassment associated with offering non-normative or stigmatized

[19] Several scholars have suggested similar processes for attitudinal items, whereby repeated questioning prompts respondents to reflect on their attitudes and beliefs, and then realign them with what they determine to be mainstream opinion (see, e.g., Sturgis et al. 2009)

responses by bringing their *reported* behaviors into closer conformity with what they perceive to be normative or appropriate (Dholakia 2010).[20] A survey of respondents' drug use, for example, may serve as an opportunity for individuals to reflect on that behavior; respondents who view their drug use as excessive or embarrassing may subsequently use less or else report inaccurately lower levels of drug use on follow-up surveys (see, e.g., Torche et al. 2012).[21] In either case, apparent declines in levels of drug use over time would represent a form of panel conditioning. I term this the *social stigma hypothesis*.[22]

> ***Hypothesis 3:*** Panel conditioning is more likely to occur when surveys are spaced more closely together in time.

Although the questions asked on a survey may make an immediate impression on respondents, there is no guarantee that this impression will last until their next interview (Chandon, Morwitz, and Reinartz 2004; Zwane et al. 2011). The longer the interval between surveys, the more that intervening life events, subject maturation and change, historical events, forgetfulness, and other factors may overwhelm, counteract, or mute the effects of

---

[20] Lacking external validating information, it is hard to determine whether this form of panel conditioning changes respondents' actual behaviors, their reported behaviors, or both.

[21] This sort of behavior—in which respondents deny, contradict, or "take back" their earlier reports—is often referred to as "recanting" in the drug-use literature (see, e.g., Percy et al. 2005).

[22] It is important not to confuse this with social desirability bias. In the absence of panel conditioning, I would expect the magnitude of social desirability bias to be the same at time 1 and time 2 of a longitudinal survey. Inferences about change over time in focal attributes would be *accurate*, but inferences about levels would not. If the social stigma hypothesis is correct, then inferences about change over time would be *inaccurate* because the reported change would not have occurred had the time 1 survey not taken place.

answering baseline questions on respondents' answers to identical items in subsequent waves. Psychologists often describe this process in terms of accessibility: when interview occasions are separated by longer periods of time, it becomes more difficult for respondents to *access* the ideas and/or feelings that were elicited during their prior interview, thus reducing the chances of panel conditioning (Feldman and Lynch 1988).

Although I have not carried out a thorough meta-analysis, my reading of the literature is that when baseline and follow-up surveys are separated by a month or less, panel conditioning effects are usually observed (see, e.g., Bailar 1989; De Amici et al. 2000; Halpern-Manners and Warren 2012). When baseline and follow-up surveys are separated by between a month and a year, the results are quite split: many studies find panel conditioning (see, e.g., Porst and Zeifang 1987; Shack-Marquez 1986; Solon 1986; Torche et al. 2012), but many do not (see, e.g., Corder and Horvitz 1989; Underwood et al. 2006). Panel conditioning effects are rarely observed—although they have rarely been sought—when baseline and follow-up surveys are separated by more than a year (but see the results presented in Chapter 2).[23] I refer to this as the *survey spacing hypothesis*.

The present study

In order to examine these hypotheses, I conducted an experiment using respondents from

[23] Although I cannot say for sure, it could be the case that the panel conditioning effects I observed in the previous chapter would have been larger (and more widespread) had the time interval between waves been less than two years.

large-scale internet survey. In this section, I describe the design of the experiment, the questionnaire that I used, and my basic estimation strategy.

Study design

The subjects in my study were members of a probability-based, non-volunteer random sample of Americans (ages 18 and older) participating in a GfK KnowledgePanel online survey.[24] Panel members were recruited using a statistically valid sampling method based on a published sampling frame of residential addresses that covers approximately 97% of U.S. households (Yeager et al. 2011).[25] To ensure recruitment of a nationally representative sample, panelists without an internet connection were granted free internet access via a netbook computer. The panel typically includes around 50,000 adults, although the actual size of the sample fluctuates over time due to voluntary withdrawal, involuntary retirement of longtime members, and the addition of new panelists from on-going recruitment (for further details on the KnowledgePanel sample see Dennis 2010).

My experiment began with a random subset of 7,654 members of the larger KnowledgePanel sample.[26] In September of 2011, these respondents completed a short online survey that included two modules of questions. The first module asked respondents *two*

---

[24] GfK was formerly known as Knowledge Networks.

[25] Recent research on the efficacy of internet-based probability samples suggests that their reliability is similar to, and sometimes better than, the reliability of random digit dialing telephone surveys (Yeager et al. 2011).

[26] The response rate for the baseline survey was 61.6%.

randomly selected questions (from a list of four possible items) concerning their lifetime experiences with marijuana use, petty theft, drunk driving, and arrest (see the Appendix for more details about each item, including the wording and response options); the ordering of the questions was also randomized so as to avoid potential sequencing effects.[27] The second module collected basic socioeconomic, demographic, and other background information from all panelists. Table 3.1 provides descriptive statistics for these measures disaggregated by survey ballot (e.g., questions 1 and 2 at baseline, questions 3 and 4 at baseline, and so on); there appear to be few significant differences between groups.

Respondents to the September 2011 baseline survey were randomly assigned to be re-interviewed either one month *or* one year after the baseline interview (but not both). In October 2011, those selected for the one-month follow-up (n = 2,852) were invited to complete a survey that included *all four* questions from the first module (again, in random order). In September 2012, those selected for the one-year follow-up (n = 4,802) were asked to complete a survey that likewise asked *all four* questions (also in random order). The response rates to the two follow-up surveys were 76.6% and 65.0%, respectively; supplementary analyses suggest that attrition from the panel was unrelated to treatment condition (i.e., the ballot received in the baseline interview had no effect on respondents' propensity to remain in the sample until the follow-up wave, regardless of whether the

[27] To maintain the integrity of the experiment, I only selected participants who had *not* previously been asked questions in the KnowledgePanel survey about the topics covered by my four focal survey questions.

Table 3.1 Descriptive statistics, by survey ballot at baseline

	Questions received at baseline					
	1 and 2	1 and 3	1 and 4	2 and 3	2 and 4	3 and 4
Age	50.5	50.0	50.2	50.3	49.7	48.9
	(16.7)	(16.6)	(16.5)	(16.1)	(16.4)	(16.7)
Female (%)	52.4	48.1	48.9	48.3	51.7	50.2
White (%)	74.4	76.2	77.2	76.7	73.4	75.2
Black (%)	10.0	9.2	9.9	8.6	8.9	9.8
Hispanic (%)	8.5	7.9	7.6	8.2	9.7	8.3
Other (%)	7.2	6.7	5.4	6.4	8.0	6.7
Married (%)	60.8	57.5	59.9	58.4	59.5	57.1
Household size	2.7	2.6	2.7	2.7	2.7	2.7
	(1.4)	(1.4)	(1.4)	(1.5)	(1.5)	(1.5)
Head of household (%)	86.6	82.43	82.66	84.24	82.93	83.46
Employed (%)	54.31	58.12	57.45	55.53	67.09	55.32
Income	70293.9	71607.8	69353.8	66709.8	67529.4	66378.3
	(47349.7)	(48933.3)	(47027.8)	(46179.3)	(45896.0)	(46484.8)
High school dropout (%)	7.4	7.8	8.3	8.5	7.7	8.2
High school graduate (%)	27.5	27.9	28.9	29.1	28.0	28.8
Some college (%)	30.5	29.9	30.1	30.6	30.2	30.0
BA or higher (%)	34.6	34.4	32.7	31.8	34.2	33.1
Lives in metro area (%)	85.74	84.86	84.71	83.76	84.26	84.33
Home owner (%)	76.8	74.8	76.0	75.2	74.9	75.6

Note: Questions are as follows: 1 (“Not counting minor traffic violations, have you ever been arrested and booked for breaking the law?”), 2 (“Have you ever used marijuana, for example: grass or pot, in your lifetime?”), 3 (“Have you ever stolen something from a store or something that did not belong to you worth less than 50 dollars?”), 4 (“Have you ever driven a vehicle while you were under the influence of alcohol?”). Note that question assignment and question order were fully randomized.

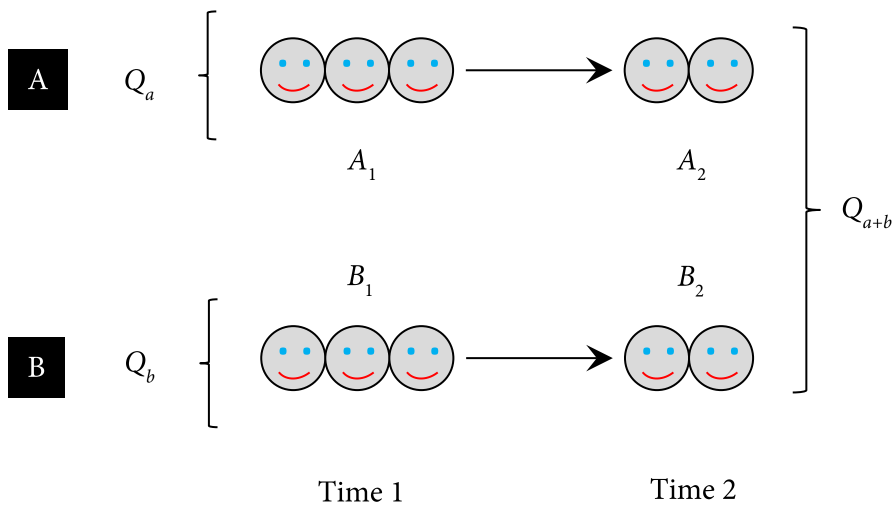

Figure 3.1: Identifying panel conditioning effects using an experimental design. In this simplified example, respondents are randomly assigned to one of two treatment groups (in the actual experiment there are six different groups and three time points). At baseline, respondents in Group A receive survey ballot Q_a; respondents in Group B receive ballot Q_b. In the follow-up wave, respondents from both groups receive the full battery of questions, denoted Q_{a+b}. If there are no panel conditioning effects, there should be no differences (beyond sampling error) between groups at Time 2 with respect to Q_{a+b} (i.e., $A_2 - B_2 = 0$). Respondents in Group A should provide similar answers to Q_b as respondents Group B, and respondents in Group B should provide similar answers to Q_a as respondents Group A.

respondent was in the one-month or one-year group). Results from these analyses can be found in Appendix Table B1.

Does exposure to survey questions that activate previously forgotten memories affect subsequent answers to the same questions? If the memory activation hypothesis has merit, I would expect to see differences across groups (i.e., respondents who were randomly assigned the focal questions as a part of the baseline survey versus those who were not) with respect to some items but not others. In particular, I would expect to see *higher* rates of drunk driving and petty theft among "treated" respondents and *lower* rates among individuals in the "control" group. These events are not always as memorable or clear cut as the others in my study and may be more difficult to recall within the context of a short online survey—especially if they occurred sporadically or at a much earlier point in the respondent's life.[28] I would not expect to see similar effects for marijuana use or arrests; these events are likely to be remembered by most panelists regardless of how long ago they occurred.[29]

Do questions that elicit non-normative, stigmatizing, or embarrassing responses lead to panel conditioning? If the social stigma hypothesis has merit, I would expect to see *lower levels*

---

[28] If the theft question had asked respondents to describe their lifetime experiences with *grand theft*—as opposed to the theft of items or goods that are valued at *less than $50*—issues related to memory and recall might be less of a concern.

[29] The findings reported by Harrison and Hughes (1997), for instance, suggest that recall errors are typically small for questions that ask respondents to provide information about their lifetime experiences with illicit drugs, including marijuana. The same is not true for questions that ask about alcohol consumption—a behavior that is often less distinct in respondents' minds.

of illicit behavior among the treatment group for all four items. Each of the questions that I asked on the survey provided the respondent with two response options, one that is clearly normative (e.g., "No, I've never been arrested—*because I follow the law*") and one that carries with it potentially negative connotations (e.g., "Yes, I have been arrested—*because I'm a criminal*"). For respondents who see themselves as upstanding and moral adults, the experience of selecting the second response option could create a "behavioral discrepancy" between the way they see themselves and the way they actually behave (Dholakia 2010). As I noted earlier, one way for respondents to resolve this discrepancy is to select the first response option after their baseline interview. Theory suggests that this should occur for each item on the survey.[30]

Does the incidence and magnitude of panel conditioning depend on the amount of time that elapses between survey waves? Because participants in the study were assigned at random to be re-interviewed after one month or after one year, I can compare the incidence and magnitude of panel conditioning effects over time. Although most social science surveys interview respondents annually or less often, other important surveys (e.g., the Current Population Survey, the Survey of Income and Program Participation, and the National Crime

[30] Because all of the participants in the experiment took the survey the same number of times, I can rule out possible familiarity effects (i.e., respondents providing more truthful answers as they become more comfortable with the persons or organizations collecting their information) and effects from increased survey experience (i.e., respondents providing more accurate answers as they learn what constitutes a valid and useful response). This is fortuitous for my purposes, as it allows for a cleaner assessment of the mechanisms in question.

Victimization Survey) contact respondents more frequently and with substantially less lag-time between interview waves. The design of the study allows me to consider the possibility that panel conditioning is *most* pronounced when survey follow-ups occur more closely together in time. If I find panel conditioning effects among the one-month group, but *not* the one-year group, this hypothesis would be supported.

Estimation strategy

To obtain estimates of panel conditioning effects, I regressed respondents' time 2 answers (at the time of their one-month or one-year follow-up surveys) onto dichotomous variables indicating whether they were exposed to the focal question during their initial interview; I did this separately for each time-in-survey group.[31] Although there is good covariate balance across the various treatment conditions, I included a number of "pre-treatment" variables in my models.[32] These covariates include common socio-demographic indicators (e.g., education, income, age, race/ethnicity, gender, marital status, householder status, home ownership, employment status, and family size), a measure of political ideology (as observed

---

[31] Because my outcomes are all binary "yes/no" variables (e.g., "Have you ever stolen something from a store or something that did not belong to you worth less than 50 dollars?"), I relied on a series of linear probability models to recover estimates of the underlying treatment effects (e.g., the effect of receiving the theft question in the baseline interview). This estimator has drawbacks—especially the possibility of predicting probabilities that are less than zero or greater than one—but it has also been shown to be more robust than limited dependent variable methods for randomized experiments if the treatment effect is small and the probability of the event is not extreme (Angrist 2001). My experiment satisfies both of these conditions.

[32] Doing so reduces residual variance and increases the precision of my estimates.

at baseline), and an indicator of whether or not the respondent lived in an urban or rural area (again, as observed at baseline). Later in the chapter, I discuss alternative specifications that allow panel conditioning effects to vary along some of these dimensions.

Results

In Table 3.2, I present the results of linear probability models that regress each outcome on treatment condition and controls that adjust for random differences across treatment groups in socioeconomic and demographic characteristics, as I described above. Table 3.2 reports coefficients and standard errors for the key independent variable in each model—whether respondents were asked the focal survey question in their baseline interview.[33] At the one-year follow-up (as shown in the second row of results), I find no significant differences between the several treatment conditions. For these respondents, the probability of answering "yes" or "no" to a particular item did *not* depend on whether they had previous exposure to the survey question. All of the parameter estimates tend toward zero and all are small relative to their standard error.

The same is not true among respondents who participated in the one-month follow-up (as shown in the first row of estimates in Table 3.2). For this group, I find two significant results: individuals who were asked the baseline question about theft were *more* likely to subsequently report ever having stolen something ($\hat{\beta} = .045$; $p < .10$), and individuals who

[33] Throughout the analysis, I estimate robust standard errors to correct for heteroskedasticity.

Table 3.2 Linear probability models predicting deviant behavior, by follow-up group

	Answers "yes" in the follow-up interview to the question…							
	Have you ever been arrested?		Have you ever smoked pot?		Have you ever stolen anything?		Have you ever driven drunk?	
Follow-up group	β	(SE)	β	(SE)	β	(SE)	β	(SE)
Received the focal question at baseline and during the **one-month** follow-up	-0.003	(0.021)	0.017	(0.027)	0.045 *	(0.025)	0.056 **	(0.027)
Received the focal question at baseline and during the **one-year** follow-up	0.005	(0.019)	-0.020	(0.024)	0.007	(0.022)	0.005	(0.023)

Note: Model estimates represent percentage point changes in the probability of answering "yes" to survey items at the time of the follow-up interview. To reduce residual variance and increase the precision of the estimated coefficients, all models include the covariates listed in Table 1. The sample size for the one-month group was 2,185; the sample size for the one-year group was 3,122. Cluster-robust standard errors are shown in parentheses. See text for more details.

* $p < .10$; ** $p < .05$; *** $p < .01$

were asked the baseline question about drunk driving were *more* likely to subsequently report ever having driven a vehicle while under the influence of alcohol ($\hat{\beta} = .056$; $p < .05$). Both effects are modestly-sized—the estimated coefficients correspond to a 12% increase in the probability of answering "yes" to the theft question, and an 11% increase in the probability of answering "yes" to the drunk driving question—and both are consistent with the memory activation hypothesis that I outlined earlier.[34]

To more formally test whether these estimates varied significantly between the one-month and one-year follow-ups, I pooled the data and entered interactions between the treatment indicators and a measure of respondents' time-between-waves group (1 = one month; 0 = one year). The results, which I display at the bottom of Table 3.3, are consistent with the patterns described above: for the questions dealing with theft and drunk driving, there are significant and positive interactions ($p < .10$) between the time-between-waves indicator and the indicator of whether or not the respondent was exposed to the focal item at baseline. This implies that panel conditioning is *more* likely to occur when surveys are separated by *shorter* amounts of time (e.g., one month versus one year), a finding that would seem to support my hypothesis concerning survey spacing.

---

[34] An alternate explanation, which has been suggested by some psychologists, is that exposure to questions about illicit behaviors triggers respondents' *latent* desire to engage in those behaviors, resulting in *actual* behavioral changes between waves (Fitzsimons and Moore 2008; Williams et al. 2006). This sort of "license to sin" effect has been observed among college students for questions like "Do you cheat on tests?" and "How often do you drink?" (Fitzsimons, Nunes, and Williams 2007), but there is little evidence that it exists within the broader population. For this reason, I tend to favor the explanation that I provide above.

Table 3.3 Pooled models, with interactions between treatment condition and survey spacing

	Answers "yes" in the follow-up interview to the question…							
	Have you ever been arrested?		Have you ever smoked pot?		Have you ever stolen anything?		Have you ever driven drunk?	
Follow-up group	β	(SE)	β	(SE)	β	(SE)	β	(SE)
Main effects								
Received the focal question at baseline	0.006	(0.019)	-0.014	(0.024)	0.004	(0.017)	0.007	(0.017)
One-month group	0.005	(0.020)	-0.026	(0.026)	0.018	(0.018)	-0.009	(0.019)
Interaction								
Received the focal question at baseline x one-month follow-up group	-0.014	(0.029)	0.044	(0.037)	0.049 *	(0.026)	0.048 *	(0.027)

Note: Model estimates represent percentage point changes in the probability of answering "yes" to survey items at the time of the follow-up interview. To reduce residual variance and increase the precision of the estimated coefficients, all models include the covariates listed in Table 1. The total sample size for this analysis was 5,307. Cluster-robust standard errors are shown in parentheses. See text for more details.

* $p < .10$; ** $p < .05$; *** $p < .01$

Robustness check

Because members of the KnowledgePanel sample often take many surveys over the course of multiple years, there is some concern that they may become "professional respondents" (Toepoel, Das, and van Soest 2008). Although I only sampled individuals who had not previously been asked the focal items on the survey, it is still possible that their prior experience responding to other KnowledgePanel surveys influenced their susceptibility to panel conditioning in my experiment.[35] In order to examine this possibility, I restricted the sample to individuals whose tenure in the panel was less than or equal to 6 months at the time of the baseline interview (n = 404 for the one-month group and n = 467 for the one-year group); I then re-estimated all of the models described above. The results, which I summarize in Table 3.4, suggest that my prior estimates may actually be conservative. In general, the effects that I observed earlier are *more* pronounced (and more highly significant) when the sample is restricted to *less* seasoned respondents.

Heterogenous treatment effects

Does the severity of panel conditioning depend on who is responding to the survey? To this point, my primary interest has been in identifying the average treatment effect (ATE) taken across all respondents in the sample. I have found significant (and theoretically-sensible)

[35] The mean tenure time for members of my sample was 38 months at the time of the baseline survey, with a standard deviation of 32. The longest tenured member of the sample had been a member of the KnowledgePanel for more than 11 years.

Table 3.4 Robustness check using “less seasoned” respondents

	Answers “yes” in the follow-up interview to the question…							
	Have you ever been arrested?		Have you ever smoked pot?		Have you ever stolen anything?		Have you ever driven drunk?	
Follow-up group	β	(SE)	β	(SE)	β	(SE)	β	(SE)
Received the focal question at baseline and during the **one-month** follow-up	-0.004	(0.045)	0.018	(0.061)	0.110 **	(0.054)	0.175 ***	(0.055)
Received the focal question at baseline and during the **one-year** follow-up	0.026	(0.044)	-0.043	(0.054)	-0.022	(0.055)	0.032	(0.058)

Note: Model estimates represent percentage point changes in the probability of answering “yes” to survey items at the time of the follow-up interview. To be included in this analysis, respondents had to have been in the KnowledgePanel sample for *6 months or less* at the time of the baseline survey. To reduce residual variance and increase the precision of the estimated coefficients, all models include the covariates listed in Table 1. The sample size for the one-month group was 404; the sample size for the one-year group was 467. Cluster-robust standard errors are shown in parentheses. See text for more details.

* $p < .10$; ** $p < .05$; *** $p < .01$

for particular items among individuals in the one-month follow-up group, but have said nothing about the stability of these effects across different subsets of respondents (i.e., *conditional* average treatment effects or CATEs). In this section, I present results from *exploratory analyses* examining whether the effects of panel conditioning vary according to respondents' socioeconomic and/or demographic characteristics.[36] The characteristics that I examine are age (coded continuously), gender (1 = female; 0 = male), and family income (coded continuously using midpoints).[37]

To obtain estimates of treatment effect heterogeneity, I entered interactions between these variables and the treatment indicators, and then re-estimated the models using respondents in the one-month group.[38] The key results are shown in Figures 3.2 through 3.4.[39] Across questions and interactions, I find a total of three significant effects—one involving age (older people tend to report *higher* rates of marijuana use if they were

---

[36] My purpose in conducting these analyses is simply to illustrate that treatment effect heterogeneity *is* possible. To more definitively establish how panel conditioning effects vary across different members of the population, I would need a larger sample and a more sophisticated methodological approach. I return to this idea briefly in the next chapter.

[37] To avoid data dredging, I did not fit alternative models with other treatment-by-covariate interactions (e.g., treatment status x family size or treatment status x urbanicity). I only considered interactions with age, gender, and income.

[38] There was no treatment effect heterogeneity for the one-year group, presumably because there were no treatment effects after the longer layoff.

[39] Because there is no reason to assume that the CATEs for the continuous variables are linear, I also experimented with generalized additive models (GAMs) when fitting treatment interactions with age and income (Hastie and Tibshirani 1987). In general, these models did not yield significant improvements with respect to model fit (according to ANOVA tests), and typically produced results that looked approximately linear. For these reasons, I elected to stick with a simpler parametric specification.

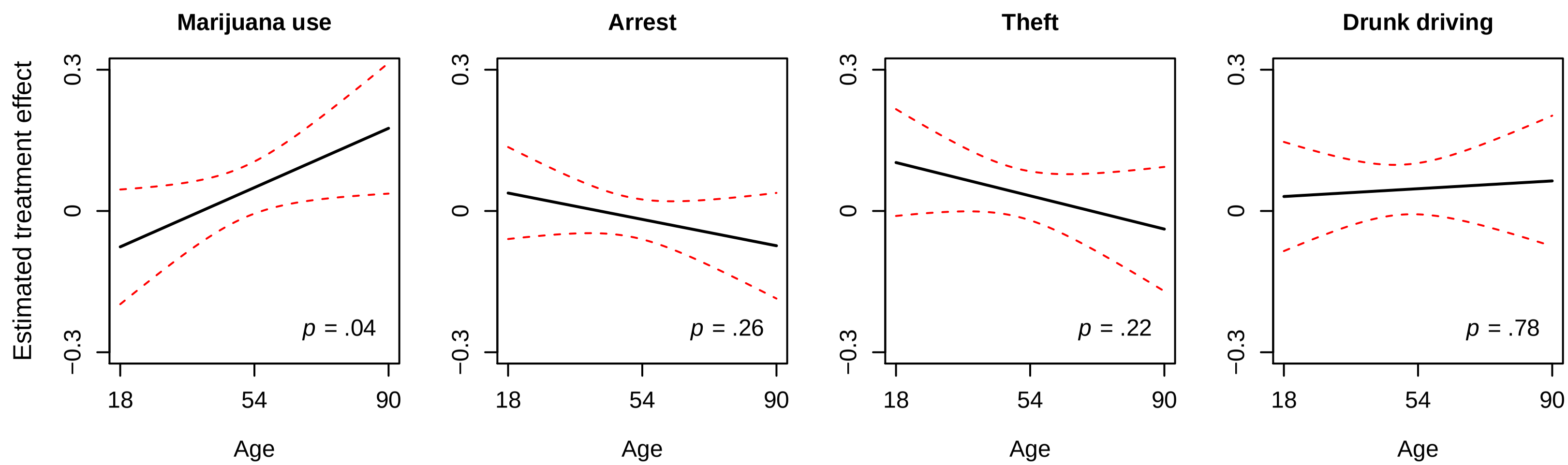

Figure 3.2: Treatment effect interactions with age. Each plot provides results for a separate question (e.g., "Have you ever smoked marijuana?"), as indicated in the heading. The black line provides an estimate of the treatment effect (e.g., receiving the marijuana question at baseline) by age; the red lines give 95% confidence intervals. The *p*-values for each interaction are displayed in the lower right-hand corner. See text for more details.

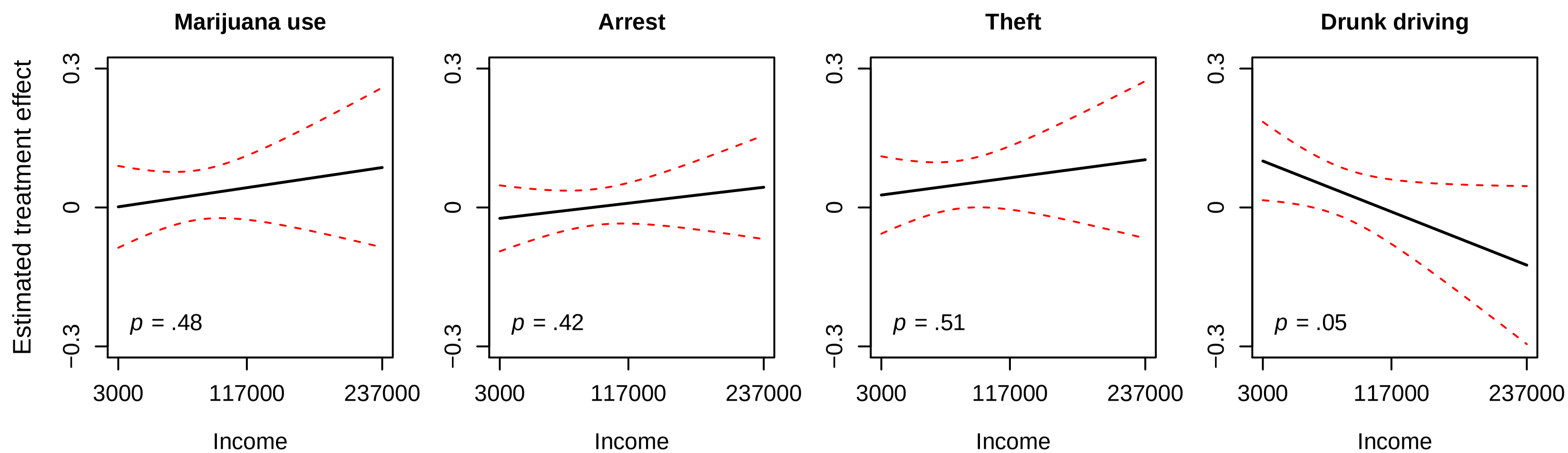

Figure 3.3: Treatment effect interactions with income. Each plot provides results for a separate question (e.g., "Have you ever smoked marijuana?"), as indicated in the heading. The black line provides an estimate of the treatment effect (e.g., receiving he marijuana question at baseline) by age; the red lines give 95% confidence intervals. The p-values for each interaction are displayed in the lower left-hand corner. See text for more details.

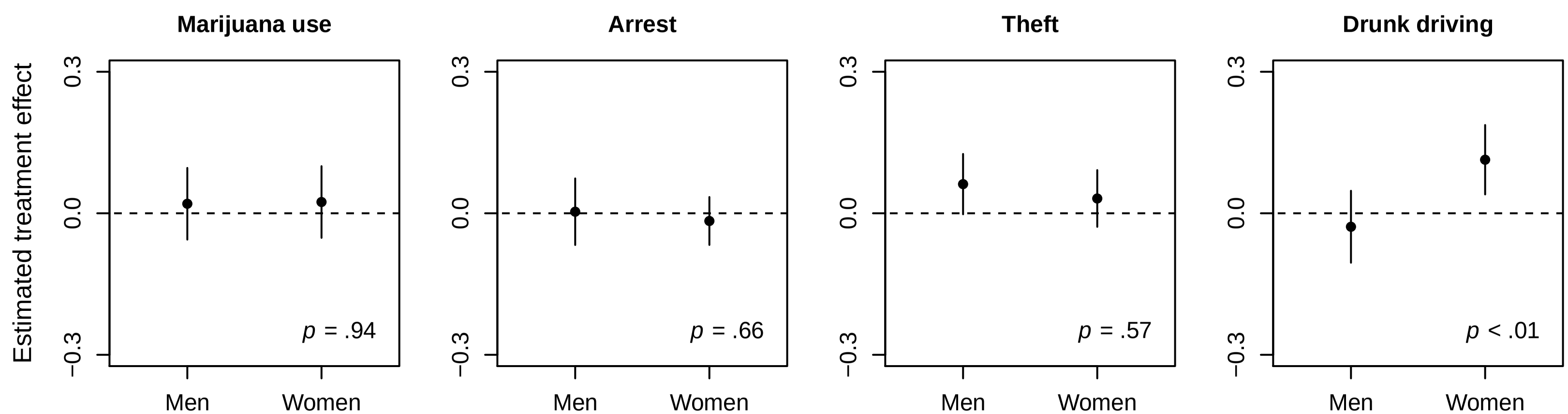

Figure 3.4: Treatment effect interactions with gender. Each plot provides results for a separate question (e.g., "Have you ever smoked marijuana?"), as indicated in the heading. The black circles provide point estimates for men and women, and the attached segments give the 95% confidence intervals. The p-values for each interaction are displayed in the lower right-hand corner. See text for more details.

previously exposed to the marijuana question), one involving gender (women are *more* likely to have driven drunk if they received the drunk driving question at baseline), and one involving income (wealthier individuals are *less* likely to admit to driving drunk if they had prior experience with that item). Although I cannot say for sure what is driving this heterogeneity, it could reflect the operation of different mechanisms within different subgroups of the population.

The income interaction for the drunk driving question provides a useful case in point. Although I found little empirical support for the social stigma hypothesis when examining results for the full sample (i.e., the likelihood of admitting to drunk driving did not decline if the respondent received the drunk driving question at baseline), it is possible (and perhaps even likely) that the degree of stigma associated with such a behavior varies according to social position. If high-income earners are more prone to embarrassment or cognitive dissonance after admitting to drunk driving (because it represents a greater departure from what they view as normative or appropriate), then the negative interaction shown in Figure 3.2 makes good theoretical sense. Different mechanisms are operating for different respondents, producing different kinds of effects across income groups.

Discussion

Economists, cognitive psychologists, and (at least some) sociologists have come to recognize that the act of responding to survey questions can affect the attitudes, behaviors, or statuses being measured and/or the quality of respondents' reports about those things. In this chapter, I considered various hypotheses about *why* these changes occur and *when* they are most likely to take place. Using an experimental design and a nationally-representative sample, I was able to show that panel conditioning tends to be a problem when (1) surveys are separated by shorter amounts of time (e.g., one month versus one year); and (2) the questions that are asked encourage respondents to think more carefully about their past experiences. In the remainder of the chapter, I provide some thoughts about what these findings mean from both a theoretical and applied perspective.

In previous work, I have described specific circumstances in which panel conditioning effects are theoretically most likely to occur. In the current study, I subjected some of these hypotheses to empirical testing by embedding a methodological experiment within a well-known web-based survey. My results suggest that *survey spacing* and *memory activation* play a key role in determining whether (and in what ways) respondents alter their reports from one survey wave to the next. After the baseline interview, treated respondents were *more likely* to say that they had previously driven drunk or stolen something of little value (behaviors that may have been difficult to recall during the initial interview), *but only when the lag between surveys was limited to one month.* Similar effects were not observed for arrests or marijuana use (events that may have been easier to recall for many respondents).

These findings suggest that some respondents may actually become better survey takers over time. What might look like increasingly risky behavior on the part of the individual (e.g., going from never having driven drunk to having driven drunk over the course of two interviews that are separated by only one month) may actually reflect their improved awareness of how they behaved in the past. Why this occurred when survey waves were separated by one month, and not one year, is not immediately clear. My suspicion is that the longer layoff between interviews counteracted any improvements that respondents were able to make with respect to memory, but this is only conjecture. Future work should investigate this issue further using a wider range of intervals between waves (e.g., one week, one month, six months, one year, and two years).

Future work should also investigate my findings concerning heterogeneity. If panel conditioning affects all respondents in the same way and to the same degree, then it will bias estimates of how frequently individuals engage in behaviors, but it will not affect our understanding of differences between groups. My results suggest that this scenario is unlikely to hold. Although I only considered a small subset of possible treatment-by-covariate interactions, I found significant variation across individuals both in terms of the severity and direction of panel conditioning effects. It seems safe to conclude that panel conditioning, like panel attrition, is an issue that may or may not bias conclusions in particular contexts or among particular subsets of the population. Indeed, the existence and magnitude of panel conditioning may vary across respondents within the same survey and substantive topic.

Ultimately, I believe that the best way to examine these issues is to field a larger-scale panel conditioning experiment. The basic research strategy that I employed here could easily be scaled up to accommodate a longer and more varied survey with bigger samples and additional follow-up waves. This would allow researchers to investigate a broader assortment of hypotheses concerning the nature and magnitude of panel conditioning effects, the sorts of surveys that produce them, and the kinds of respondents that seem to be the most susceptible. The costs of fielding such an experiment would likely be high (at least relative to the current study), but the potential payoff cannot be overstated. If we can convincingly demonstrate when and why panel conditioning occurs, it is not unrealistic to think that we may be able to correct for it in future longitudinal research.

Appendix

As described above, in the baseline survey I implemented a split ballot design in which respondents were asked a randomly selected two out of four survey questions, and in a randomly determined order. At follow-up, all four questions were asked of all respondents (again, in random order). The question wording and response options for each question appear below.[40] The first question was derived from the *National Survey on Drug Use and Health,* and the second and third questions were derived from the *1997 National Longitudinal Survey of Youth*.

> Question wording: "Not counting minor traffic violations, have you ever been arrested and booked for breaking the law? Being 'booked' means that you were taken into custody and processed by the police or by someone connected with the courts, even if you were then released."
>
> Response options: (1) Yes; (2) No; (3) Don't Know

> Question wording: "Have you ever used marijuana, for example: grass or pot, in your lifetime?"
>
> Response options: (1) Yes; (2) No; (3) Don't Know

---

[40] Respondents who answered "don't know" in the follow-up wave were coded to missing for that item. There were simply too few of them ($n < 10$ for each variable) to include in my analysis.

Question wording: "Have you ever stolen something from a store or something that did not belong to you worth less than 50 dollars?"

Response options: (1) Yes; (2) No; (3) Don't Know

Question wording: "Have you ever driven a vehicle while you were under the influence of alcohol?"

Response options: (1) Yes; (2) No; (3) Don't Know

Table B1. Models predicting the probability of attrition between interview occassions

	Model 1			Model 2		
	β		(SE)	β		(SE)
Treatment condition						
Ballot group 2 (ref: Ballot group 1)	-0.02		(0.02)	-0.02		(0.02)
Ballot group 3	0.01		(0.02)	0.01		(0.02)
Ballot group 4	-0.00		(0.02)	-0.01		(0.02)
Ballot group 5	-0.00		(0.02)	-0.01		(0.02)
Ballot group 6	0.01		(0.02)	0.01		(0.02)
Background characteristics						
Female				0.04	***	(0.01)
Age				-0.00	***	(0.00)
Black (ref = White)				0.07	***	(0.02)
Hispanic				0.07	***	(0.02)
Other race				0.02		(0.02)
High school graduate (ref: < high school degree)				-0.05	**	(0.02)
Some college				-0.03		(0.02)
BA or higher				-0.06	**	(0.02)
ln(Income)				-0.01		(0.01)
Household size				0.02	***	(0.00)
Widowed (ref: Married)				0.04		(0.03)
Divorced				0.04	*	(0.02)
Separated				0.00		(0.04)
Never married				-0.02		(0.02)
Living with partner				0.07	***	(0.02)
Living in metro area				-0.01		(0.01)
Head of household				-0.02		(0.02)
Renter (ref: Owns home)				0.03	**	(0.01)
Not employed (ref: Employed)				-0.01		(0.01)
Constant	0.31	***	(0.01)	0.54	***	(0.09)

Note: Ballot groups refer to the two questions received at baseline (e.g., the theft and the arrest question). Cluster-robust standard errors are in parentheses.

* $p < .10$; ** $p < .05$; *** $p < .01$

CHAPTER 4: A PROPOSAL FOR FUTURE RESEARCH

In the previous chapter, I carried out a small-scale panel conditioning experiment using data from a well-known internet panel. In this chapter, I describe a larger and more ambitious project that would involve new data collection and a lengthier survey instrument. The organization of the chapter adheres to the guidelines that are provided by the National Institutes of Health (NIH) for grants submitted under the R01 funding mechanism. The first section outlines my specific aims; the second section emphasizes the significance of the project; the third section highlights key innovations; and the fourth section provides information about my research design and analytic approach. I intend to submit a version of this chapter to NIH within the next year.

A. Specific aims

I propose to carry out a large-scale experiment on "panel conditioning," or the bias that is introduced when participating in one wave of a survey alters respondents' reports of their attitudes and/or behaviors in subsequent survey interviews. Prior research—conducted by myself and by researchers in public health, cognitive psychology, and political science—suggests that panel conditioning has the potential to substantially and negatively affect the quality of data from any number of widely-used surveys, including cornerstone resources like the Health and Retirement Study, the National Longitudinal Study of Adolescent Health, and the Panel Study of Income Dynamics. In this project, I will systematically investigate the incidence, magnitude, and nature of panel conditioning effects in an effort to determine *why they occur* and *what can be done to prevent them.*

Aim #1: *Test theoretically-motivated hypotheses concerning the nature and magnitude of panel conditioning effects. Determine whether the severity of panel conditioning depends on the substantive focus of the survey question, the manner in which it is administered, and/or the characteristics of the respondents.*

Aim #2: *Formulate a set of "best practices" for researchers who work with and/or collect longitudinal data. Identify specific settings in which researchers should be especially concerned about panel condition as a form of bias. Identify other settings in which the threat to inference is less extreme.*

If partaking in a longitudinal survey alters participants' responses to future attitudinal or behavioral questions—either because their actual attitudes or behaviors change or because the quality of their reports about their attitudes or behaviors changes—then researchers may be mischaracterizing the existence, magnitude, and correlates of changes across survey waves in respondents' characteristics. *Unfortunately, the evidence that I have produced to date suggests that this is exactly what is happening.* In my prior work, I have detected significant (and substantively large) panel conditioning effects in surveys that ask about unemployment (Halpern-Manners and Warren 2012); in surveys that elicit information about drug and alcohol use (Torche et al. 2012); and in surveys that pose questions about important social and political issues (see, e.g., Chapter 2).[41]

Researchers in the social and behavioral sciences should find these results extremely disconcerting. If panel conditioning affects such disparate topics as labor force participation and substance use, there is reason to think that the problem may be fairly widespread. To examine this possibility, I plan to test for panel conditioning effects across a wide range of substantive domains using a specially designed survey instrument. The topics that I will consider include health and nutrition (drug use, exercise, and diet); psychological well-being (happiness, affective state, and stress); sexual behavior (number of partners, use of contraceptives, and abortion); social participation (volunteering and friendship ties); and economic security (job loss, welfare receipt, and poverty). Results from this work should be of keen interest to researchers from a variety of academic and applied fields.

Identifying panel conditioning effects requires a strong analytic approach. In the present study, I propose to use a modified version of the Solomon four-group design. Respondents will be assigned at random to different treatment groups, with each treatment group receiving a slightly different survey ballot at baseline. All respondents will then receive the full battery of questions during their follow-up interview. By making comparisons across groups in the follow-up wave, I will be able to determine when panel conditioning is most likely to occur

[41] Respondents who are participating in the CPS for the first time, for example, have an unemployment rate that is around three-quarters of a percentage point higher, on average, than otherwise identical respondents with prior survey experience (Halpern-Manners and Warren 2012). This suggests that official estimates of the unemployment rate may be substantially biased.

and who is most likely to be affected by it. Again, this is no small matter for users of longitudinal data. Although researchers have carefully developed methods to deal with other problems associated with longitudinal research (e.g., panel attrition), much less attention has been paid to what may be an equally serious source of bias.

B. Significance

Researchers' awareness on panel conditioning dates back until at least 1940, when Lazarfeld (1940: 128) noted that "the big problem yet unsolved is whether repeated interviews are likely, in themselves, to influence a respondent's opinions." Since that time, scholars in several disciplines have sought to measure panel conditioning effects in what have often been disconnected lines of inquiry. In this section, I review and critique this body of research in an effort to demonstrate both the significance of the problem and the key contributions of the study that I am proposing to carry out.

(B.1) Prior research on panel conditioning and behaviors

The earliest attempts to address Lazarfeld's "big problem" focused not on opinions but instead on voting. Initially spurred on by Clausen's (1968) serendipitous finding that participation in a pre-cursor to the National Election Study (NES) increased voter turnout in the 1964 U.S. presidential election, political scientists and others have repeatedly shown that participation in the NES and other political opinion polls increases actual political participation in the United States and elsewhere (Anderson, Silver, and Abramson 1988;

Bartels 1999; Kraut and McConahay 1973; Traugott and Katosh 1979). In one study, Kraut and McConahay (1973) found that being interviewed about intended voting behaviors *more than doubled* respondents' probability of actually voting in that election—a finding that they validated using precinct records. In another study, Traugott and Katosh (1979) showed that surveying respondents about their candidate preferences led to higher rates of registration and voter turnout, with effect sizes increasing for each additional interview that the respondent completed.

Unfortunately, these behavioral changes do not appear to be limited to respondents' political participation. Among economists, for example, it has long been understood that the panel design of the Current Population Survey (CPS) has the consequence of leading to some degree of panel conditioning bias. Bailar (1989), Solon (1986), and Halpern-Manners and Warren (2012) have demonstrated that for any particular calendar month, unemployment rates for respondents in their initial month in the CPS rotation are *higher* than those for respondents who have been in the panel for longer. This issue has made its way into documentation about the design of the CPS (U.S. Bureau of Labor Statistics 2000), where Table 16-10 shows that in September 1995 the unemployment rate for CPS respondents in their first month in the sample was 8.6 percent higher than for CPS respondents as a whole in that month.[42] The implication is that CPS respondents are less likely to be unemployed, or

[42] Although CPS respondents are interviewed in person in some months and by phone in others, the differences reported in Table 16-10 are not due to mode effects. CPS respondents are interviewed in person in months 1 and 5

less willing to report unemployment, after their first of eight interviews.

Panel conditioning effects have also been observed in surveys that ask about respondents' health behaviors. Zwane et al. (2011), for example, showed that being interviewed about water quality and medical coverage increased the rate at which respondents used water treatment products *and purchased health insurance*. Battaglia, Zell, and Ching (1996) found evidence that participating in a survey about childhood vaccinations increased the chances that parents *subsequently immunized their children* against various viruses. And Levav and Fitzsimmons (2006), O'Sullivan et al. (2004), and Torche et al. (2012) found large panel conditioning effects on rates of *flossing, colorectal screening, and adolescents' substance use*, respectively. Taken together, these findings suggest that respondents' behaviors (or their ability and/or willingness to describe their behaviors accurately within the context of a longitudinal survey) may be influenced by the questions that they received in previous waves. Researchers who analyze longitudinal data should find this quite troubling.

(B.2) Prior research on panel conditioning and attitudes

Although methodologists were initially interested in the effects panel conditioning has on respondents' reported behaviors, researchers have also demonstrated effects for attitudinal measures (Sturgis et al. 2009). Analyses of the German Socio-Economic Panel (SOEP), for example, show that respondents are significantly more likely to select the top three scale

---

of their rotations, and differences between those two months are robust.

points to the life satisfaction question when being interviewed for the first time relative to those who had been asked the question in prior waves (Frick et al. 2006; Jürges 2005; Landua 1991). Frick et al. (2006) argue that this difference stems from respondents becoming more familiar with their satisfaction by reflecting on the question after completing the initial interview. Similar effects have been found in recent waves of the General Social Survey (GSS) for questions dealing with social status (see the results presented in Chapter 2), and in studies that ask about marital satisfaction (Veroff et al. 1992) and the quality of respondents' relationships (Wilson and Kraft 1993).

(B.3) Critique of existing panel conditioning research

The findings reviewed above demonstrate that panel conditioning can occur, at least in some instances. But there are also many instances in which researchers have sought to identify panel conditioning effects but have not found them. Although many political scientists have demonstrated that participating in opinion polls can increase voter turnout, Mann (2005) has disputed the basis of these findings and Smith et al. (2003) found no such effects. In contrast to O'Sullivan et al.'s (2004) finding that panel conditioning affects rates of *colorectal* cancer screening, Sutton et al. (1994) found no effects on rates of *breast* cancer screening. And despite half a century of evidence that panel conditioning biases CPS-based estimates of the unemployment rate, there do not appear to be similar effects in the Survey of Income and Program Participation (McCormick, Butler, and Singh 1992; Pennell and Lepkowski 1992) or the National Medical Care Utilization and Expenditure Survey (Corder and Horvitz 1989). I contend that this confusing mix of results has to do with two things: (1) weak research designs; and (2) a lack of well-validated theoretical expectations for when panel conditioning effects will be large, small, or non-existent.

(B.3a) *Weaknesses of typical research designs*

Most researchers have proceeded by comparing survey responses from members of a longitudinal panel to those from members of an independent cross-sectional sample drawn from the same population. Details of the problems with this design have been carefully laid out elsewhere (Cantor 2008; Das et al. 2011; Halpern-Manners and Warren 2012; Holt 1989; Sturgis et al. 2009; Torche et al. 2012; Warren and Halpern-Manners 2012; Williams and Mallows 1970; Zwane et al. 2011). Most importantly, such a design risks confounding biases from panel *conditioning* with biases from panel *attrition* (and in some cases, mode effects). Whereas the fresh cross-sectional sample may be representative of the original target population, the panel sample may have suffered from non-random attrition over time. Efforts to adjust for the resulting selectivity within the panel sample (via post-stratification weighting and/or related techniques) may or may not resolve the problem.[43]

Other researchers have estimated the magnitude of panel conditioning effects using data from rotating and/or overlapping panel survey designs such as the CPS or SIPP (see, e.g., Bailar 1989; Halpern-Manners and Warren 2012; Pennell and Lepkowski 1992). Most of the

[43] Post-stratification weights account for observable differences between groups with different levels of prior survey experience, but the groups may still differ in terms of unobserved characteristics.

evidence for panel conditioning effects in the CPS, for example, is derived from a comparison of reported unemployment rates in a given calendar month by respondents in their first month in the panel to those in later months in the panel. Unless the sample weighting procedures fully account for panel attrition, such differences may also be partly attributable to selective attrition across CPS waves. Although I have demonstrated in my prior work that this problem can be minimized by comparing respondents who (1) remained in sample for the same amount of time (e.g., for two consecutive months) *and* (2) entered the panel in different waves (e.g., March and April), this technique has only rarely been applied (Halpern-Manners and Warren 2012).[44]

To my knowledge, only a small handful of studies have randomly assigned individuals to treatment (i.e., received the focal questions during the baseline interview) and control (i.e., did not receive the focal questions during the baseline interview) groups in order to estimate the effects associated with panel conditioning (Kraut and McConahay 1973; Spangenberg et al. 2012; Torche et al. 2012; Veroff et al. 1992; Zwane et al. 2011). All of these studies found substantial evidence of bias—but each focused on a limited set of outcomes (e.g., voting in a primary election, donating to certain charities, marital satisfaction, substance use, and health insurance take-up) and none used data from a nationally representative sample. *The study that I describe below will be the first to embed a rigorous panel conditioning experiment within*

[44] Note that this approach is only effective when there are *overlapping* panels (like the GSS or the CPS). It cannot be applied when there are *rotating* panels that are not interviewed concurrently (like the Medical Expenditure Panel Survey).

an omnibus survey of the sort that sociologists, economists, public health researchers, and others frequently use. This will allow me to produce information about panel conditioning effects that is both broadly relevant and statistically sound.

(B.3b) *Lack of well-validated hypotheses*

Another reason for mixed results may have to do with the atheoretical nature of most panel conditioning research. Various scholars have made propositions about the circumstances in which panel conditioning will occur and the form that it will take (see, e.g., Sturgis et al. 2009; Warren and Halpern-Manners 2012; Waterton and Lievesley 1989). Some of these propositions are grounded in theories about the cognitive processes that underlie attitude formation and change, and the relationship between attitudes and behaviors (Feldman and Lynch 1988); others emerge from common sense and field experience. Testing these propositions in a rigorous way would help to shed light on why panel conditioning seems to take place in some survey settings but not others. *Unfortunately, this is not something that has typically been done in existing panel conditioning research*. I believe that this inattention to theory testing has greatly limited our ability to anticipate when panel conditioning effects will and will not occur. *I seek to overcome this limitation by evaluating a series of theoretically-motivated hypotheses concerning the scope and severity of panel conditioning effects in longitudinal data; these hypotheses are described in detail below*.

(B.4) Theoretical background and hypotheses

Are there specific situations where we should expect to observe significant panel conditioning effects? Are there other situations where we should *not* expect to observe such effects? In this sub-section, I derive a series of hypotheses concerning the nature of panel conditioning and the circumstances that are most likely to produce it.[45]

(B.4a) *Real changes in attitudes as a result of panel conditioning*

Many researchers conceive of attitudes as "crystallized" or fixed in people's minds (Eagly and Chaiken 2005; Fazio 1989; Wilson and Hodges 1992). Schuman and Presser (1981:271) define "crystallized attitudes" as those dispositions "that exist independently of our measurement, and that when appropriately measured show high reliability." In contrast, social psychologists, political scientists, and survey methodologists view attitudes as varying a great deal in their degree of crystallization and stability (see, e.g., Krosnick 1989; Krosnick and Abelson 1992; Schwarz 2007; Sudman, Bradburn, and Schwartz 1996; Tourangeau et al. 2000; Wilson and Hodges 1992; Zaller and Feldman 1992). As Sturgis et al. (2009:4) note, "opinions are often weakly held, easily influenced, and founded on a rather thin informational base." The implication is that respondents' attitudes may actually be quite fluid and malleable, particularly when the topic is one about which they have given relatively little thought or are not particularly knowledgeable (Wilson and Kraft 1989).

With this latter conceptualization in mind, researchers in survey methodology and social

[45] I present (but do not formally test) similar hypotheses in Halpern-Manners and Warren (2012) and Warren and Halpern-Manners (2012).

psychology have shown that thinking about an attitude as a result of being interviewed has the potential to change that attitude (Millar and Tesser 1986; Sudman et al. 1996). Respondents who lack crystallized attitudes about a topic will nonetheless offer a response to a question about that attitude in a baseline survey, often by drawing on a general set of values or predispositions (Schwarz 2007; Sudman et al. 1996; Tourangeau et al. 2000). This experience may set in motion a series of thoughts or actions (e.g., introspection, discussions with others, and so on) that change that attitude by the time of the follow-up survey (Krosnick 1999). When asked about that attitude again the next time that they are interviewed, some respondents may answer differently than they did in the previous wave. The result would be *real* change over time in respondents' attitudes or beliefs.

> ***Hypothesis 1:*** *Real changes in an attitude will occur as the result of responding to questions about that attitude, especially when respondents' initial attitudes are less crystallized and when the issue at hand is salient for the respondent.*

I expect that these changes will occur more frequently when the question asks about a topic that is important or salient to the respondent (Tourangeau et al. 2000). If the topics that the survey asks about are trivial, irrelevant, or uninteresting, it seems unlikely that respondents will take the time to reflect on, reconsider, or discuss their attitudes after completing the initial interview.

(B.4b) *Real changes in behaviors as a result of panel conditioning*

To explain why participating in pre-election surveys increased respondents' chances of voting, Clausen (1968) speculated that the experience of being surveyed "stimulated" respondents' interest in the election, leading to higher turnout at the polls. In replicating and re-evaluating Clausen's (1968) findings, Kraut and McConahay (1973) developed two alternative hypotheses. First, their "alienation reduction" hypothesis suggests that the personal, one-on-one contact with the survey interviewer changed respondents' sense of alienation from the political process, reducing psychological barriers toward voting in many respondents' minds. Second, their "self-concept" hypothesis suggests that participating in the pre-election survey led many respondents to view themselves as more politically involved than they were prior to the pre-election survey; respondents who saw themselves as more politically involved were thus more likely to vote.

Although each of these hypotheses pertains to a very specific substantive topic (the likelihood that respondents will cast a vote in an upcoming election), their commonalities are useful for thinking more broadly about the conditions under which participating in a survey might alter respondents' subsequent behaviors. In their own way, each of Clausen's (1968) and Kraut and McConahay's (1973) proposed mechanisms imply that responding to survey questions about a behavior has the potential to increase respondents' willingness and/or ability to engage in that behaviors after the survey interview is over. Based on this and other research (see, e.g., Cantor 2008; Das et al. 2007; Fitzsimons and Moore 2008; Nancarrow and Cartwright 2007; Perkins et al. 2008; Toepoel, Das, and van Soest 2009; Tourangeau et al.

2000; Waterton and Lievesley 1989), I believe that there are two basic ways in which behavioral change due to panel conditioning could occur.

First, answering survey questions about a behavior may increase respondents' interest in, awareness of, or information about that behavior. Beyond Clausen's (1968) "stimulation" hypothesis—which focuses on the role of a survey in heightening respondents' interest in a behavior—it seems clear that the survey experience may simply provide respondents with *information* that informs their future actions. A survey of low-income individuals that focuses on their utilization of the State Children's Health Insurance Program (SCHIP), for example, may alert respondents to the fact that the SCHIP exists and that it is something for which they may be eligible. Likewise, a survey of individuals' end of life legal preparations may have the consequence of informing (or reminding) some respondents that having a living will and assigning durable power of attorney are things that they should probably consider doing. This suggests the following hypothesis:

> ***Hypothesis 2:*** *Changes in a behavior will occur as the result of responding to questions about that behavior, especially when the questions serve to increase respondents' interest in, awareness of, or information about that behavior. This effect will be more pronounced when respondents are inclined to view the behavior as having some positive utility.*

Some questions may cause respondents to become more aware of the benefits of engaging in certain behaviors (Nancarrow and Cartwright 2007; Toepoel et al. 2009). As noted in the hypothesis given above, I suspect that these changes will be more likely to occur when the behavior in question has some perceived utility to the respondent (Fitzsimons and Moore 2008). A survey of respondents' cocaine use, for example, may make individuals more aware of the risks of cocaine use and may make them better informed about the methods of using cocaine, but it is unlikely to motivate respondents to begin to use cocaine—unless they are already positively predisposed to doing so (Fitzsimons et al. 2007). In general, this stipulation suggests that the magnitude and direction of panel conditioning effects may vary across focal behaviors and may be heterogeneous across respondents with different utilities or latent attitudes toward those behaviors.

> ***Hypothesis 3:*** *Respondents' behaviors will (at least appear to) change over time when survey questions force them to provide socially non-normative answers or portray themselves in a way that does not fit with their desired self-image.*

Second, answering survey questions about a behavior may lead to critical self-evaluation by some respondents that, in turn, influences their chances of engaging in that behavior in the future. In many surveys, respondents are asked to provide information about potentially sensitive and/or stigmatized topics—like drug use or crime or promiscuity. Reporting engagement in these sorts of "socially undesirable" behaviors may elicit negative emotions such as guilt or fear, triggering a reflective process that could lead to an altered response in subsequent waves (Epstude and Roese 2008). The negative emotions caused by answering questions about alcohol consumption, for example, may lead heavy-drinking respondents to drink less or else to report inaccurately lower levels of alcohol consumption on follow-up surveys (see, e.g., Torche et al. 2012). In either case, apparent declines in levels of alcohol consumption across waves would represent a form of panel conditioning.

(B.4c) *Panel conditioning effects when attitudes and behaviors remain constant*

In the preceding sections, I developed hypotheses about the conditions under which participation in a baseline survey of attitudes or behaviors may actually change those attitudes or behaviors prior to a follow-up survey. Although there is evidence that attitudes and behaviors may actually change as a result of participating in a survey, there is also good reason to suppose that another form of panel conditioning is at least as pervasive: respondents' attitudes or behaviors may remain unchanged between waves of a longitudinal survey, but the act of answering questions in a baseline survey may affect the answers that respondents provide in follow-up surveys (as I noted briefly in the example above concerning alcohol consumption). In this section, I articulate three reasons why the quality of respondents' reports might vary as a function of their prior survey experience. The first has to do with the relationships that form over time between interviewers and respondents.

> ***Hypothesis 4:*** *Respondents' attitudes and behaviors will appear to change across survey waves as they become more comfortable with and trusting of the survey experience, interviewers, and the organizations collecting their personal information.*

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 (Willis, Sirken, and Nathan 1994). As respondents become more comfortable with the survey process and the organizations that are collecting their information, their level of suspicion may decreases and their confidence in the confidentiality of their responses may grow (Fowler 1995). One consequence of this dynamic is that respondents' tendency to provide guarded, safe, and/or socially desirable responses (Krosnick et al. 1996; McFadden et al. 2005) may diminish as they gain more survey-taking experience (Cantor 2008; Warren and Halpern-Manners 2012; Waterton and Lievesley 1989). If this occurs, it could lead to the appearance of change across waves even when respondents' underlying attributes remain stable.

The fact that one form of panel conditioning (i.e., stigma-induced change) leads to the opposite empirical expectation as another form of panel conditioning (i.e., trust-induced change) means that research on the nature and magnitude of panel conditioning needs to consider both the social context of the focal topic (i.e., whether the topic is socially stigmatized in the target population) and the social relations surrounding the interviewing process (i.e., whether respondents are become more or less trusting over time). Simply asking survey questions for which there *might* be socially stigmatizing responses may or may not induce panel conditioning, and the direction of that bias may be positive or negative. The magnitude and direction of those biases may be conditional on both the social meaning of the focal topic and the nature of the relationship between respondents and interviewers. I will return to these ideas latter on in my discussion of heterogeneous panel conditioning effects (see section D.2f).

Hypothesis 5: *Respondents' attitudes and behaviors will appear to change across survey waves as they learn to manipulate the survey instrument in order to minimize their burden.*

Respondents sometimes find surveys to be tedious, cognitively demanding, and/or undesirably lengthy. One way that they can speed up the process is by answering "trigger" questions in a way that prevents additional follow-up items (Cohen and Burt 1985; Duan et al. 2007; Halpern-Manners and Warren 2012; Hernandez et al. 1999; Wang et al. 2000). In my prior work, I have speculated that this may be one of the reasons why individuals in the CPS tend to switch their answers from "unemployed" to "employed" or "out of the labor force" after their initial month-in-sample (Halpern-Manners and Warren 2012). Rather than reporting that they are still without work (and then answering a series of follow-up questions about unemployment and their job seeking behaviors), respondents simply select a different response option the next time that they are interviewed. This again leads to the false impression of change over time when no change has actually occurred.

Hypothesis 6: *Apparent changes across survey waves in respondents' attitudes and behaviors will occur as respondents learn to provide more accurate and complete responses. These changes will be especially evident for questions that require detailed information or* *considerable thought.*

Answering questions accurately within the context of a survey often requires a certain amount of skill (Tourangeau et al. 2000). Questions about previous events (e.g., "Have you ever experienced an episode of back pain that lasted one week or more?") or concepts that are vaguely defined (e.g., "What industry do you work in?") can be challenging to answer—either because it is difficult to recall the relevant information (a memory issue) or because it is hard to know what constitutes a valid response (a reporting issue). As respondents become more familiar with the survey instrument, these problems may become less pronounced. Vaguely defined concepts could become easier to describe (because respondents have a better sense for the response options that are available to them) and people's ability to recall past events may increase (because they will have had additional time to consider the questions). This suggests that some respondents may actually become *better* survey takers over time.

(B.4d) *Panel conditioning effects and survey spacing*

Hypothesis 7: *Panel conditioning is more likely to occur when surveys are spaced more closely together in time.*

In different ways, the six hypotheses described above each involve mechanisms whose strength depends, in part, on the length of time between baseline and follow-up surveys. In each case, the longer the interval between waves, the more that intervening life events, subject maturation, historical events, forgetfulness, and other factors may overwhelm, counteract, or mute potential panel conditioning effects. Although I have not carried out a thorough meta-analysis, my reading of the literature is that when baseline and follow-up surveys are separated by a month or less, panel conditioning effects are usually observed. When baseline and follow-up surveys are separated by between a month and a year, the results are quite split: many studies find panel conditioning (see, e.g., Torche et al. 2012), but many do not (see, e.g., Underwood et al. 2006). Panel conditioning effects are rarely observed when baseline and follow-up surveys are separated by more than a year (but see Chapter 2).

(B.5) Summary

Previous assessments of panel conditioning have suffered from two major shortcomings: (1) they have generally used weak research designs; and (2) they have failed to demonstrate why panel conditioning occurs and who it is most likely to affect. In this proposal, I describe a new data collection effort that is designed to overcome these limitations. By fielding a specially-designed survey with embedded methodological experiments, I will be able to distinguish panel conditioning effects from other forms of bias *and* systematically test the theoretical hypotheses that I spelled out above. *The results of this project will vastly improve our understanding of panel conditioning as a methodological problem, and will allow me to produce practical guidelines for researchers who work with and/or collect longitudinal data*. In the remainder of this proposal, I highlight the key innovations of my study and provide a detailed overview of the survey that I intend to field.

C. Innovation

The study that I am proposing to carry out contains four key innovations:

1. I plan to test for panel conditioning effects using a wide-ranging, multi-topic survey. The breadth and variety of the questions that I ask (as described in section D.2) will allow me to pinpoint specific substantive areas where panel conditioning effects are especially likely (or unlikely) to occur. Researchers who work with longitudinal data should find this information extremely valuable.
2. I will test hypotheses concerning the nature, scope, and magnitude of panel conditioning effects. Scholars have proposed several theories for why panel conditioning exists, but these theories have not been tested in a systematic way (because the necessary data do not exist). By carefully tailoring the questions that I ask to specific theoretical propositions, I will be able to address this issue.
3. I will maximize internal *and* external validity by embedding a rigorous panel conditioning experiment within a nationally representative longitudinal survey. Most investigators have utilized research designs that limit their ability produce valid estimates of panel conditioning effects and/or generalizable findings. My research design (see Section D.1) will allow me to overcome both of these problems.
4. I will provide guidelines for researchers who work with longitudinal data. Although scholars have argued that panel conditioning represents a serious methodological problem, very few have said anything about how to avoid it. Identifying when panel

conditioning is most likely to occur and what causes it to happen will allow me to speak to this question. I see this as the next frontier in panel conditioning research.

D. Approach

In the following section, I (1) review the basic research design that I plan to use for the proposed analyses; (2) detail my strategy for addressing each specific aim; and (3) provide a timeline for the completion of my analyses and for the dissemination of the results.

(D.1) Basic research design

In the study that I am proposing to carry out, I will use a variant of the Solomon four-group design to measure panel conditioning effects. At baseline, a nationally-representative sample of respondents will be contacted and asked to participate in a short telephone interview.[46] The baseline survey will include four modules of questions (hereafter referred to as A, B, C, and D) and a standard set of core demographic items (e.g., age, gender, race/ethnicity, urbanicity). One-fourth of the sample will receive modules A, B, and C in their initial interview; one-fourth will receive modules A, B, and D; one-fourth will receive modules A, C, and D; and one-fourth will receive modules B, C, and D. Assignment to the modules will be randomized to ensure that the groups are balanced with respect to background characteristics, and the order with which the modules appear within the survey (e.g., A-B-C versus B-C-A) will be

[46] I will use pre-notification letters, small monetary incentives, and repeated phone calls to achieve maximal response rates (Dillman 1978).

determined by chance. All respondents will receive the core set of demographic items as a part of their baseline survey.

One month later, half of the sample will be re-contacted and asked to participate in a follow-up survey (which I will administer over the phone using identical procedures). In this wave of interviews, the survey instrument will include *all four modules* of questions (A, B, C, and D). This sets up a comparison—in the second wave—between respondents who do and do not have prior exposure to a given module (respondents who *do* have prior exposure can be thought of as the "treatment" group). In the absence of panel conditioning, I would expect to see an equivalent distribution of responses across all respondents—*regardless of which modules they received during their baseline interview*.[47] The answers that are given to questions in Module A, for example, should look the same regardless of whether or not the respondent's baseline survey included *that particular set of items* (since all respondents are drawn from the same population). If there are large and systematic deviations from this result, it would suggest that panel conditioning has in fact occurred.

To examine panel conditioning effects in settings where the time interval between waves is longer than one month, I plan to repeat this exercise after one year using respondents who were *not* re-interviewed in the first follow-up wave. The basic mechanics will be exactly the same: respondents will be contacted by telephone and asked to participate in a second survey

[47] This can be tested within the context of a simple regression model, where the key predictor is an indicator of whether or not the respondent received the focal question during the previous wave.

interview. Each respondent will receive the full battery of questions, including items from *all four modules*. If I observe differences between groups (e.g., differences between respondents who received Module B at baseline and respondents who only received Modules A, C, and D) with respect to the focal questions (e.g., Module B), I will conclude that panel conditioning occurred for that particular subset of items. In general, I expect the severity of panel conditioning to be greater among those whose follow-up survey took place *one month* after the baseline interview; I plan to test this hypothesis using the modeling strategy described below (in section D.2e).

(D.1a) *Issues related to identification*

Identifying panel conditioning effects is difficult, even in a well-controlled experimental setting. The main concern in my case has to do with isolating attrition. Although all of the respondents in the sample will have demonstrated the same propensity to persist in the study at the time of the follow-up interview (because they will all be participating in the survey for the second time), there could still be *differential attrition* across treatment groups (e.g., respondents who received Module C at baseline may be more likely to leave the study than respondents who received Modules A, B, and D due to the nature of the questions that are involved). If this occurs, it could lead to imbalances between groups with respect to observed and/or unobserved characteristics. To evaluate this possibility, I plan to model respondents' attrition status using indicators of which modules they received during their baseline interview. If the results show that attrition is orthogonal to module assignment, I will be able to more confidently attribute differences between groups to panel conditioning.

(D.1b) *Sample sizes and power analyses*

In order to properly design a study producing experimental data, one must know whether the scale of the study is large enough to produce effects that are detectable over experimental error (Feiveson 2002). For the purposes of this study, I propose to use the following sample sizes: 5,260 for the one-month follow-up group and 6,012 for the one-year follow-up group. If I assume that the rate of attrition between waves will be 20% for the one-month group and 30% for the one-year group, I will be left with an analytic sample of approximately 4,208 for each set of analysis. With this many cases (and with a 3:1 ratio between the treatment and control group for any given measure) I will be able to detect differences in proportions that are at least 10 percent, and differences in means that are greater than or equal to one-tenth of a standard deviation (assuming an α-level of 0.05 and a β-level of 0.80).[48] These effect sizes are generally in line with the estimates I have produced in my prior panel conditioning research (see, e.g., the preliminary results given in section D.1c).

(D.1c) *Preliminary analyses*

To evaluate the feasibility of the research design described in section D.1, I conducted a small-scale version of my experiment using GfK's KnowledgePanel internet survey. The

[48] To simplify these calculations, I am assuming that the proportions will be around 0.5. For variables where the distribution of responses is more highly skewed, I will be able to detect differences in proportions that are much smaller in magnitude.

KnowledgePanel is a probability-based, non-volunteer random sample of Americans ages 18 and older. Panel members are recruited using a statistically valid sampling method that is based on a published sampling frame of residential addresses that covers approximately 97% of U.S. households. The panel typically includes around 50,000 adults, although the actual size of the sample fluctuates over time due to voluntary withdrawal, involuntary retirement of longtime members, and the addition of new panelists from on-going recruitment. In order to carry out preliminary analyses, I selected a random subset of 7,654 members of the larger KnowledgePanel sample. I then implemented a simplified two-module version of the research design described above, where each module contained two questions about deviant and/or illicit behaviors (like marijuana use, theft, drunk driving, and arrest).

Among respondents who were re-interviewed after one month, I found large and statistically significant differences between those who received the "treatment" (i.e., respondents who answered the focal question at baseline) and those who did not. Individuals who were asked the baseline question about theft were 12% more likely to report ever having stolen something the second time they were interviewed; and individuals who were asked the baseline question about drunk driving were 11% more likely to subsequently report ever having driven drunk. *Similar effects were not observed among respondents who were randomly assigned to the one-year follow-up group*. For these individuals, the probability of answering "yes" or "no" to a particular item did not depend on whether they had previous exposure to that question. All of the parameter estimates tended toward zero and all were small relative to their standard error.[49]

(D.2) Addressing my specific aims

In this section, I provide *examples* of how I will test each of the seven hypotheses described in sections B.4a through B.4d, including *exemplary* items from each of the four modules that I intend to field. Space constraints prevent me from listing all of the survey items that would be useful for testing each of my hypotheses.

(D.2a) *Module A: Testing the attitude crystallization hypothesis*

Hypothesis 1 suggests that respondents in the treatment group who are presented with an attitudinal item concerning an important or salient topic about which they had no well-developed or crystallized attitude may have subsequently reflected on, thought about, or otherwise reconsidered the attitude they reported the first time they were presented with the question. In Module A, I will evaluate this hypothesis by asking a series of questions that deal with topics that are salient to respondents, but about which they are unlikely to have formed strong opinions prior to first encountering the questionnaire items. The following two questions serve as relevant examples:

> ***Question #1***: "If you had a serious illness TODAY with very low chances of survival, and were mentally intact but in severe and constant physical pain, what would you do?" Response

[49] In conducting these analyses, I found that module assignment was unrelated to attrition between waves. This is reassuring for my purposes, as it suggests that differential attrition may pose less of a concern.

options: (1) Continue all treatments in an effort to prolong life; (2) Stop all life-prolonging treatments; (3) Have some treatments but not others; (4) Depends on the situation; (5) Leave it up to family, doctor, or God; (6) Don't know; (7) Refused. *[Adapted from the Wisconsin Longitudinal Study]*

Question #2: "Today, tests are being developed that make it possible to detect serious genetic defects before a baby is born. But so far, it is impossible either to treat or to correct most of them. If [you/your partner] were pregnant, would you want [her] to have a test to find out if the baby has any serious genetic defects?" Response options: (1) Yes, I would want [her] to have a test; (2) No, I would not want [her] to have a test; (3) Don't know; (4) Refused. *[Adapted from the General Social Survey]*

If Hypothesis 1 is correct, then members of the treatment and control groups *should* differ in their responses to Questions 1 and 2, but they should *not* differ in their responses to attitudinal items that ask about well-known or highly-visible topics (e.g., premarital sex or abortion or gun control).

(D.2b) *Module B: Testing the behavioral adoption hypothesis*

Hypothesis 2 suggests that respondents in the treatment group who are presented with questions concerning behaviors with which they are unfamiliar, *but that may have positive utility for them,* will be more likely to report engaging in those behaviors in subsequent waves (again, as compared to members of a randomly selected control group who do not receive such items during their baseline interview). In Module B, I will evaluate this hypothesis by asking a series of questions that deal with behaviors that (1) are new to many respondents; and (2) carry real and tangible benefits with respect to their quality of life. The following two items provide useful examples:

> ***Question #3***: "Have you made any legal arrangements for someone to make decisions about your care or medical treatment if you become unable to make those decisions yourself? This is sometimes called a Durable Power of Attorney for Health Care." Response options: (1) Yes; (2) No; (3) Don't know. *[Adapted from the Health and Retirement Survey]*

> ***Question #4***: "Do you or anyone in your family have a Flexible Spending Account for health expenses? These accounts are offered by some employers to allow employees to set aside pre-tax dollars of their own money for their use throughout the year to reimburse themselves for their out-of-pocket expenses for health care." Response options: (1) Yes; (2) No; (3) Don't know. *[Adapted from the National Health Interview Survey]*

If Hypothesis 2 is correct, then members of the treatment and control groups *should* differ in their responses to Questions 3 and 4 at the time of their second interview, but they should *not* differ in their responses to questions that ask about behaviors that most people are already well aware of (e.g., exercise or employment or marriage).

(D.2c) *Module C: Testing the stigma and trust hypotheses*

Hypothesis 3 suggests that respondents in the treatment group who are presented with questions that force them to provide non-normative answers or answers that do not conform to their desired self-image will be less likely to offer such answers in subsequent surveys. In contrast, Hypothesis 4 suggests that members of the treatment group may actually be *more* likely to provide such answers, due to their familiarity with the interviewing organization and their increased confidence in the confidentiality of their responses. Both of these hypotheses can be tested by asking questions about deviant, non-normative, or potentially embarrassing topics. Questions 5 and 6 provide two illustrative examples:

> ***Question #5:*** "The last time you had sex, was it with someone you were in an on-going relationship with, or was it with someone else? By 'sex' we mean only vaginal, oral, or anal sex. <u>Response options</u>: (1) Yes, in a relationship; (2) No, not in a relationship; (3) Refused. *[Adapted from the General Social Survey]*

> ***Question #6***: "Have you ever had an eating disorder, a learning or emotional problem, or a mental condition that limited your ability to do school work, attend school regularly, or work at a job for pay?" <u>Response options</u>: (1) Yes; (2) No; (3) Refused. *[Adapted from the National Longitudinal Survey of Youth, 1997]*

If Hypothesis 3 is correct, then members of the treatment group should be *less* likely to provide non-normative answers to Questions 5 and 6. If Hypothesis 4 is correct, then members of the treatment group should be *more* likely to provide non-normative answers when asked these same questions (e.g., "No, I was not in a relationship" or "Yes, I do have an emotional problem that limits what I can do").

(D.2d) *Module D: Testing the burden and survey taking hypotheses*

Hypotheses 5 and 6 suggest that the answers respondents provide will change over time as they become more familiar with the survey instrument and with the expectations involved. These changes can manifest themselves in two ways: respondents can either become better survey takers (by learning what constitutes an accurate response and/or thinking more carefully about the answers that they provide) or they can become more strategic in their approach (by identifying skip patterns and avoiding stem questions). In Module D, I will test these hypotheses by (1) including response options that trigger a lengthy set of follow-up items; and (2) asking open-ended questions that require respondents to elaborate on their answers. Questions 7 and 8 serve as useful examples:

> ***Question #7:*** "From time to time, most people discuss important personal matters with other people. Looking back over the last year, who are the people with whom you discussed an important personal matter? Please give their first names or initials." [Allow the respondent to report up to 10 names.]
>
> > Follow-up item: FOR EACH NAME PROVIDED: "How old is [name]?" → "Is [name] Asian, Black, Hispanic, White, or something else?" → "As far as you know, what is

[name]'s highest level of education?" → "What is [name]'s religious preference? Is it Protestant, Catholic, Jewish, some other religion, or no religion?"

Question #8: "Do you agree or disagree? Homosexual couples should have the right to marry one another?" <u>Response options</u>: (1) Strongly agree; (2) Agree; (3) Neither agree nor disagree; (4) Disagree; (5) Strongly disagree. *[Adapted from the General Social Survey]*

Follow-up item: IF THE RESPONDENT SAYS THAT THEY AGREE/ DISAGREE: "Why is that—why do you [favor/oppose] gay marriage?"

Follow-up item: IF THE RESPONDENT NEITHER AGREES NOR DISAGREES: "Why is that—why do you have no opinion about gay marriage?"

PROBE EACH OPEN-ENDED RESPONSE FOR CLARITY.

If Hypothesis 5 is correct, experienced respondents (i.e., respondents who were given Module D at baseline) should list fewer confidants the second time they encounter the name generator contained in Question 7. If Hypothesis 6 is correct, respondents in the treatment group will be more prepared for the open-ended questions concerning gay marriage, will provide clearer answers, and will require fewer probes in the follow-up wave.

(D.2e) *Testing the survey spacing hypothesis*

My final hypothesis pertains to the amount of spacing between interview occasions. Prior work—including the preliminary analyses that I described in section D.1c—suggests that panel conditioning is more likely to occur when survey waves are separated by shorter periods of time. In my analysis, I plan to evaluate this hypothesis by modeling interactions between the treatment indicator (i.e., did the respondent receive the item in question during their baseline interview?) and an indicator of which follow-up group the respondent participated in (i.e., was the respondent in the one-month follow-up group or the one-year follow-up group?). If Hypothesis 7 is correct, I would expect to see significantly larger panel conditioning effects for members of the one-month group.

(D.2f) *Variation in panel conditioning effects across respondents*

Theory suggests that panel conditioning is unlikely to affect all respondents in the same way and to the same degree (see, e.g., the hypotheses described in section B.4b). Identifying who *is* affected, and how the effects vary, is essential if we are to develop effective techniques for avoiding this problem in future longitudinal research. *With this objective in mind, I plan to conduct a series of supplementary analyses that are capable of identifying particular socio-demographic subgroups for which panel conditioning may be especially problematic.* These analyses will draw on the methodology outlined by Green and Kern (2012), who describe an automated Bayesian procedure (Bayesian Additive Regression Trees) that can be used to analyze treatment effect heterogeneity in experimental settings such as mine.

(D.2g) *Formulating a set of best practices*

The research strategy outlined above will yield detailed information about panel conditioning effects and the survey settings that are most likely to produce them. In order to ensure that this information is disseminated as widely as possible, I plan to author a series of articles aimed at scholars in sociology, public health, economics, and elsewhere (see the dissemination plan outlined below). These articles will provide researchers with a sense for when panel conditioning is most likely to occur and how its effects might bias their analyses. These articles will also contain a set of easily-implemented strategies for detecting panel conditioning effects both during the data collection process and after the data are already in hand. These diagnostic tools should prove useful to organizations that collect longitudinal data (e.g., NORC and the Bureau of Labor Statistics) and to the researchers who use them.

(D.3) Timeline to completion and dissemination plan

Field operations for this project will commence in the fall of 2014, with the baseline interviews, and will continue through the fall of 2015. I anticipate presenting preliminary results in a conference paper at the August 2016 meetings of the American Sociological Association; and more final results in papers at the April 2017 meetings of the Population Association of America and the May 2017 of the American Association for Public Opinion Research.[50] After receiving feedback and making revisions, I will submit these papers to journals like the *Proceedings of the National Academies of Sciences,* the *American Sociological Review, Sociological Methodology,* the *American Journal of Public Health,* and *Public Opinion*

[50] In accordance with NIH's guidelines concerning data sharing, I plan to make all of my data publicly available through an online data repository (like ICPSR).

Quarterly. I anticipate making these submissions during the summer and fall of 2017.

REFERENCES

Anderson, Barbara A., Brian D. Silver, and Paul R. Abramson. 1988. “The Effects of Race of the Interviewer on Measures of Electoral Participation by Blacks in SRC National Election Studies.” *The Public Opinion Quarterly* 52:53-83.

Angrist, Joshua A. 2001. “Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice.” *Journal of Business and Economic Statistics* 19:2-28.

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

—. 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: Wiley.

Barber, Jennifer S., Heather H. Gatny, and Yasamin Kusunoki. 2012. “The Results of an Experiment: Effects of Intensive Longitudinal Data Collection on Pregnancy and Contraceptive Use.” Working Paper 12-781, Population Studies Center, University of Michigan.

Bartels, Larry M. 1999. “Panel Effects in the American National Election Studies.” *Political Analysis* 8:1-20.

Battaglia, Michael P., Elizabeth Zell, and Pam Ching. 1996. “Can Participating in a Panel Sample Introduce Bias into Trend Estimates?”: National Immunization Survey Working Paper. Washington, D.C.: National Cenrter for Health Statistics.

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.

Benjamini, Yoav, Abba M. Krieger, and Daniel Yekutieli. 2006. “Adaptive Linear Step-Up Procedures that Control the False Disovery Rate.” *Biometrika* 93:491-507.

Borle, Sharad, Utpal M. Dholakia, Siddharth S. Singh, and Robert A. Westbrook. 2007. "The Impact of Survey Participation on Subsequent Customer Behavior: An Empirical Investigation." *Marketing Science* 26:711-726.

Cantor, David. 2008. "A Review and Summary of Studies on Panel Conditioning." Pp. 123-138 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.

Chandon, Pierre, Vicki G. Morwitz, and Werner J. Reinartz. 2004. "The Short- and Long-Term Effects of Measuring Intent to Repurchase." *Journal of Consumer Research* 31:566-572.

Clausen, Aage R. 1968. "Response Validity: Vote Report." *The Public Opinion Quarterly* 32:588-606.

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

Cohen, Steven B. and Vicki L. Burt. 1985. "Data Collection Frequency Effect in the National Medical Care Expenditure Survey." *Journal of Economic and Social Measurement* 13:125-151.

Corder, Larry S. and Daniel G. Horvitz. 1989. "Panel Effects in the National Medical Care Utilization and Expenditure Survey." Pp. 304-318 in *Panel Surveys*, edited by D. Kasprzyk, G. J. Duncan, G. Kalton, and M. P. Singh. New York: Wiley.

Crespi, Leo P. 1948. " The Interview Effect in Polling." *Public Opinion Quarterly* 12:99-111.

Das, Marcel, Vera Toepoel, and Arthur van Soest. 2007. *Can I Use a Panel? Panel Conditioning and Attrition Bias in Panel Surveys*. CentER Discussion Paper Series No. 2007-56. Tilburg, Netherlands: Tilburg University CentER.

—. 2011. “Nonparametric Tests of Panel Conditioning and Attrition Bias in Panel Surveys.” *Sociological Methods & Research* 40:32-56.

De Amici, Donatella , Catherine Klersy, Felice Ramajoli, Loretta Brustia, and Pierluigi Politi. 2000. “Impact of the Hawthorne Effect in a Longitudinal Clinical Study: The Case of Anesthesia.” *Controlled Clinical Trials* 21:103-114.

Dennis, J. Michael. 2010. “KnowledgePanel Design Summary.” Knowledge Networks, Palo Alto.

Dholakia, Utpal M. 2010. “A Critical Review of Question-Behavior Effect Research.” *Review of Marketing Research* 7:147-199.

Dillman, Don A. 1978. *Mail and Telephone Surveys: The Total Design Method*. New York: John Wiley & Sons.

—. 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.

Eagly, Alice H. and Shelly Chaiken. 2005. “The Advantages of an Inclusive Definition of Attitude.” *Social Cognition* 25:582-602.

Epstude, Kai and Neal J. Roese. 2008. “The Functional Theory of Counterfactual Thinking.” *Personality and Social Psychology Review* 12:168-192.

Fazio, Russell H. 1989. “On the Power and Functionality of Attitudes: The Role of Attitude Accessibility.” in *Attitude Structure and Function*, edited by A. Pratkanis, S. Breckler, and A. Greenwald. Hillsdale: Erlbaum.

Feiveson, A. H. 2002. “Power by Simulation.” *The Stata Journal* 2:107-124.

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* 73:421-435.

Fitzsimons, Gavan J. and Sarah G. Moore. 2008. "Should We Ask Our Children about Sex, Drugs and Rock & Roll? Potentially Harmful Effects of Asking Questions about Risky Behaviors." *Journal of Consumer Psychology* 18:82-95.

Fitzsimons, Gavan J., Joseph C. Nunes, and Patti Williams. 2007. "License to Sin: The Liberating Role of Reporting Expectations." *Journal of Consumer Research* 34:22-31.

Fowler, Floyd J. Jr. 1995. *Improving Survey Questions: Design and Evaluation*. Thousand Oaks, CA: Sage Publications.

French, David P. and Stephen Sutton. 2010. "Reactivty of Measurement in Health Psychology: How Much of a Problem is It? What Can Be Done About It?" *British Journal of Health Psychology* 15:453-468.

Frick, Joachim R., Jan Goebel, Edna Schechtman, Gert G. Wagner, and Shlomo Yitzhaki. 2006. "Using Analysis of Gini (ANOGI) for Detecting Whether Two Subsamples Represent the Same Universe: The German Socio-Economic Panel Study (SOEP) Experience." *Sociological Methods & Research* 34:427-468.

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* 27:179-184.

Green, Donald P. and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76:491-511.

Halpern-Manners, Andrew and John Robert Warren. 2012. "Panel Conditioning in the Current Population Survey: Implications for Labor Force Statistics." *Demography* 49:1499-1519.

Harrison, Lana and Arthur Hughes. 1997. "The Validity of Self-Reported Drug Use: Improving Accuracy of Survey Estimates." Rockville: National Institutes of Health.

Hastie, Trevor and Robert Tibshirani. 1987. "Generalized Additive Models: Some Applications." *Journal of the American Statistical Association* 82:371-386.

Hernandez, Lyla M., Jane S. Durch, Dan G. Blazer, and Isabel V. Hoverman. 1999. *Gulf War Veterans: Measuring Health*. Committee on Measuring the Health of Gulf War Veterans, Division of Health Promotion and Disease Prevention Institute of Medicine. Washington, D.C.: National Academies Press.

Holt, D. 1989. "Panel Conditioning: Discussion." Pp. 340-347 in *Panel Surveys*, edited by D. Kasprzyk, G. J. Duncan, G. Kalton, and M. P. Singh. New York: Wiley.

Jürges, Hendrik. 2005. "First Reply Effects (aka Repeated Measurement Effects)." Presentation at the DIW Workshop, Methodology and Measurement of Subjective Variables, Berlin.

Kraut, Robert E. and John B. McConahay. 1973. "How Being Interviewed Affects Voting: An Experiment." *The Public Opinion Quarterly* 37:398-406.

Krosnick, Jon A. 1989. "Attitude Importance and Attitude Accessibility " *Personality and Social Psychology Bulletin* 15:295-306.

—. 1999. "Survey Research." *Annual Review of Psychology* 50:537-567.

Krosnick, Jon A. and Robert P. Abelson. 1992. "The Case For Measuring Attitude Strength in Surveys." Pp. 177-203 in *Questions about Questions: Inquiries Into the Cognitive Bases of Surveys*, edited by J. Tanur. New York: Russell Sage.

Krosnick, Jon A., Sowmya Narayan, and Wendy R. Smith. 1996. "Satisficing in Surveys: Initial Evidence." Pp. 29-44 in *Advances in Survey Research*, edited by M. Braverman and J. Slater. San Francisco: Jossey-Bass.

Landua, Detlef. 1991. "An Attempt to Classify Satisfaction Changes: Methodological and Content Aspects of a Longitudinal Problem." *Social Indicators Research* 26:221-241.

Lazarsfeld, Paul F. 1940. "'Panel' Studies." *The Public Opinion Quarterly* 4:122-128.

Mathiowetz, Nancy A. and Tamra J. Lair. 1994. "Getting Better? Changes or Errors in the Measurement of Functional Limitations." *Journal of Economic & Social Measurement* 20:237-262.

McFadden, Daniel L., Albert C. Bemmaor, Francis G. Caro, Jeff Dominitz, Byung-Hill Jun, Arthur Lewbel, Rosa L. Matzkin, Francesca Molinari, Norbert Schwarz, Robert J. Willis, and Joachim K. Winter. 2005. "Statistical Analysis of Choice Experiments and Surveys." *Marketing Letters* 16:183-196.

Millar, Murray G. and Abraham Tesser. 1986. "Thought-Induced Attitude-Change - The Effects of Schema Structure and Commitment." *Journal of Personality and Social Psychology* 51:259-269.

Nancarrow, Clive and Trixie Cartwright. 2007. "Online Access Panels and Tracking Research: The Conditioning Issue." *International Journal of Market Research* 49:573-594.

Nukulkij, Poom, Joe Hadfield, Stefan Subias, and Evan Lewis. 2007. "An Investigation of Panel Conditioning with Attitudes toward U.S. Foreign Policy." Accessed at http://www.knowledgenetworks.com/ganp/docs/aapor2007/Panel-Conditioning-AAPOR07.pdf on April 25, 2012.

Pearson, Robert W., Michael Ross, and Robyn M. Dawes. 1994. "Personal Recall and the Limits of Retrospective Questions in Surveys." Pp. 65-94 in *Questions About Questions: Inquiries Into the Cognitive Bases of Surveys*, edited by J. M. Tanur. New York: Russell Sage Foundation.

Pearson, Thomas A. and Teri A. Manolio. 2008. “How to Interpret a Genome-wide Association Study.” *JAMA* 299:1775-1877.

Pennell, Steven G. and James N. Lepkowski. 1992. “Panel Conditioning Effects in the Survey of Income and Program Participation.” Pp. 566-571 in *Proceedings of the Survey Research Methods Section of the American Statistical Association*.

Percy, Andrew, Siobhan McAlister, Kathryn Higgins, Patrick McCrystal, and Maeve Thornton. 2005. “Response Consistency in Young Adolescents' Drug Use Self-Reports: A Recanting Rate Analysis.” *Addiction* 100:189-196.

Perkins, Andrew, Ronn J. Smith, David E. Sprott, Eric R. Spangenberg, and David C. Knuff. 2008. “Unverstanding the Self-Prophecy Phenomenon.” *European Advances in Consumer Research* 8:462-467.

Porst, Rolf and Klaus Zeifang. 1987. “A Description of the German General Social Survey Test-Retest Study and a Report on the Stabilities of the Sociodemographic Variables.” *Sociological Methods & Research* 15:177-218.

Schaeffer, Nora Cate and Stanley Presser. 2003. “The Science of Asking Questions.” *Annual Review of Sociology* 29:65-88.

Schuman, Howard and Stanley Presser. 1981. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. New York: Academic Press.

Schwarz, Norbert. 2007. “Attitude Construction: Evaluation in Context.” *Social Cognition* 25:638-656.

Shack-Marquez, Janice. 1986. “Effects of Repeated Interviewing on Estimation of Labor-Force Status.” *Journal of Economic and Social Measurement* 14:379-398.

Sherman, Steven J. 1980. “On the Self-Erasing Nature of Errors of Prediction.” *Journal of Personality and Social Psychology* 39:211-221.

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

Solon, Gary. 1986. “Effects of Rotation Group Bias on Estimation of Unemployment.” *Journal of Business & Economic Statistics* 4:105-109.

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* 18:102-106.

Spangenberg, Eric R., David E. Sprott, David C. Knuff, Ronn J. Smith, Carl Obermiller, and Anthony G. Greenwald. 2012. “Process Evidence for the Question-Behavior Effect: Influencing Socially Normative Behaviors.” *Social Influence* 7:211-228.

Sturgis, Patrick, Nick Allum, and Ian Brunton-Smith. 2009. “Attitudes Over Time: The Psychology of Panel Conditioning ” Pp. 113-126 in *Methodology of Longitudinal Surveys*, edited by P. Lynn. New York: Wiley.

Sudman, Seymour, Norman M. Bradburn, and Norbert Schwartz. 1996. *Thinking about Answers: The Application of Cognitive Processes to Survey Methdology*. San Francisco: Jossey-Bass.

Toepoel, Vera, Marcel Das, and Arthur van Soest. 2008. “Effects of Design in Web Surveys: Comparing Trained and Fresh Respondents.” *Public Opinion Quarterly* 72:985-1007.

—. 2009. “Relating Question Type to Panel Conditioning: Comparing Trained and Fresh Respondents.” *Survey Research Methods* 2:73-80.

Toh, Rex S., Eunkyu Lee, and Michael Y. Hu. 2006. “Social Desirability Bias in Diary Panels is Evident in Panelists' Behavioral Frequency.” *Psychological Reports* 99:322-334.

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. 2000. “Remembering What Happened: Memory Errors and Survey Reports.” Pp. 29-48 in *The Science of Self-Report: Implications for Research and Practice*, edited by A. Stone, J. Turkkan, C. Bachrach, J. Jobe, H. Kurtzman, and V. Cain. Englewood Cliffs: Lawrence Erlbaum.

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

Traugott, Michael W. and John P. Katosh. 1979. “Response Validity in Surveys of Voting-Behavior.” *Public Opinion Quarterly* 43:359-377.

U.S. Bureau of Labor Statistics. 2000. *Current Population Survey: Design Methodology*. Technical Paper 63. Washington, D.C.: U.S. Bureau of the Census.

Underwood, Martin R., Suzanne Parsons, Sandra M. Eldridge, Anne E. Spencer, and Gene S. Feder. 2006. “Asking Older People About Fear of Falling Did Not Have a Negative Effect.” *Journal of Clinical Epidemiology* 59:629-634.

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.

Veroff, Joesph, Shirley Hatchett, and Elizabeth Douvan. 1992. “Consequences of Participating in a Longitudinal-Study of Marriage.” *Public Opinion Quarterly* 56:315-327.

Wang, Kevin, David Cantor, and Adam Safir. 2000. “Panel Conditioning in a Random Digit Dial Survey.” Pp. 822-827 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. 319-339 in *Panel Surveys*, edited by D. Kasprzyk, G. J. Duncan, G. Kalton, and M. P. Singh. New York: Wiley.

Williams, Patti, Lauren G. Block, and Gavan J. Fitzsimons. 2006. “Simply Asking Questions About Health Behaviors Increases Both Healthy and Unhealthy Behaviors.” *Social Influence* 1:117 - 127.

Williams, William H. and Colin L. Mallows. 1970. “Systematic Biases in Panel Surveys Due to Differential Nonresponse.” *Journal of the American Statistical Association* 65:1338-1349.

Willis, G., M. Sirken, and G. Nathan. 1994. *The Cognitive Aspects of Responses to Sensistive Survey Questions*. Working Paper Series, No. 9. Hyattsville, MD: National Center for Health Statistics.

Wilson, Timothy D. and Sara D. Hodges. 1992. “Attitudes as Temporary Constructions.” Pp. 37-66 in *The Construction of Social Judgements*, edited by L. L. Martin and A. Tesser. New York: Springer-Verlag.

Wilson, Timothy D. and Dolores Kraft. 1989. “The Disruptive Effects of Explaining Attitudes: The Moderating Effect of Knowledge About the Attitude.” *Journal of Experimental Social Psychology* 25:379-400.

—. 1993. “Why Do I Love Thee: Effects of Repeated Introspections about a Dating Relationship on Attitudes toward the Relationship.” *Personality and Social Psychology Bulletin* 19:409-418.

Yeager, David S., Jon A. Krosnick, LinChiat Chang, Harold S. Javitz, Matthew S. Levendusky, Alberto Simpser, and Rui Wang. 2011. "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples." *Public Opinion Quarterly* 75:709-747.

Zaller, John and Stanley Feldman. 1992. "A Simple Theory of the Survey Response: Answering Questions versus Revealing Preferences." *American Journal of Political Science* 36:579-616.

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* 108:1821-1826.