doi:10.1017/S0003055410000092

Are Survey Experiments Externally Valid?

JASON BARABAS and JENNIFER JERIT *Florida State University*

Researchers use survey experiments to establish causal effects in descriptively representative samples, but concerns remain regarding the strength of the stimuli and the lack of realism in experimental settings. We explore these issues by comparing three national survey experiments on Medicare and immigration with contemporaneous natural experiments on the same topics. The survey experiments reveal that providing information increases political knowledge and alters attitudes. In contrast, two real-world government announcements had no discernable effects, except among people who were exposed to the same facts publicized in the mass media. Even among this exposed subsample, treatment effects were smaller and sometimes pointed in the opposite direction. Methodologically, our results suggest the need for caution when extrapolating from survey experiments. Substantively, we find that many citizens are able to recall factual information appearing in the news but may not adjust their beliefs and opinions in response to this information.

Social science researchers seek to establish causal relationships that are generalizable—that is, they try to maximize internal and external validity. Survey experiments are becoming more popular among scholars because they seem to possess both properties. The random assignment of respondents to treatment and control conditions reveals whether one factor causes another, whereas the use of a representative sample allows generalization to the larger population.

However, even in nationally representative survey experiments, external validity may still be a concern if the treatments do not resemble the relevant phenomena in question or if the experimental setting exaggerates the effect of the stimulus. This article investigates a question that recent studies have raised (e.g., Gaines, Kuklinski, and Quirk 2007; Kinder 2007) but that has not been examined empirically: are the causal findings of survey experiments reliable predictors of how opinion changes in the wake of actual political events?

For researchers using survey experiments, the implicit assumption is that a significant treatment effect says something about the direction, if not the rough magnitude, of effects that might be expected to occur in the real world (Gaines, Kuklinski, and Quirk 2007, 5). Our study investigates this assumption. We do so by comparing the effects of two naturally occurring political events to contemporaneous survey experiments that delivered similar information to diverse samples of the American public. In the first study, we examine the effects of news coverage of the 2007 Medicare trust fund warning and compare it to a survey experiment that provided the same key facts about Medicare's financial status. In the second study, we analyze media coverage of a new citizenship test launched by the U.S. Citizenship and Immigration Services (USCIS) in the fall of 2008. Once again, we conducted a concurrent survey experiment that delivered similar information to an adult national sample. Taken together, the two studies offer a rare opportunity to compare the treatment effects from survey experiments with the effects of real-world political events.

This research is timely, if only because survey experiments have overturned much of the conventional wisdom on the nature of public opinion (e.g., Gibson 1998; Kinder and Sanders 1996; Krosnick and Schuman 1988; Prior and Lupia 2008; Sniderman and Piazza 1993; Zaller and Feldman 1992). Indeed, one scholar recently proclaimed, "survey experiments that integrate representative samples with the experimental control of questions represent the most valuable tool for gaining access to the processes that underlie opinion formation" (Lavine 2002, 242). With the expansion of Internet surveys and multi-investigator studies such as Time-sharing Experiments in the Social Sciences, researchers from a variety of subfields are now using survey experiments (Druckman et al. 2006). Thus, the number of scholars affected by this issue is large and growing.[1]

We draw two lessons from our study—one is primarily methodological, the other is more substantive. Methodologically, we show that survey experiments generate effects that are observable among particular subgroups, not necessarily the entire population. Insofar as researchers keep this point in mind, we believe that survey experiments can be a valuable tool

Jason Barabas is Associate Professor, Department of Political Science, Florida State University, 531 Bellamy Building, Tallahassee, FL 32306 (jason.barabas@fsu.edu).

Jennifer Jerit is Associate Professor, Department of Political Science, Florida State University, 531 Bellamy Building, Tallahassee, FL 32306 (jjerit@fsu.edu).

The authors appreciate the helpful comments they received from Scott Allard, Charles Barrilleaux, Bill Berry, John Bullock, Brian Gaines, Cindy Kam, Gary King, Jon Krosnick, Jim Kuklinski, Skip Lupia, Cherie Maestas, Rose McDermott, Betsy Sinclair, Mark Souva, John Transue, Mathieu Turgeon, Gabor Toka, Paul Quirk, Will Shadish, Joe Young, and participants in colloquia at Florida State University, the University of North Carolina at Chapel Hill, Texas A&M University, and Vanderbilt University. Earlier versions of this article were presented at the Visions in Methodology Conference at Ohio State University, and at annual meetings of the American Political Science Association, the Canadian Political Science Association, and the International Society of Political Psychology. Ben Gaskins provided valuable research assistance. Data, replication code, and an online Appendix are available at http://polisci.fsu.edu/people/faculty/index.htm.

[1] Indeed, scholars in other fields are engaged in an analogous line of inquiry (e.g., Benz and Meier 2008; Cook, Shadish, and Wong 2008; Levitt and List 2007).

for studying public opinion. Substantively, we find that people in the natural experiments do not integrate new information and adjust their political beliefs to the degree that they do in survey experiments. Thus, scholars might come to different conclusions about the nature of public opinion depending on the manner in which they study it.

VARIETIES OF EXPERIMENTS AND VALIDITY

Years ago, scholars rarely undertook social science experiments (Kinder and Palfrey 1993). Today, experimental research is conducted in many different ways (McDermott 2002): laboratory experiments, carried out in controlled settings with students or members of the local community as subjects (e.g., Grosser and Schram 2006; Iyengar and Kinder 1987; Kam, Wilking, and Zechmeister 2007; Mutz and Reeves 2005); field experiments, in which randomized treatments are delivered outside the lab (e.g., Arceneaux and Kolodny 2009; Gerber and Green 2000); and natural experiments, where analysts take advantage of variation in real-world phenomena (e.g., Huber and Arceneaux 2007; Lassen 2005; Mondak 1995).[2] Increasingly, public opinion researchers employ survey experiments that deliver randomized treatments in telephone or Internet polls. This approach is common in the area of citizen competence, where researchers use survey experiments to examine whether people are capable of learning (e.g., Kuklinski et al. 2000) and to analyze how the presentation of information alters beliefs or opinions (e.g., Berinsky 2007; Kuklinski et al. 2001; Lupia et al. n.d.; Sniderman and Theriault 2004; Turgeon 2009).

Because of their randomized treatments, experiments have the advantage of a high degree of internal validity. Survey experimenters, who typically rely on nationally representative adult samples, often claim the mantle of external validity as well (for discussion, see Kellstedt and Whitten 2009, 75). But the issue goes beyond the representativeness of the subjects. According to Shadish, Cook, and Campbell (2002), "External validity concerns inferences about the extent to which a causal relationship holds over variations in persons, settings, treatments, and outcomes" (83; also see Campbell and Stanley 1963). Thus, when scholars embed experiments in opinion surveys, they must consider whether the treatments themselves are externally valid (see Druckman 2004, 684–85 or Gilens 2002, 248). Kinder and Palfrey (1993, 27) worry that a study's findings may be "the product of an unrealistically powerful manipulation, one that rarely occurs in natural settings." For example, in a critique of the framing literature, Sniderman and Theriault (2004) argue that it is unrealistic to examine the effects of a single frame when citizens typically experience competing frames (also see Chong and Druckman 2007). To the extent that treatments in survey experiments are overly strong or atypical, the observed effects may not generalize beyond the particular study at hand.

A second and related issue is the manner in which treatments are received (i.e., the experimental setting). Kinder (2007, 157) is concerned that

> experimental results can always be questioned on their generalizability, and framing effects are no exception. The major worry in this respect is that framing experiments—like experiments in mass communication generally—typically obliterate the distinction between the supply of information, on the one hand, and its consumption, on the other. That is, experiments are normally carried out in such a way that virtually everyone receives the message. The typical experiment thereby avoids a major obstacle standing in the way of communication effects, namely, an inattentive audience, lost in the affairs of private life. By ensuring that frames reach their intended audiences, experiments may exaggerate their power.

Likewise, in their framework, Gaines, Kuklinski, and Quirk (2007, 16) introduce an "inflation parameter" in recognition of the fact that "the artificially clean environment of the survey question makes treatment easier to receive than in real life." Unlike the natural world, which contains competing messages and other distractions that make exposure to the treatment probabilistic, exposure is essentially forced in a survey experiment. This feature may also limit the extent to which the findings are generalizable, even with nationally representative samples.

EXPECTATIONS

Based on the preceding discussion, we expect that the treatment effects in our survey experiments will be larger than those observed among the public as a whole in the wake of actual political events on the same topic. In other words, the effects should be in the same direction but magnified. Yet, this comparison is only a starting point—one that we will refine momentarily. We begin with the contrast between survey experiment and general population because we believe that this is the comparison researchers have in mind when they interpret the results of a survey experiment. After all, why implement a survey experiment on a nationally representative adult sample—a costly and time-consuming endeavor—unless one intended to generalize to that population? As Gaines, Kuklinski, and Quirk (2007) observe, "If those in the treatment group differ, on average, from those in the control group, the researcher normally concludes that the treatment works in a politically significant way in the real world" (5).

However, individuals are affected by messages in the mass media to the degree that they are exposed to that

[2] Shadish, Cook, and Campbell (2002, 17) define a natural experiment as "a naturally occurring contrast between a treatment and a comparison condition," and they give the example of property values before and after an earthquake (also see Cook and Campbell 1979). Some scholars reject this distinction, arguing that a study either is a randomized experiment or is not (King, Keohane, and Verba 1994, 7, n1). Even though some controversy surrounds the term "natural experiment," we employ it because it is used by some methodologists (e.g., Kennedy 2003, 405; Wooldridge 2009, 453) and because the events described later in this article meet the criteria for a natural experiment as outlined in Robinson, McNulty, and Krasno (2009; also see Dunning 2008).

information (Zaller 1992). Thus, we expect that the treatment effects in the natural setting will be largest for those who were exposed to media coverage. Indeed, it is possible that the difference in effect sizes across the two contexts (i.e., survey and natural world) may disappear completely once we focus on people who were exposed to the stimulus in the natural setting. Such a pattern would provide guidance for interpreting the results of survey-based experiments. It would suggest that the typical survey experiment generates effects likely to be observed only among the highly attentive in the real world (also see Hovland 1959). We test these expectations in two separate studies on different political issues.

STUDY 1: MEDICARE TRUST FUND WARNING

Each spring in the United States, the board of trustees overseeing Medicare and Social Security releases its future funding estimates. Unlike past years, the 2007 report triggered a warning calling for legislation to be introduced because Medicare's finances fell below a specific threshold.[3] This major policy event focused national attention on Medicare in the spring of 2007. We leveraged this situation by combining a natural experiment—a comparison of public opinion before and after the trustees' 2007 report—with survey experiments that delivered information about Medicare's finances. The survey treatments were similar to the real-world government announcement. As a result, this design allows us to contrast treatment effects from Medicare information provided in the real world with the treatment effects from survey experiments providing the same key facts about Medicare.

Study 1 uses data from two separate survey firms, Polimetrix and Knowledge Networks (KN), which administered national opinion surveys in the spring of 2007.[4] The Polimetrix and KN studies both feature Internet-based survey experiments, as well as comparisons before and after the Medicare announcement (see the Appendix for details on the two surveys). Because several features of Studies 1 and 2 are the same, we provide a detailed discussion of the research design in the context of the Medicare announcement and then briefly review the design when we get to Study 2.

Overview of Research Design

Figure 1 summarizes the design, relying on the symbolic notation for experimental designs from Campbell and Stanley (1963; also see Shadish, Cook, and Campbell 2002). Between March 1 and 21, 2007, researchers at Polimetrix randomly assigned 346 individuals to a treatment group. This group received a *survey treatment* consisting of information about the relative financial status of Medicare and Social Security (a stimulus denoted X_1 in Figure 1). The survey treatment was designed to present the most pertinent facts expected to be announced by the government—in particular, information about the projected exhaustion dates of Medicare and Social Security.[5] An additional 163 respondents were randomly assigned to the control group that did not receive this information. Respondents in both groups answered questions about their knowledge and opinions regarding Medicare. These observations are represented by the O symbols in Figure 1. We determine the *effect of the survey treatment* by comparing treatment group responses (denoted O_1 in Figure 1) with control group responses (O_2).[6]

A similar design was used in a survey experiment administered by KN. From March 2 to March 10, 2007, a randomly assigned group of 201 adults received the same Medicare information stimulus used in the Polimetrix study (X_1) and were then asked a similar set of questions about Medicare. Their responses (denoted O_3) are compared with those of a control group (O_4), offering a second measure of the survey treatment effect. The control group ($n = 604$) is bigger than the treatment group because we combine several conditions from the larger study.[7]

The trustees overseeing Medicare and Social Security made their announcement on April 23, 2007. Media coverage of this announcement is the second stimulus, a *natural treatment*, given to individuals in the adult population from which both Polimetrix and KN samples were drawn. Following the government announcement, both organizations conducted postannouncement surveys asking the same questions that were posed in the first time period; these began in late April and continued through early May.[8] Polimetrix reinterviewed the subjects that were in the control group at time 1. This permits us to compare the preannouncement survey responses of the individuals not receiving the survey treatment, O_2, with the responses of these same individuals after the natural

[3] Provisions for a "funding warning" were included in the Medicare Modernization Act of 2003.

[4] Each survey has more than 1,000 adult respondents. However, we analyze a subset of the data that pertain directly to the survey versus natural experiment comparisons.

[5] The "exhaustion date" is the year in which either program will no longer be able to pay full benefits (see Jerit and Barabas 2006 for more). This study focuses on Medicare, but because the trustees often deliver information about each program in the same public announcement, our treatment condition provides the exhaustion dates for *both* Medicare and Social Security.

[6] The treatment group combines two conditions ($n = 195$ and $n = 151$). These conditions were *identical* at time 1 (i.e., respondents in both groups received the stimulus). Respondents in the two groups differed only in terms of the information received at time 2, which is not part of this study. For ease of presentation, we combine the two conditions into a single treatment group, but we obtain the same substantive results with separate comparisons.

[7] To increase the power of our statistical tests, we combine respondents from an untreated control condition with individuals from two partially treated conditions. This second group of respondents received information about *either Social Security or Medicare, but not both*. Baseline knowledge and beliefs in these partially treated conditions is statistically indistinguishable ($p > .10$, two-tailed) from responses in the untreated control group (see the Appendix for additional details).

[8] Both postannouncement surveys were launched three days after the Medicare report on April 26, 2007 and continued until May 16 for Polimetrix and May 3 for KN.

FIGURE 1. Research Design Overview

	Condition Description	Time 1 (t_1) March 2007	Government Announcement April 23, 2007	Time 2 (t_2) Late April–Early May 2007
Polimetrix	Survey Experiment Treatment Group	$R \quad X_1 \quad O_1$		
Polimetrix	Survey Exp. Control Group & Natural Experiment Group t_1 and t_2 Observations	$R \qquad O_2 \left\langle \begin{matrix} O_2^{low-\exp} \\ O_2^{hi-\exp} \end{matrix} \right.$	X_2	$O_5 \left\langle \begin{matrix} O_5^{low-\exp} \\ O_5^{hi-\exp} \end{matrix} \right.$
Knowledge Networks (KN)	Survey Experiment Treatment Group	$R \quad X_1 \quad O_3$		
Knowledge Networks (KN)	Survey Exp. Control Group & Natural Experiment Group t_1 Observations	$R \qquad O_4 \left\langle \begin{matrix} O_4^{low-\exp} \\ O_4^{hi-\exp} \end{matrix} \right.$		
Knowledge Networks (KN)	Natural Experiment Group t_2 Observations	R	X_2	$O_6 \left\langle \begin{matrix} O_6^{low-\exp} \\ O_6^{hi-\exp} \end{matrix} \right.$

Notation

R indicates random assignment to a group.

X indicates introduction of a stimulus: X_1 (information provided in the survey) or X_2 (information provided in the mass media).

O indicates survey observation.

Definition of Treatment Effects

Survey experiment treatment effect = $O_1 - O_2$ and $O_3 - O_4$

Natural experiment treatment effect for entire sample = $O_5 - O_2$ and $O_6 - O_4$

Natural experiment treatment effect among high media exposure group = $(O_5^{hi-\exp} - O_2^{hi-\exp}) - (O_5^{low-\exp} - O_2^{low-\exp})$ and $(O_6^{hi-\exp} - O_4^{hi-\exp}) - (O_6^{low-\exp} - O_4^{low-\exp})$

treatment—denoted O_5—to measure the *effect of the natural treatment*. KN did not reinterview individuals in the control group from time 1, but the firm did interview a new nationally representative sample after the Medicare announcement. The responses of this group—denoted in the figure as O_6—can be compared with the responses of the time 1 control group, O_4, for a second measurement of the effect of the natural treatment.

We assume that respondents who were subjected to the survey treatment (X_1) received the stimulus. We cannot make the same assumption when it comes to media coverage of the government's Medicare announcement, our natural treatment (X_2). Some individuals saw media coverage; others did not. The Polimetrix and KN surveys included questions that permit us to distinguish individuals who are likely to have seen the media coverage (a "high exposure" group) from those who are unlikely to have been exposed (a "low exposure" group). (This is why, in Figure 1, we disaggregate O_2, O_5, O_4, and O_6 into two components, denoting survey responses for the low exposure and high exposure groups separately.) Consider Polimetrix respondents in the natural experiment. By comparing the change in survey responses experienced by individuals in the high exposure group ($O_5^{hi-\exp} - O_2^{hi-\exp}$) with the same change by subjects in the low exposure group ($O_5^{low-\exp} - O_2^{low-\exp}$), we are able to measure the effect of the natural treatment on the subset of the population most likely to have

actually experienced the treatment. In the KN study, we measure treatment effects in the same manner [i.e., $(O_6^{hi-\exp} - O_4^{hi-\exp}) - (O_6^{low-\exp} - O_4^{low-\exp})$], providing another measure of the effect of the natural treatment on individuals most likely to have heard or seen media coverage of the Medicare announcement.[9]

Survey Experiment Treatment

The same stimulus appeared in the Polimetrix and KN studies. It was the following factual information on the fiscal situations confronting Medicare and Social Security:

> As you may know, the board of trustees overseeing the Medicare and Social Security programs regularly releases financial estimates. These estimates provide information about the condition of both programs over the next several decades and often are featured in the media. The following passage is from a news story on the financial estimates from the most recent report: [Farther down on the same screen] According to the trustees, the financial status of Medicare is particularly problematic. Due to the growing size of the elderly population, trust fund reserves will be exhausted in the year **2018**. Social Security also faces financial problems, but the trust fund for Social Security is projected to be exhausted in the year **2040**.

This information captured the true state of the world as of March 2007, according to the panel of experts overseeing the programs.[10] The treatment in our study is similar in length and level of detail to stimuli in past studies providing factual information in survey experiments (e.g., Berinsky 2007; Gilens 2001; Kuklinski et al. 2000). More important, it conveys the same essential facts as news reports of the Medicare announcement (i.e., $X_1 \approx X_2$).[11]

Measuring Medicare Knowledge and Opinions

The empirical analysis focuses on two outcome measures. The first is knowledge, determined by a question that asked, "According to news reports, both Social Security and Medicare are facing financial problems in the future. If Congress doesn't take any action, which of these two programs is expected to be the first to not have enough money to cover all benefits—Medicare or Social Security?" The answer choices provided in random order were "Medicare," "Social Security," or "Both programs will exhaust their funds within the same year." The correct answer at the time of the first study was unambiguously Medicare. According to widely cited funding estimates, Medicare's trust fund was projected to run out of funds in 2018, and Social Security's in 2040. The knowledge outcome question asks for the relative status of the programs, not specific dates, because a primary concern among policy makers is that Medicare, not Social Security, will be the first to face a funding dilemma.

The second measure is respondents' beliefs about Medicare's fiscal health. In the Polimetrix survey, this item read: "Which of the following four statements comes closest to your own view of the Medicare program." The answer choices were: "The program is in crisis," "The program has major problems, but is not in crisis," "The program has minor problems," or "The program has no problems." In contrast, the KN sample was asked: "How confident are you that the Medicare program will continue to provide benefits of at least equal value to the benefits received by retirees today?" The response options were "Very confident," "Somewhat confident," "Not too confident," or "Not at all confident." Both questions were intended to assess beliefs about Medicare's fiscal health, and this is the only time in which the outcome measures for Polimetrix and KN respondents are not identical. Questions like these appear regularly in surveys about Medicare and Social Security (e.g., Cook, Jacobs, and Kim 2010; Page 2000; Shaw and Mysiewicz 2004), and they tap into the underlying notion of whether a problem exists apart from what reforms, if any, an individual supports.[12]

Study 1 Empirical Results

How well do survey treatment effects correspond to what occurs in the real world? To answer that question, we must first characterize news coverage of the Medicare announcement.

The trustees' 2007 report triggered a funding warning, the first in the program's history. Nevertheless, the warning generated a moderate amount of media coverage, in part because of other competing news events such as the death of former Russian President Boris Yeltsin and ongoing coverage of the Bush Administration's firing of U.S. attorneys. Whatever the source of the editorial decisions on the newsworthiness

[9] Note that $(O_6^{hi-\exp} - O_4^{hi-\exp}) - (O_6^{low-\exp} - O_4^{low-\exp})$ is equivalent to $(O_6^{hi-\exp} - O_6^{low-\exp}) - (O_4^{hi-\exp} - O_4^{low-\exp})$ (Wooldridge 2009, 453–54). The distinction between the intent-to-treat (ITT) effect and the average-treatment-on-treated (ATT) effect is instructive. The ATT represents the treatment effects for those who were actually treated, whereas the ITT effect represents the difference between treatment and control, regardless of compliance. Our examination of the pre–post difference among all respondents in our natural experiment corresponds to the ITT effect, whereas the subgroup analyses estimate the ATT effect.

[10] Even though Medicare's financial difficulties are more severe and immediate (Marmor 2000; Oberlander 2003), most people are more concerned about Social Security (Gramlich 1998). Thus, the treatment counters what are likely to be mistaken views in the minds of many Americans. We treated respondents on factual information (as opposed to, say, a persuasive argument) because policy-relevant facts are important in the formation of people's beliefs and, ultimately, their policy preferences (Delli Carpini 2009; Gilens 2001).

[11] In the 2007 report, the exhaustion dates moved back (2019 for Medicare and 2041 for Social Security). The relative financial status of the two programs remained unchanged. The issue of treatment equivalence is related to construct validity, or "the match between study operations and the constructs used to describe those operations" (Shadish, Cook, and Campbell 2002, 72).

[12] The distribution of both variables was also highly skewed, with very few people selecting the "Minor problems/No problems" and "Very confident/Somewhat confident" response options. As a result, we use a median split to create a dichotomous measure of respondents' Medicare beliefs (see Druckman 2004, 667 or Druckman and Nelson 2003, 739 for similar procedures). See the Appendix for details regarding question order.

TABLE 1. Survey and Natural Experiment Effects: 2007 Medicare Announcement

	Survey Experiment			Natural Experiment: Before Media Coverage (t_1)			Natural Experiment: After Media Coverage (t_2)			Natural Experiment: Over Time $\Delta_{t2} - \Delta_{t1}$	
	Mean	(s.e.)	*n*	Mean	(s.e.)	*n*	Mean	(s.e.)	*n*	Mean	(s.e.)
Knowledge (Polimetrix)											
Treatment group	.61	(.03)	346	.42	(.07)	53	.49	(.07)	53	.08	(.10)
Control group	.47	(.04)	163	.49	(.05)	110	.44	(.05)	110	−.05	(.07)
Difference	.15	(.05)	509	−.08	(.08)	163	.05	(.08)	163	.13	(.09)
| Sig. test value |	3.11			.91			.65			1.50	
p value	.00			.18			.26			.07	
Beliefs (Polimetrix)											
Treatment group	.40	(.03)	344	.28	(.06)	53	.34	(.07)	53	.06	(.09)
Control group	.29	(.04)	163	.29	(.04)	110	.27	(.04)	110	−.02	(.06)
Difference	.11	(.04)	507	−.01	(.08)	163	.07	(.08)	163	.08	(.08)
| Sig. test value |	2.46			.10			.88			.96	
p value	.01			.54			.19			.17	
Knowledge (KN)											
Treatment group	.58	(.03)	201	.42	(.04)	154	.52	(.04)	151	.09	(.06)
Control group	.35	(.02)	604	.33	(.02)	450	.33	(.02)	455	.00	(.03)
Difference	.22	(.04)	805	.09	(.05)	604	.18	(.05)	606	.09	(.07)
| Sig. test value |	5.56			2.04			4.00			1.37	
p value	.00			.02			.00			.09	
Beliefs (KN)											
Treatment group	.26	(.03)	201	.27	(.04)	154	.23	(.03)	151	−.04	(.05)
Control group	.22	(.02)	602	.20	(.02)	448	.22	(.02)	453	.02	(.03)
Difference	.04	(.04)	803	.07	(.04)	602	.01	(.04)	604	−.06	(.05)
| Sig. test value |	1.20			1.92			.34			1.08	
p value	.11			.03			.37			.14	

Notes: Cell entries represent values on the outcome measures for the treatment and control groups, with standard errors in parentheses. The entries in gray shading highlight the survey experiment and natural experiment comparisons across each outcome. All variables are scaled on a zero-to-one interval so that the highest value of 1 represents knowing that Medicare will exhaust its trust fund before Social Security (knowledge), believing that Medicare is in a crisis (Beliefs-Polimetrix), or having no confidence that Medicare will continue to provide benefits of at least equal value to the benefits received by retirees today (Beliefs-Knowledge Networks). One-tailed *p* values are shown. Entries may not sum perfectly due to rounding. The significance tests are differences in proportions, except in the case of the over time difference-in-differences estimates, which are the marginal effects from the t_2*X*exposure interaction in probit models (see Wooldridge 2009, 450–55). The standard errors for the models with interactions have been clustered to account for repeated observations of the panel respondents.

of the Medicare event (Bennett 2006), 54 stories appeared in newspapers and television newscasts during a 5-week period surrounding the Medicare announcement.[13] Nearly all media coverage was concentrated in a one-week period beginning on the day of the announcement. Therefore, although the coverage was fairly broad in that it appeared in a wide range of news outlets (Barabas and Jerit 2009), the event did not generate multiple stories in the same news source over several weeks. The important point, from our perspective, is that nearly every story led with the Medicare trust fund warning and conveyed information about the relative financial condition of Medicare and Social Security, specifying that Medicare was going to run out of money sooner than Social Security.[14]

Treatment Effects in Survey and Natural Experiments

Table 1 shows the treatment effects from our two survey experiments and the natural experiment. Cell entries represent values on the knowledge and belief outcome measures across treatment and control

[13] The count would have been higher had we included instances in which the Associated Press story was reprinted in local and regional newspapers (approximately 50 stories). The content analysis was based on a search of the LexisNexis and NewsBank archives from April 9 to May 16, and it covered all major papers and television networks, and a variety of cable sources. Two coders analyzed the transcripts to identify stories containing the exhaustion date information. Intercoder agreement was assessed on a randomly selected subset (35%) of the data (kappa = .84).

[14] In a comparison with other events, the trustees' 2007 report received roughly the same amount of coverage as the March 2007 meeting of the Federal Reserve and an April 2007 report from the Commerce Department on the U.S. trade deficit. It received twice as much coverage as either that month's unemployment report or the quarterly estimates of the nation's gross domestic product (GDP) issued in February 2007. Compared with events from previous years, the trustees' 2007 report received about half as much coverage as the October 2004 report on Iraq's nuclear capabilities. These comparisons emerged from a search of five newspapers (*New York Times*, *Washington Post*, *USA Today*, *Seattle Times*, and *Chicago Sun Times*) on LexisNexis.

conditions (all scaled to the 0–1 interval); standard errors and *n*s appear in the adjacent columns. Beginning with the upper-left portion of the table, we report the results from the survey experiment administered by Polimetrix. Subjects in the treatment group received information on the exhaustion dates for Medicare and Social Security. Subjects in the control conditions received no information before answering the outcome measures. Both groups were asked which program, Medicare or Social Security, would be the first to be unable to cover all benefits. The survey experiment treatment effect is represented by $O_1 - O_2$ from Figure 1.

In the Polimetrix data, there is a statistically significant treatment effect, with a higher proportion of treated respondents providing the correct answer than control group respondents. Much as one would expect, when respondents are given the exhaustion date information, roughly 6 in 10 can later recall which program will run out of money first. The box in gray shading shows the 15 point difference in knowledge between the treatment and control conditions ($.61 - .47 = .15$; $| z | = 3.11$; $p < .01$).[15]

We look for evidence of a treatment effect in the natural experiment in two ways. We begin by examining the *undifferentiated* control group, which entails comparing the preannouncement survey responses of the 163 people not receiving the survey treatment, O_2, with the responses of these same individuals after the natural treatment (O_5). When we do this, we find little evidence of learning (results not shown).[16] Forty-seven percent of respondents provided a correct answer to the knowledge item before the Medicare announcement; only 45% were able to do so afterward. The change over time is insignificant ($| z | = .22$; $p \leq .59$).

Recall, though, that we can refine our analysis of the natural experiment by examining the change in survey responses experienced by people in the high exposure group ($O_5^{hi-\exp} - O_2^{hi-\exp}$) relative to those in the low exposure group ($O_5^{low-\exp} - O_2^{low-\exp}$). We distinguish levels of media exposure with a question that asked respondents which media source they used most.[17] Drawing on the media content analysis described previously, we created a term, *High Exposure*, that was scored as 1 if a respondent's news source mentioned the Social Security *and* Medicare dates (i.e., they were exposed to the exhaustion date information like the subjects in the survey experiment). All other responses were given a zero.[18]

The results of the subgroup analyses appear in the remaining portion of the top rows of Table 1, where we compare changes in knowledge for high and low exposure individuals in the natural experiment.[19] High exposure people experienced a .08 change in their level of knowledge across the two time points (going from .42 to .49), whereas low exposure people moved in the opposite direction (−.05). The difference between these two differences, known as a "difference-in-differences" estimate (see Wooldridge 2009, 451), is .13, and it is statistically significant ($p \leq .07$). Even though our analysis of the undifferentiated control group turned up little in the way of learning effects, significant effects emerge when we examine a subset of people in the natural world who are highly attentive and therefore are most likely to be exposed to information carried in the mass media. In this instance, then, the survey experiment is a reasonable guide to what learning might look like in the real world among those who are exposed to naturally occurring media treatments. The question is whether we see a similar pattern with our attitudinal measure.

The survey experiment revealed evidence of belief change when respondents received information on the fiscal status of the two programs. The control group mean is .29, whereas it is .40 in the treatment group (resulting in the .11 treatment effect reported in Table 1). Relative to the controls, individuals in the treatment condition were more likely to state that the program was in a crisis ($| z | = 2.46$; $p < .01$). In contrast, the evidence for belief change in the natural experiment is weaker, regardless of whether we examine the undifferentiated control group ($.29 - .29 = 0$; $| z | = .12$; $p < .45$), or compare high and low exposure people over time (the remaining portion of the Polimetrix belief entries in Table 1). Focusing on the comparison between high and low exposure respondents, the difference-in-differences estimate is .08, and it is statistically insignificant ($p < .17$).

The null results from the natural experiment imply that in the real world, beliefs were unaffected by the Medicare announcement. Caution is warranted, however, because the comparison groups are small ($n = 53$ and $n = 110$). Indeed, if one looks at the direction of change for high and low exposure individuals, the pattern in the natural experiment mimics that of the survey experiment (with high exposure individuals becoming more likely to state Medicare is in a crisis). This raises the question of whether there is sufficient power to establish statistical *in*significance in the natural experiment. Auxiliary power analyses show that the probability of detecting a significant belief effect in the natural experiment, given an effect of the size found

[15] This *p* value is significant in a two-tailed test, but we report one-tailed *p* values throughout the article because of our expectations of learning and directional belief change (Blalock 1979, 163).

[16] To simplify the tables, patterns for the undifferentiated control group appear only in the text.

[17] The question wording was "How have you been getting most of your information about current events?" If respondents replied television, they were asked which channel from a list of network and cable sources. If they replied newspapers, they were asked to indicate which one. In other words, we have information about the particular source respondents were using (e.g., *Los Angeles Times*, *Chicago Sun Times, Wall Street Journal,* FOX, CNN, CSPAN, NPR, CBS).

[18] To reiterate, the *High Exposure* variable is an individual-level measure of media exposure, one that indicates whether a respondent was using a news source that provided the exhaustion date information. Such a measure is superior to traditional media use variables, which only indicate that a person reports paying attention to the news in general.

[19] We refer to high and low exposure groups as "treatment" and "control," even though they are commonly thought of as observational "comparison" groups (see footnote 2).

in the survey experiment, is .54.[20] Standards vary, but researchers generally strive for a power level of .90 (see Cohen 1988). Thus, statistical power is a concern in the natural experiment.

In light of the potential power issues in the Polimetrix survey, it is useful to turn to our second data source. The KN survey was administered at roughly the same time as the Polimetrix survey and has the advantage of larger comparison groups. Unfortunately, the KN survey lacks a measure of media exposure, and so we use a person's level of education as a proxy (e.g., Abrajano 2005; Nadeau and Niemi 1995). Respondents with a college degree or above (25% of the sample) were treated as having "high media exposure," whereas those with less than a college degree were coded as "low media exposure." The results for the KN survey are shown in the bottom half of Table 1. Here we observe a statistically significant treatment effect for knowledge in the survey experiment ($.58 - .35 = .22$ after rounding; $| z | = 5.56$; $p < .01$). Like the pattern observed in the Polimetrix data, subjects who were treated with exhaustion date information were able to state which program would be the first to not provide full benefits.[21]

Moving on to the pattern in the natural experiment, there is little evidence of learning in the undifferentiated control group over time ($.35 - .38 = .03$; $| z | = .91$; $p < .18$). Once again, however, there is greater correspondence between the survey and natural experiments when we focus on high exposure respondents. The high exposure group in the KN survey increased their level of knowledge from .42 to .52, which translates into a .09 change (with rounding). The corresponding change for low exposure individuals was zero (i.e., their level of knowledge was the same [.33] at both time points). The difference-in-differences estimate is .09 ($p < .09$).

Despite the variation in how we operationalize media exposure, the results for belief change in the KN data follow the general pattern established in the Polimetrix data: a survey treatment effect that is not easily replicated in the natural experiment. Among respondents in the survey experiment, the treatment makes a person more likely to state that they are not confident Medicare will continue to provide benefits ($.26 - .22 = .04$; $| z | = 1.20$; $p \leq .11$). In the natural experiment, there is no change in beliefs when one examines the undifferentiated control group over time ($.22 - .22 = 0$; $| z | = .18$; $p < .43$). Similarly, we cannot reproduce the survey treatment effect among high exposure respondents—in fact, the highly exposed are *less* likely to voice concerns about Medicare's ability to provide benefits, leading to a difference-in-differences estimate of $-.06$ that is nearly significant in the "wrong" direction ($p \leq .14$; power = .38).[22]

Taken together, the analyses of the survey experiments suggest that when people are given factual information about the relative status of Medicare and Social Security, they exhibit higher levels of knowledge and adjust their beliefs about the fiscal health of the two programs. The picture that emerges from the natural experiment is more complicated. When it comes to the undifferentiated control group, there is no evidence that the public registered the kinds of changes we observed in the survey experiment. In fact, there are rather large differences between the size of the treatment effects in the survey experiment and the overall control group for both outcome measures. There is greater correspondence between the survey and natural experiments when we focus on the high exposure group in the natural experiment. Here, however, the correspondence is limited to the knowledge item. The highly exposed were able to recall the essential fact that appeared in real-world news stories about the trustees' 2007 report. But coverage of the Medicare announcement did not have a corresponding effect on respondents' beliefs. Thus, our two survey experiments seem to induce a wider range of effects—in this case, belief effects—than were observed in the natural world. We explore this idea further with one final set of analyses.

Beliefs are commonly thought of as the "building blocks" of attitudes (Eagly and Chaiken 1993, 103). Thus, we probed for additional belief effects in the survey experiment with a third outcome measure, one that gauges willingness to make Medicare more like private sector managed care. The question asks: "There is a proposal to make Medicare health insurance benefits more like a private health plan, such as a PPO (preferred provider organization) or an HMO (health maintenance organization)? How do you feel about this proposal?" The four choices were: "I strongly support it," "I support it somewhat," "I oppose it somewhat," or "I strongly oppose it." Our expectation, based on past research (e.g., Jacobs and Shapiro 1998; Jerit and Barabas 2006; Oberlander 2003; Page 2000), was that individuals who believed Medicare was in a state of crisis would be more likely to embrace large-scale policy change, such as privatization.

The key question is whether there is a difference in the belief–opinion relationship among respondents

[20] The power analyses assume a .11 belief change (the survey experiment effect) with sample sizes of 53 and 110 and standard deviations of .45 and.46. Power is evaluated using a one-tailed, alpha = .10 level for repeated observations on the panelists (the correlation between samples is .45). Here, we run the risk of making a Type II error (i.e., failing to identify a meaningful effect as significant, which is a "false negative"). Statisticians refer to the risk of a Type II error as beta (where power = 1 − beta). The higher the power, the lower the risk of a Type II error.

[21] The mean level of knowledge in the KN treatment group (.58) is similar to the Polimetrix group (.61), but the magnitude of the treatment effect varies (.15 vs. .22). This difference reflects the higher level of baseline knowledge among Polimetrix respondents. Effect sizes in both surveys are similar to findings in previous studies (e.g., Norris and Sanders 2003). It is notable, however, that across both surveys, recall of the information was hardly perfect, with roughly 40% of treated respondents unable to give a correct answer to the knowledge question.

[22] The calculations in Table 1 are based on unweighted data. We obtain similar results when we employ the survey weights provided by KN. The only notable exception is that the nine-point natural experiment learning effect has a p value of .13 rather than .09. This pattern is consistent with Brehm's (1993, 119–21) observation that weighting can attenuate effects. We discuss the representativeness of the samples in the conclusion.

in the treatment and control conditions of the survey experiment. To the extent that the survey experiment primed respondents to think of Medicare in a state of crisis, it also might have brought individual's crisis beliefs into alignment with their policy preferences—in effect, making their Medicare attitudes more constrained (Converse 1964). Such a pattern would be consistent with research showing that the determinants of a person's attitude can change, depending on the considerations that are most accessible (e.g., Zaller 1992, 80–84).

In the final analysis of Study 1, we explore whether crisis beliefs became more strongly associated with Medicare policy preferences for respondents in the treatment condition. We observe the expected pattern in both survey experiments, with crisis beliefs positively and significantly related to privatization opinions in the treatment conditions of the Polimetrix and KN surveys ($p < .01$ and $p < .09$, respectively).[23] These results suggest that when presented with new information, respondents adjusted their beliefs in a way that made sense in light of the stimulus. Their beliefs and policy preferences also became more closely linked.

In the natural experiment, we looked for changes in belief integration in the undifferentiated control group as well as in the subset of highly exposed respondents. In both instances, there was no increase in constraint following the Medicare announcement. In the Polimetrix and KN surveys, crisis beliefs were generally unrelated to Medicare reform preferences. The only exception occurred in KN among the high exposure respondents, where we observed a significant but *negative* relationship between beliefs and policy preferences (coeff. $= -.41$, s.e. $= .21$). We believe that this last finding underscores our basic point—that beliefs and opinions became integrated as a result of the survey experiment, but that it was difficult to observe the same degree of belief integration in the natural setting.

When it comes to the correspondence between our survey experiments and the natural experiment, the analyses in Study 1 suggest that the survey experiment would be a good guide to predicting how levels of knowledge might change in response to real-world information flows. In particular, the group of people most likely to be exposed to news about the trustees' report were affected by it and showed significant gains in knowledge. At the same time, however, survey experiments seem to induce a wider range of effects than were observed in the natural world. To determine whether we could replicate these patterns, we conducted a second study.

Study 2: New Citizenship Test

In our second study, we turn to data that we collected as part of the 2008 Cooperative Campaign Analysis Project (CCAP), an Internet survey administered by YouGov/Polimetrix (see the Appendix for additional details). These data have many of the same features as our original study—in particular, a survey experiment designed to coincide with a naturally occurring political event. However, Study 2 has the advantage of larger comparisons groups and, hence, greater statistical power. It also considers another political issue at a different moment in time.

On October 1, 2008, the USCIS rolled out a new version of the test that immigrants must take in order to become U.S. citizens. The redesign of the test, the first since 1986, followed years of criticism of the old test, which critics believed was too easy. Instead of focusing on civics facts and trivia about the United States, the new test probed immigrants' understanding of concepts and key moments in the country's history (e.g., the Cold War). At the crux of the debate surrounding this event was the content of the questions on the new test. Officials at the USCIS said the goal was to encourage immigrants to learn about the country's civic values, but critics said the new test required a more sophisticated understanding of the United States, and they worried that it might be too hard.

As was the case with Study 1, we conducted a content analysis, this time examining coverage of the new citizenship test. The event was covered in approximately 50 news stories across a variety of media outlets.[24] During this time, the mass media also were covering the 2008 presidential election and the federal government's $700-billion rescue package for the financial industry. Like the trustees' 2007 report, then, the citizenship test had to compete with several other newsworthy events and therefore received a moderate amount of news coverage. There was, however, one notable difference across the two issues. In the case of the trustees' 2007 report, nearly all news coverage was concentrated in the one-week period surrounding the Medicare announcement. In contrast, stories about the new citizenship test appeared throughout the entire 5-week content analysis period. Thus, even though the number of stories was similar across the two studies, coverage of the citizenship test was distributed over a longer time period.

Overview of Research Design

We estimate survey treatment effects by comparing the responses of individuals randomly assigned into treatment and control conditions ($n = 280$ and $n = 304$, respectively) in the September wave of the CCAP study. This corresponds to the survey experiment treatment effect estimated in Study 1 ($O_1 - O_2$ in Figure 1). Like the first study, the survey treatment was designed to present information that we expected to appear in media coverage about the citizenship test—namely, that there was going to be a new version of the test and that

[23] The p values are from ordered probit analyses that examine the relationship between beliefs and policy opinions in the treatment conditions (coeff. = .29, s.e. = .12 for Polimetrix and coeff. = .24, s.e. = .18 for KN). There was no corresponding belief–opinion relationship for control group respondents at time 1 (coeffs. $= -.01$ and $-.10$; s.e. = .20 and .11, respectively).

[24] The nature of the content analysis was similar to Study 1. We consulted the LexisNexis and ProQuest news archives for a 5-week period surrounding the news event (i.e., September 15, 2008 – October 20, 2008).

the content of the questions was going to change from a focus on civics facts to more abstract concepts.

The USCIS launched the new test on October 1, between the September and October waves of the CCAP study. Because this is a panel study, we can compare the responses of individuals in the control group in September (time 1) with the responses of these same people in the October wave (time 2). This quantity of interest corresponds to the O_2 vs. O_5 comparison from Study 1 (Figure 1).[25] We also have a series of media exposure questions that allow us to identify respondents who were using news sources that provided information about the new citizenship test. The question wording was "How have you been getting most of your information about current events?" Answer choices were television, newspapers, radio, the Internet, discussion, and other. Individuals were then asked, "Please provide the name of your most used media source. Try to be as specific as possible (i.e., provide the name of your television station, newspaper, radio station, website, etc.)." We used the open-ended responses to this second question and a detailed content analysis to determine whether a respondent's news source provided the correct information to the knowledge question. We created a term, *High Exposure*, that was scored as 1 if a respondent's news source mentioned that the USCIS was administering a new citizenship test. All other responses were given a zero.

Survey Experiment Treatment

The treatment in the survey experiment was designed to look like a news story. There was a title at the top of the screen that read, "New Tests Asks: What Does American Mean?" The stimulus also included a color photograph taken from an actual news story about the new citizenship test. The image showed 9 adults reading the oath of allegiance, with a caption beneath the picture ("New citizens taking the oath of allegiance."). Further down on the same screen was the following text:

> Federal authorities unveiled new questions immigrants will have to answer to become naturalized American citizens. The redesign of the test follows years of criticism. Immigration officials want to move away from civics to emphasize basics on the structure of government, American history, and geography.

As was the case with Study 1, the survey treatment presented the same basic information as media coverage about the new citizenship test.

Measuring Immigration Knowledge and Beliefs

Study 2 features two outcome measures. Knowledge is assessed with a question that asked, "The U.S. Citizenship and Immigration Services (USCIS) is in charge of the naturalization process and applications to become an American citizen. Did this government agency take any of the following actions recently: Design a new naturalization test for immigrants to become U.S. citizens?" Response options were "Yes" (the correct answer), "No," and "Don't Know." On the very next screen was an item that measured respondents' beliefs. The question read: "Now I'd like to ask you about immigration in recent years. How likely is it that recent immigration levels will take jobs away from people already here?" Answer choices were: "Extremely likely," "Very likely," "Somewhat likely," and "Not at all likely." Answer choices for both outcomes were dichotomized as in Study 1 (i.e., correct vs. all other responses and a median split on beliefs).

Study 2 Empirical Results

In Table 2, we report the treatment effects from the survey and natural experiments. Like the presentation in Table 1, cell entries represent values on the knowledge and belief outcome measures across treatment and control conditions (scaled to the 0–1 interval), with standard errors and *n*s appearing in the remaining columns.

Beginning with knowledge, there is a 15-point treatment effect in the survey experiment ($.40 - .25 = .15$; $|z| = 3.88$; $p < .01$). Respondents who were exposed to the stimulus had a greater chance of providing the correct answer to the knowledge question. At the same time, knowledge among treated subjects did not reach the levels observed in Study 1. The baseline level of knowledge (.25) also was lower, suggesting that immigration was a more difficult issue.

We again look for treatment effects in the natural world in two ways: first by considering the undifferentiated control group over time, and then by comparing high and low exposure respondents in the natural experiment. When we examine the overall control group over time, there is little evidence of learning: 25% of the controls were able to provide the correct answer to the question at time 1, whereas 26% were able to do so in the second wave ($|z| = .39$; $p < .35$; not shown in Table 2). The comparison between high and low exposure respondents appears in the remaining portion of the top rows of Table 2. High exposure individuals experienced a .10 change in their level of knowledge across the two time points (going from .27 to .37), whereas low exposure people moved in the opposite direction (−.01). The difference between these two differences is .11, and it is statistically significant ($p < .05$). Consistent with the results from Study 1, there is correspondence between the survey and natural experiments when it comes to learning effects. Like the first study, however, significant learning effects emerge only when we examine the subgroup of people in the natural world who were most likely to be exposed to information about the new citizenship test.

When it comes to beliefs, the treatment in the survey experiment made respondents less likely to state that

[25] Although the CCAP study is based on a panel design, new cases were added at each wave. This is why, in Table 2, the *n* for our time 2 measures is larger than the *n* at time 1.

TABLE 2. Survey and Natural Experiment Effects: 2008 Immigration Announcement

	Survey Experiment			Natural Experiment: Before Media Coverage (t_1)			Natural Experiment: After Media Coverage (t_2)			Natural Experiment: Over Time $\Delta_{t2} - \Delta_{t1}$	
	Mean	(s.e.)	*n*	Mean	(s.e.)	*n*	Mean	(s.e.)	*n*	Mean	(s.e.)
Knowledge											
Treatment group	.40	(.03)	280	.27	(.05)	82	.37	(.05)	98	.10	(.07)
Control group	.25	(.02)	304	.24	(.03)	222	.23	(.02)	357	−.01	(.04)
Difference	.15	(.04)	584	.03	(.06)	304	.14	(.05)	455	.11	(.07)
| Sig. test value |	3.88			.53			2.75			1.66	
p value	.00			.30			.00			.05	
Beliefs											
Treatment group	.23	(.03)	280	.20	(.04)	82	.17	(.04)	98	−.02	(.06)
Control group	.29	(.03)	304	.32	(.03)	222	.32	(.02)	357	−.01	(.04)
Difference	−.06	(.04)	584	−.13	(.05)	304	−.14	(.05)	455	−.02	(.06)
| Sig. test value |	1.68			2.20			2.78			.35	
p value	.05			.02			.00			.36	

Notes: Cell entries represent values on the outcome measures for the treatment and control groups, with standard errors in parentheses. The entries in gray shading highlight the survey experiment and natural experiment comparisons across each outcome. All variables are scaled on a zero-to-one interval so that the highest value of 1 represents knowing that the U.S. Citizenship and Immigration Service (USCIS) developed a new test for citizenship or believing that immigrants are taking away jobs. One-tailed *p* values are shown. Entries may not sum perfectly due to rounding. The significance tests are differences in proportions, except in the case of the over time difference-in-differences estimates, which are the marginal effects from the $t_2 X$exposure interaction in probit models (see Wooldridge 2009, 450–55). The standard errors for the models with interactions have been clustered to account for repeated observations of the panel respondents.

immigrants take jobs away from people (.23 vs. .29 for a −.06 treatment effect; $p \leq .05$). This is to be expected because the stimulus presented a positive view of immigrants (e.g., people taking the oath of allegiance, an American flag in the foreground). Having just been exposed to an image that highlights legal (as opposed to illegal) immigrants, respondents in the treatment group were less likely to voice concern about job loss.

Consistent with the pattern in Study 1, though, there was little evidence that real-world media coverage of the new citizenship test altered people's beliefs. In the overall control group, beliefs remained constant across the time periods at .29 ($|z| = .11; p < .54$). Even among highly exposed respondents, the change in beliefs was slight (.20–.17), with no corresponding movement for the unexposed. This results in a difference-in-differences estimate of −.02, which is statistically insignificant ($p \leq .36$).

Even with the larger comparison groups in the CCAP survey, statistical power may still be a concern given the size of the treatment effect in the survey experiment (.06). Auxiliary analyses indicate that the power of our natural experiment to detect changes of the magnitude found in the survey experiment is low (power = .40). Fortunately, the fact that we conducted three independent comparisons in the natural setting provides additional leverage. The likelihood that the null belief effects from Studies 1 and 2 are due to inadequate power is low (i.e., less than 20%). More specifically, we can represent this probability as $(1 - .54) * (1 - .38) * (1 - .40) = .17$, where .54, .38, and .40 are the power levels in Study 1 (Polimetrix and KN) and Study 2, respectively (see Keppel 1982, ch. 4). When viewed across multiple studies, the null belief effects are unlikely to be an artifact of low statistical power.[26]

We also replicated the constraint analyses from Study 1. In the case of immigration, there was no evidence that the survey experiment induced constraint among respondents. Beliefs were closely related to immigration policy opinions, even among individuals in the control group. Absent a significant treatment effect for constraint in the survey experiment, there is no expectation that one would be found in the real world—and that is exactly what the data show.

All in all, the findings were consistent with those from Study 1. Across separate analyses, we observed statistically significant information and belief effects in the survey experiments. In contrast, the natural experiments showed evidence of learning only. This implies that in the real world, it is possible to get people to absorb information, but they do not process this information in the same manner as the survey experiment. Indeed, survey experiments appear to induce a series of downstream attitudinal effects (e.g., belief change, increased constraint) that have no apparent counterpart in the real world. Thus, scholars might come to different conclusions about citizen competence depending on how they examine public opinion.

[26] In another attempt to find belief effects in Study 2, we took advantage of the larger number of cases and explored the interaction between our media exposure measure and a person's level of education. We found no evidence of belief effects among this subgroup of respondents.

TABLE 3. Comparison of Survey and Natural Experiment Effects

		Survey Experiment Effect	Natural Experiment Effect	SE – NE Difference	*t* Ratio	*p* Value
	Medicare Polimetrix	.15 (.05)	.13 (.09)	.02 (.10)	.19	.42
Learning effects	Medicare KN	.22 (.04)	.09 (.07)	.13 (.08)	1.61	.05
	Immigration YouGov/Polimetrix	.15 (.04)	.11 (.07)	.04 (.08)	.50	.31
	Average effect size	.17	.11			
	Medicare Polimetrix	.11 (.04)	.08 (.08)	.03 (.09)	.34	.37
Belief effects	Medicare KN	.04 (.04)	−.06 (.05)	.10 (.06)	1.56	.06
	Immigration YouGov/Polimetrix	.06 (.04)	.02 (.06)	.04 (.07)	.55	.29
	Average effect size	.07	.01			
	Medicare Polimetrix	.10 (.05)	.12 (.14)	−.02 (.15)	−.13	.55
Opinion integration	Medicare KN	.14 (.08)	−.14 (.10)	.28 (.13)	2.19	.01
	Average effect size	.12	−.01			

Notes: Cell entries for learning and belief effects in the first two columns reproduce the effect sizes reported in Tables 1 and 2. Standard errors appear in the parentheses. The bottom portion of Table 3 shows the marginal effects from the constraint analyses described in the text. For ease of presentation, the marginal effects are based on a dichotomous dependent variable. Average effect sizes are shown in the boxes with gray shading.

Overview of Findings in Studies 1 and 2

We close by summarizing the findings from both studies. Table 3 reproduces the effect sizes from the survey and natural experiments (based on the analyses reported in Tables 1 and 2). For ease of interpretation, all effects are on the 0–1 interval.[27]

As the top panel shows, the average learning effect in the survey experiment was .17, roughly 50% larger than the average effect in the natural experiment (avg. = .11). The middle panel shows that, on average, respondents changed their policy beliefs by about .07 units in the survey experiment, but exhibited virtually no belief change in the natural experiments. Similarly, in the bottom panel of Table 3, respondents increased the association among policy-relevant beliefs by about .12 units in the survey experiment, whereas individuals in the natural experiment show essentially no average increase in opinion integration.

Of course, it is one thing to show a difference in effect size across the two contexts; it is quite another to demonstrate that those differences are statistically significant. We address this issue in the right-hand side of Table 3, where we present the difference in effect size across the survey and natural experiments (third column) and the corresponding *t* ratios and *p* values (fourth and fifth columns). For illustrative purposes, we focus on learning effects in the Polimetrix survey (Medicare). Here the difference between the survey and natural experiments is .02, with a *t* ratio and one-tailed *p* value of .19 and .42, respectively. This level of statistical significance is low. The difference between the survey and natural experiments is larger for the other two learning comparisons, achieving statistical significance in the KN survey (difference in effect size = .13; $p = .05$). Taken together, we conducted three independent tests of learning, and all show the same pattern. According to a sign test, the probability of obtaining these three results by chance alone is .125. Overall, then, we have differences between the survey and natural experiments that are substantively large but statistically marginal.[28] Although we have documented discrepancies in effect size across three domains (learning effects, belief effects, and opinion

[27] We reversed the coding on the belief item in Study 2 so that all survey experiment effects run in the same direction. At the bottom of Table 3, we show the marginal effects from the constraint analyses described previously.

[28] We thank the editors of the *APSR* for their help with these calculations. On the various ways to combine significance tests from multiple studies, see Loughin (2004). In auxiliary analyses not shown here, we obtain the same pattern of results with coefficient tests from statistical models that include terms for the survey and natural experiments. See Achen (1982) on the importance of not disregarding results that are substantively large but statistically marginal. Achen's point is especially important in cases, like this one, in which data are difficult to obtain.

integration), additional research with large samples will be needed to confirm that the differences between survey and natural experiments are real.[29]

Before placing the findings from Studies 1 and 2 in broader context, it is important to comment on the robustness of our results. In the case of learning, the findings from the survey experiment always exceed high levels of statistical significance (i.e., z test critical values of 1.96, the benchmark for a two-tailed test at alpha = .05). The results from the natural experiment are more tenuous in that the learning findings are statistically significant but at a lower threshold (i.e., z values of 1.28, the $p < .10$ critical value for a one-tailed test). In that sense, different choices regarding the data and analysis could weaken the information effects in the natural experiment. In an attempt to determine whether differences in sample composition play a role in the findings we observe, we include controls for background characteristics (e.g., gender, race, income, age, education, partisanship) along with the treatment indicators. In those auxiliary analyses, we obtain the same pattern of results. That is, we find strong learning and belief effects in the survey experiments, but in the natural experiments, we find learning among the highly exposed with no corresponding belief effects.

DISCUSSION

Although the real world does not look so different as to throw into doubt the validity of survey experiments, there is drop-off in terms of both the size of treatment effects and the population experiencing those effects (i.e., the significant findings are observed only among a subgroup highly exposed to the media). This attenuation may stem from several sources: (1) the relatively modest level of attention the national media devoted to our two government announcements; (2) the differences in reception of the treatment across the survey and natural experiments; or (3) because of competing stimuli, real-world media treatments may have a lower impact than those appearing in a survey experiment. In the rest of this section, we consider each source in more detail.[30]

Coverage of our two news events appeared in dozens of news stories. This is not a trivial amount of coverage, but it also did not reach the level of the 2009–10 debate over health care reform or the run-up to the 2003 Iraq war. Instead, and as noted previously, the trustees' 2007 report and the new citizenship test received about the same level of media coverage as other routine political events. It is important to study the effects of ordinary news events because they, along with extraordinary events, form the raw ingredients of public opinion. The variation in effect sizes across the survey and natural experiments (Table 3) suggests that the typical treatment in a survey experiment might better approximate news events receiving a substantial amount of coverage (i.e., in multiple outlets and over a period of weeks, not days).[31]

Another source of attenuation is the lower levels of exposure that occur in the real world when compared with the experimental setting. As we documented in Study 1, only a minority of the Polimetrix sample (about 30%) reported using a source that included the exhaustion date information (the corresponding figure for Study 2 is 27%). In the experiment, in contrast, exposure to the information treatment was forced. In general, we expect the level of exposure in a survey experiment to exceed that found in the real world. This expectation implies that when researchers compare experimental and observational studies, there is likely to be a discrepancy in effect size (e.g., Hovland 1959). The challenge then becomes how to translate significant experimental findings into statements about the effects of similar treatments in a natural setting (e.g., Brader 2005, 402). In our study, for example, how do we interpret the significant treatment effects we observed in our survey experiments? The obvious answer is that they indicate the likely effect among people receiving similar messages in the real world. But how many researchers acknowledge that the treatment effects they observe in a nationally representative survey experiment may be present only among a subset of the population?[32]

A third source of attenuation is that for any given level of exposure, real-world media treatments may have a lower impact than those appearing in a survey experiment. Here the comparison of learning effects in Table 3 is instructive. Recall that the knowledge questions were the closest, in terms of substance, to the treatment. We were also able to identify subgroups of the public that were exposed to the report and found significant learning effects among this group. One might then expect high levels of correspondence between the survey and natural experiments on the knowledge questions. Yet, this was where we observed one of the largest differences in treatment effect size (.22 − .09 = .13 in KN; Table 3). In this instance, one cannot attribute the divergence to differences in exposure because we were able to identify respondents whose news source carried the exhaustion date information. Instead, we suspect that attenuation may stem from the multiple stimuli (competing messages) in the real world that reduce the effect of any given message. In contrast, the more pristine experimental setting may

[29] Differences between the survey and natural experiments generally were larger for belief effects and opinion integration than for learning. We had no a priori basis for expecting the fall-off from survey to natural experiment to be greater for beliefs than for knowledge. A two-tailed test of the significance of the observed pattern of smaller effects in the natural experiments is therefore appropriate. A sign test comparing the results of the three knowledge items with the three belief questions indicates that the p value of the difference is .25 ($2 \times .125$).

[30] Our discussion treats each factor in isolation, although each may be operating at the same time to produce a discrepancy in the findings from survey experiments and the natural world.

[31] At the same time, researchers run the risk of pretreatment bias when designing survey experiments in the wake of high-profile political events. In this instance, one might observe sizeable media effects in the real world, but muted or insignificant findings in a survey experiment because people have already been treated (e.g., Gaines, Kuklinski, and Quirk 2007).

[32] Scholars have made strides in this area by making treatments probabilistic (Arceneaux and Johnson 2007) and allowing selection mechanisms to operate (Gaines and Kuklinski 2009).

exaggerate the power of the stimulus (Kinder 2007). From this perspective, survey experiments might operate in a way that is analogous to an existence proof, demonstrating what treatment effects in the real world might look like if the entire population received a particular message.[33]

Another important finding to emerge from this study is the differential pattern for information and belief effects. In the survey experiment, we observed changes in knowledge, beliefs, and (on the Medicare issue) attitude constraint; in the natural experiment, individuals were able to recall the key facts reported by the two government announcements, but the content and organization of their beliefs remained unchanged. What accounts for this difference? Belief effects require more than just recall, and instead involve some degree of thought and integration. It may be easier to observe belief change in a survey experiment because subjects digest the treatments more thoroughly than their real-world counterparts. This raises the possibility that survey experiments may induce a variety of second-order effects (e.g., making inferences, integrating beliefs and attitudes) that are unlikely to occur in the wake of actual political events.

CONCLUSION

Researchers conduct experiments for a variety of reasons (e.g., Roth 1995), but what distinguishes political scientists' use of this method is "their attention to focal aspects of politically relevant contexts" (Druckman et al. 2006, 629). For scholars seeking to understand and to make predictions about public opinion, the relevant context often implies stimuli that have a real-world referent. Survey experimenters, in particular, go to great lengths to make their treatments correspond to features of the political world, such as persuasive arguments, campaign ads, facial displays, and the like. Despite the tremendous amount of resources devoted to designing and administering survey experiments, it has been a mystery as to whether the findings generated by survey experiments correspond to the political world. The results presented here should be encouraging to anyone devoted to the scientific study of politics because they suggest that what occurs in survey experiments resembles what takes place in the real world.

Although there was a discrepancy between the size of survey treatment effects and the general population in our natural experiment, we observed correspondence exactly where one would expect to find it—among those who were most likely to be exposed to media messages about the two government announcements. To be sure, there was some drop-off. The treatment effects in the experiments were nearly always larger than what was observed in the natural settings. Nevertheless, if researchers consider the possible sources of attenuation and carefully interpret the meaning of significant treatment effects, survey experiments will continue to be an important tool for studying public opinion. Indeed, we believe that one of the most promising ways to learn about the sources of attenuation is through the continued use of experiments.

As for the weaknesses of this study, we highlight two issues. The first pertains to the nature of the samples used in our analyses. The two Polimetrix surveys are convenience samples, made up of individuals who volunteer for research. These surveys are intended to be representative and the firm attempts to match respondents to the U.S. population, but the Polimetrix surveys are not strict probability samples. Despite this potential limitation, respondents in these surveys represent a broad cross-section of people (see online Appendix for details). Moreover, recent research has found lower levels of satisficing in Internet samples, which may improve the quality of those data vis-à-vis telephone surveys (Weisberg 2005).[34]

The second and, in our view, more serious flaw is the limited statistical power in our natural experiments. As a result of this weakness, the reader may be left with some doubt as to the credibility of the null findings reported in Studies 1 and 2. Unfortunately, there is little a researcher can do to improve statistical power after a study has been fielded. Instead, we hope future scholars will learn from the drawbacks of this study. In particular, researchers who are interested in comparing treatment effects across different settings (e.g., survey experiment and natural world) should use large samples in order to conduct the requisite subgroup analyses. This is especially true when the expected effects are small (O'Keefe 2007) or when they involve interactive specifications (McClelland and Judd 1993). In general, statistical power deserves more attention in studies of public opinion (see Zaller 2002).

When it comes to the theoretical issues at stake, we are less sanguine. Political knowledge has been described as the "the anchor that tethers attitudes to each other, to behavioral intentions, and to the empirical world" (Delli Carpini 2009, 27). Although our survey experiments give the impression of a "rational public," one that reacts to new information in a sensible way (e.g., Page and Shapiro 1992), we found less support for this proposition when we examined public opinion in the natural world. And, yet, the two government announcements we examined are precisely the sort of news events that ought to create an informed public. From what or whom, then, does the public learn? Presumably from episodic policy debates, such as the 2009 debate over health care reform, and from elections, when candidates announce various prescriptions for change. However, communication in these situations often consists of ideological rhetoric and stark predictions (Jerit 2009; Jerit, Kuklinski, and Quirk

[33] We thank Liz Gerber for this observation.

[34] The KN survey is a probability survey, but we have chosen to use unweighted data, both to be consistent in our presentation across the three sets of analyses and also because researchers have noted challenges with using weights in statistical analyses of survey data (e.g., Brehm 1993; Cameron and Trivedi 2005; Gelman 2007; Weisberg 2005; Winship and Radbill 1994). In addition, the KN weights are based on observed variables, including a person's level of education. This was another factor in our decision not to use weights because education appears in the analysis of the KN data (it serves as our proxy for media exposure). However, and as we note previously, we obtain substantively similar results when we employ survey weights.

2009)—hardly the stuff of enlightened deliberation (e.g., Barabas 2004; Jacobs, Cook, and Delli Carpini 2009; Luskin, Fishkin, and Jowell 2002). Even if elite rhetoric was of higher quality, citizens who enter campaigns and debates without any contextual understanding because they have not routinely followed the news might not make optimal choices. It becomes essential, then, to conduct the kinds of comparisons we do here, if only to place experimental research in context and to better understand the public's true capabilities.

APPENDIX

In this Appendix, we provide details on response rates, question order, and the original conditions in the KN survey. Additional information about the methodology underlying each survey, sample characteristics, randomization, and attrition are found in the online Appendix.

Response Rates

The cross-sectional survey on the Medicare announcement was conducted by Knowledge Networks (KN), an Internet opinion polling firm. To select a sample, KN identified potential respondents from a nationally representative, probability-based web panel. Between March 2 and March 10, 2007, KN completed 805 interviews out of 1,143 eligible respondents who were contacted for an interview. This results in a 70.4% completion rate. Although the American Association for Public Opinion Research (AAPOR) response rate standards have not been formally established for web panels, the completion statistic corresponds to AAPOR Response Rate 3. In the second wave of the survey, between April 26, 2007 and May 3, 2007, a new cross-section of 817 respondents completed interviews out of 1,143 eligible who were contacted (817/1,143 = 71.5%).

For the second survey on Medicare, the authors contracted with Polimetrix (which subsequently became known as YouGov/Polimetrix) to conduct a panel survey during the spring of 2007. The first wave was conducted from March 1 to 21, 2007. Polimetrix interviewed thousands of respondents from their panel—a pool of several hundred thousand individuals who volunteered or were recruited to participate in occasional online polls. For the second wave from April 26 to May 16, 2007, Polimetrix reinterviewed 64% of the respondents who had previously completed wave 1 of this study. Detailed response rates are not available.

The immigration study was conducted as a part of the Cooperative Campaign Analysis Project (CCAP), which is a six-wave panel study with an oversample in contested battleground and early primary states (FL, IA, MN, NV, WI, NH, NM, OH, PA). In that sense, the CCAP sample more closely represents states with competitive elections rather than the entire nation. There was a baseline survey in December 2007, with subsequent panel waves in January, March, September, October, and November 2008. The analyses in this study employ 1,039 respondents from the September and October waves. Detailed response rates are not available.

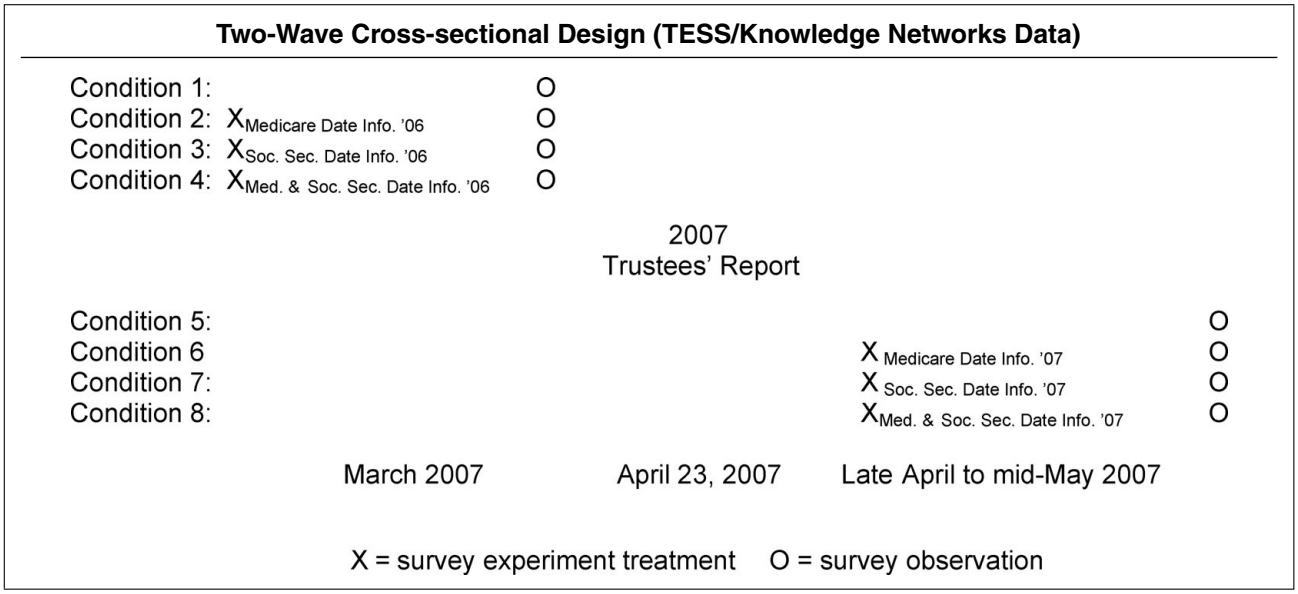
Question Order

The outcome measures are asked relatively soon after the treatments, although there is variation on this dimension across our three surveys. In the KN surveys (Study 1), the belief question is asked immediately after the treatment or with a different question in between, depending on a random rotation in the questionnaire format. The next question was the knowledge item, and the policy preference item was either the fourth or fifth question following the treatment (again depending on the randomization pattern). In Polimetrix (Study 1), the belief question was either the third or fourth question after the treatment, the knowledge item was five questions later (i.e., 8 or 9 questions after the treatment), and the policy preference measure appeared two questions after the knowledge question (i.e., 11 questions after the treatment). Finally, in the CCAP survey (Study 2), the knowledge question appeared immediately after the treatment. The belief question followed the knowledge item.

Experimental Conditions in the Knowledge Networks Study

The KN study divides a Time-sharing Experiments in the Social Sciences (TESS) survey with 1,622 subjects into two parts (see schematic below). Half of the subjects ($n = 805$) were randomly selected to be interviewed about 1 month *prior* to the release of the trust fund report. The other half was interviewed after the event.

Two-Wave Cross-sectional Design (TESS/Knowledge Networks Data)



Respondents were randomly divided into four groups. The first ($n = 206$, the control condition) was asked a series of questions about Medicare *without* being given any information about the exhaustion dates of the two programs. Respondents in the remaining treatment conditions were exposed to varying amounts of policy-relevant information before answering the outcome measures. People in the second condition ($n = 202$) were shown information about the fiscal status of Medicare. Those in the third condition ($n = 196$) were provided information about Social Security, whereas those in the fourth group ($n = 201$) received information about *both* Medicare and Social Security. The last condition mimicked actual coverage of the trustee's report, which highlighted the fact that Medicare's trust fund was projected to be exhausted in 2018 compared with the date of 2040 for Social Security. Because we did not know how the report would be covered by news organizations, we created multiple treatment conditions, some with partial date information (e.g., conditions 2 and 3 at t_1).

Much as one might expect, respondents who were given the exhaustion date of each program separately performed no better than the control group on the knowledge question. In other words, there are no significant differences in knowledge or beliefs in comparisons of condition 2 versus condition 1 or condition 3 versus condition 1 (at time 1). This pattern confirms that our central manipulation—providing both dates—worked as intended. It also allowed us to expand the size of the control group by combining the conditions with partial information with the original control condition (for a similar approach, see Chong and Druckman 2007, 649, note 21). We used the combined conditions throughout the entire analysis of the KN survey.[35]

REFERENCES

Abrajano, Marisa A. 2005. "Who Evaluates a Presidential Candidate by Using Non-policy Campaign Messages?" *Political Research Quarterly* 58 (March): 55–67.

Achen, Chris. 1982. *Interpreting and Using Regression*. Thousand Oaks, CA: Sage.

Arceneaux, Kevin, and Martin Johnson. 2007. "Channel Surfing: Does Choice Reduce Videolmalaise." Presented at the 2007 Annual Meeting of the Midwest Political Science Association, Chicago.

Arceneaux, Kevin, and Robin Kolodny. 2009. "Educating the Least Informed: Group Endorsements in a Grassroots Campaign." *American Journal of Political Science* 53 (October): 755–70.

Barabas, Jason. 2004. "How Deliberation Affects Policy Opinions." *American Political Science Review* 98 (November): 687–701.

Barabas, Jason, and Jennifer Jerit. 2009. "Estimating the Causal Effects of Media Coverage on Policy-specific Knowledge." *American Journal of Political Science* 53 (January): 73–89.

Bennett, W. Lance. 2006. *News: The Politics of Illusion*. 7th ed. New York: Longman.

Benz, Matthias, and Stephan Meier. 2008. "Do People Behave in Experiments as in the Field? Evidence from Donations." *Experimental Economics* 11: 268–81.

Berinsky, Adam J. 2007. "Assuming the Costs of War: Events, Elites, and American Public Support for Military Conflict." *Journal of Politics* 69 (November): 975–97.

Blalock, H.M., Jr. 1979. *Social Statistics*. 2nd ed. New York: McGraw-Hill.

Brader, Ted. 2005. "Striking a Responsive Chord: How Political Ads Motivate and Persuade Voters by Appealing to Emotions." *American Journal of Political Science* 49 (April): 388–405.

Brehm, John. 1993. *The Phantom Respondents: Opinion Surveys and Political Representation*. Ann Arbor: University of Michigan Press.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-experimental Designs for Research*. Chicago: Rand McNally.

Chong, Dennis, and James N. Druckman. 2007. "Framing Public Opinion in Competitive Democracies." *American Political Science Review* 101 (November): 637–56.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum.

Converse, Philip E. 1964. "The Nature of Belief Systems in Mass Publics." In *Ideology and Discontent*, ed. David E. Apter. New York: Free Press.

Cook, Fay Lomax, Lawrence R. Jacobs, and Dukhong Kim. 2010. "Trusting What You Know: Information, Knowledge, and Confidence in Social Security." *Journal of Politics* 72 (April): 397–412.

Cook, Thomas D., and Donald T. Campbell. 1979. *Quasi-experimentation: Design & Analysis Issues for Field Settings*. New York: Houghton Mifflin.

Cook, Thomas D., William R. Shadish, and Vivian C. Wong. 2008. "Three Conditions under Which Experimental and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-study Comparisons." *Journal of Policy Analysis and Management* 27 (Autumn): 724–50.

Delli Carpini, Michael X. 2009. "The Psychology of Citizen Learning." In *The Political Psychology of Democratic Citizenship*, eds. Eugene Borgida, Christopher M. Federico, and John L. Sullivan. New York: Oxford University Press, 23–51.

Druckman, James N. 2004. "Political Preference Formation: Competition, Deliberation, and the (Ir)relevance of Framing Effects." *American Political Science Review* 98 (November): 671–86.

Druckman, James N., Donald P. Green, James H. Kuklinski, and Arthur Lupia. 2006. "The Growth and Development of Experimental Research in Political Science." *American Political Science Review* 100 (November): 627–35.

Druckman, James N., and Kjersten R. Nelson. 2003. "Framing and Deliberation: How Citizens' Conversations Limit Elite Influence." *American Journal of Political Science* 47 (October): 729–45.

Dunning, Thad. 2008. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." *Political Research Quarterly* 61 (2): 282–93.

Eagly, Alice, and Shelly Chaiken. 1993. *The Psychology of Attitudes*. New York: Harcourt Brace.

Gaines, Brian J., and James H. Kuklinski. 2009. "Including Self-selection in Random Assignment Experiments." University of Illinois at Urbana–Champaign. Unpublished manuscript.

Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk. 2007. "The Logic of the Survey Experiment Reexamined." *Political Analysis* 15 (Winter): 1–20.

Gelman, Andrew. 2007. "Struggles with Survey Weighting and Regression Modeling." *Statistical Science* 22 (2): 153–64.

Gerber, Alan, and Donald P. Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." *American Political Science Review* 94 (September): 653–63.

Gibson, James. 1998. "A Sober Second Thought: An Experiment in Persuading Russians to Tolerate." *American Journal of Political Science* 42 (July): 819–50.

Gilens, Martin. 2001. "Political Ignorance and Collective Policy Preferences." *American Political Science Review* 95 (June): 379–96.

Gilens, Martin. 2002. "An Anatomy of Survey-Based Experiments." In *Navigating Public Opinion*. eds. J. Manza, F. L. Cook, and B. I. Page. New York: Oxford, 232–50.

Gramlich, Edward M. 1998. *Is It Time to Reform Social Security?* Ann Arbor: University of Michigan Press.

Grosser, Jens, and Arthur Schram. 2006. "Neighborhood Information Exchange and Voter Participation: An Experimental Study." *American Political Science Review* 100 (May): 235–48.

Hovland, Carl. 1959. "Reconciling Conflicting Results Derived from Experimental and Survey Studies of Attitude Change." *American Psychologist* 14: 8–17.

[35] In this study, we focus on the t_1 survey experiments because these observations were not contaminated by real-world coverage of Medicare and therefore represent our cleanest estimate of treatment effects in the survey experiment.

Huber, Gregory A., and Kevin Arceneaux. 2007. "Identifying the Persuasive Effects of Presidential Advertising." *American Journal of Political Science* 51 (October): 957–77.
Iyengar, Shanto, and Donald Kinder. 1987. *News That Matters*. Chicago: University of Chicago Press.
Jacobs, Lawrence, R., Fay Lomax Cook, and Michael X. Delli Carpini. 2009. *Talking Together: Public Deliberation and Political Participation in America*. Chicago: University of Chicago Press.
Jacobs, Lawrence R., and Robert Y. Shapiro. 1998. "Myths and Misunderstandings about Public Opinion toward Social Security." In *Framing the Social Security Debate: Values, Politics, and Economics*, eds. R. Douglas Arnold, Michael J. Graetz, and Alicia H. Munnell. Washington, DC: Brookings Institution Press, 355–88.
Jerit, Jennifer. 2009. "How Predictive Appeals Shape Policy Opinions." *American Journal of Political Science* 53 (April): 411–26.
Jerit, Jennifer, and Jason Barabas. 2006. "Bankrupt Rhetoric: How Misleading Information Affects Knowledge about Social Security." *Public Opinion Quarterly* 70 (Fall): 278–303.
Jerit, Jennifer, James H. Kuklinski, and Paul J. Quirk. 2009. "Strategic Rhetoric, Emotional Citizens, and the Rhetoric of Prediction." In *The Political Psychology of Democratic Citizenship*, eds. Eugene Borgida, John L. Sullivan, and Christopher M. Federico. New York: Oxford University Press, 100–24.
Kam, Cindy D., Jennifer R. Wilking, and Elizabeth J. Zechmeister. 2007. "Beyond the 'Narrow Database': Another Convenience Sample for Experimental Research." *Political Behavior* 29 (December): 415–40.
Kellstedt, Paul M., and Guy D. Whitten. 2009. *The Fundamentals of Political Science Research.* New York: Cambridge University Press.
Kennedy, Peter. 2003. *A Guide to Econometrics*. 5th ed. Cambridge, MA: MIT Press.
Keppel, Geoffrey. 1982. *Design and Analysis: A Researcher's Handbook*. 2nd ed. Englewood Cliffs, NJ: Prentice-Hall.
Kinder, Donald. 2007. "Curmudgeonly Advice." *Journal of Communication* 57 (March): 155–62.
Kinder, Donald R., and Thomas R. Palfrey, eds. 1993. *Experimental Foundations of Political Science*. Ann Arbor: University of Michigan Press.
Kinder, Donald R., and Lynn M. Sanders. 1996. *Divided by Color: Racial Politics and Democratic Ideals*. Chicago: University of Chicago Press.
King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.
Krosnick, Jon A., and Howard Schuman. 1988. "Attitude Intensity, Importance, and Certainty and Susceptibility to Response Effects." *Journal of Personality and Social Psychology* 54 (6): 940–52.
Kuklinski, James H., Paul J. Quirk, Jennifer Jerit, and Robert F. Rich. 2001. "The Political Environment and Citizen Competence." *American Journal of Political Science* 45 (April): 410–24.
Kuklinski, James H., Paul J. Quirk, Jennifer Jerit, David Schwieder, and Robert F. Rich. 2000. "Misinformation and the Currency of Citizenship." *Journal of Politics* 62 (August): 790–816.
Lassen, David Dreyer. 2005. "The Effect of Information on Voter Turnout: Evidence from a Natural Experiment." *American Journal of Political Science* 49 (January): 103–88.
Lavine, Howard. 2002. "On-line versus Memory Based Process Models of Political Evaluation." In *Political Psychology*, ed. Kristen Monroe. Mahwah, NJ: Lawrence Erlbaum Associates, 225–47.
Levitt, Steven D., and John A. List. 2007. "What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?" *Journal of Economic Perspectives* 21 (Spring): 153–71.
Loughin, Thomas M. 2004. "A Systematic Comparison of Methods for Combining *p-Values from Independent Tests.*" *Computational Statistics and Data Analysis* 47: 467–85.
Lupia, Arthur, Yanna Krupnikov, Adam Seth Levine, Casandra Grafstrom, William McMillian, and Erin McGovern. n.d. "How 'Point Blindness' Dilutes the Value of Stock Market Reports." *Political Communication*. Forthcoming.
Luskin, Robert C., James S. Fishkin, and Roger Jowell. 2002. "Considered Opinions: Deliberative Polling in Britain." *British Journal of Political Science* 32: 455–87.
Marmor, Theodore R. 2000. *The Politics of Medicare*. 2nd ed. NY: Aldine de Gruyter.
McClelland, Gary H., and Charles M. Judd. 1993. "Statistical Difficulties of Detecting Interactions and Moderator Effects." *Psychological Bulletin* 114 (2): 376–90.
McDermott, Rose. 2002. "Experimental Methods in Political Science."*Annual Review of Political Science* 5: 31–61.
Mondak, Jeffery J. 1995. "Newspapers and Political Awareness." *American Journal of Political Science* 39 (May): 513–27.
Mutz, Diana, and Byron Reeves. 2005. "The New Videomalaise: Effects of Televised Civility on Political Trust." *American Political Science Review* 99 (February): 1–15.
Nadeau, Richard, and Richard G. Niemi. 1995. "Educated Guesses: The Process of Answering Factual Knowledge Questions in Surveys." *Public Opinion Quarterly* 59 (Autumn): 323–46.
Norris, Pippa, and David Sanders. 2003. "Message or Medium? Learning during the 2001 British General Election." *Political Communication* 20: 233–62.
Oberlander, Jonathan. 2003. *The Political Life of Medicare*. Chicago: University of Chicago Press.
O'Keefe, Daniel J. 2007. "Post Hoc Power, Observed Power, A Priori Power, Retrospective Power, Prospective Power, Achieved Power: Sorting Out Appropriate Uses of Statistical Power Analyses." *Communication Methods and Measures* 1(4): 291–99.
Page, Benjamin I. 2000. "How Public Opinion Affects Reform: Is Social Security Reform Ready for the American Public?" In *Social Security and Medicare: Individual versus Collective Risk and Responsibility*, eds. Shelia Burke, Eric Kingson, and Uwe Reinhardt. Washington, DC: National Academy of Social Insurance, 183–218.
Page, Benjamin I., and Robert Y. Shapiro. 1992. *The Rational Public: Fifty Years of Trends in Americans' Policy Preferences*. Chicago: University of Chicago Press.
Prior, Marcus, and Arthur Lupia. 2008. "Money, Time, and Political Knowledge: Distinguishing Quick Recall and Political Learning Skills." *American Journal of Political Science* 52 (January): 169–83.
Robinson, Gregory, John E. McNulty, and Jonathan S. Krasno. 2009. "Observing the Counterfactual? The Search for Political Experiments in Nature." *Political Analysis* 17: 341–57.
Roth, Alvin. 1995. "Introduction to Experimental Economics." In *The Handbook of Experimental Economics*, eds. John H. Kagel and Alvin E. Roth. Princeton, NJ: Princeton University Press, 3–109.
Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.
Shaw, Greg M., and Sarah E. Mysiewicz. 2004. "Trends: Social Security and Medicare." *Public Opinion Quarterly* 68 (Fall): 394–423.
Sniderman, Paul M., and Thomas Piazza. 1993. *The Scar of Race.* Cambridge, MA: Harvard University Press.
Sniderman, Paul M., and Sean M. Theriault. 2004. "The Structure of Political Argument and the Logic of Issue Framing." In *Studies in Public Opinion: Attitudes, Nonattitudes, Measurement Error, and Change*, eds. Willem E. Saris and Paul M. Sniderman. Princeton, NJ: Princeton University Press, 133–64.
Turgeon, Mathieu. 2009. "'Just Thinkin': Attitude Development, Public Opinion, and Political Representation." *Political Behavior* 31 (September): 353–78.
Weisberg, Herbert, F. 2005. *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. Chicago: University of Chicago Press.
Winship, Christopher, and Larry Radbill. 1994. "Sampling Weights and Regression Analysis." *Sociological Methods and Research* 23 (2): 230–57.
Wooldridge, Jeffrey M. 2009. *Introductory Economics: A Modern Approach.* 4th ed. Mason, OH: South-Western Cengage Learning.
Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. New York: Cambridge.
Zaller, John R. 2002. "The Statistical Power of Election Studies to Detect Media Exposure Effects in Political Campaigns." *Electoral Studies* 21: 297–329.
Zaller, John R., and Stanley Feldman. 1992. "A Simple Theory of the Survey Response: Answering Questions versus Revealing Preferences." *American Journal of Political Science* 36 (August): 579–616.