Survey Experiments and the External Validity of Treatments

Jason Barabas[1]
Jennifer Jerit[2]

Florida State University
Department of Political Science
531 Bellamy Building
Tallahassee, FL 32306
850-644-5727 (phone)
850-644-1367 (fax)

Prepared for Presentation at the Conference on Experimentation in Political Science
at the Annual Meetings of the Canadian Political Science Association
Vancouver, British Columbia
June 6, 2008

Abstract

Survey experiments help establish causality, but scholars do not know how closely the treatments mimic natural phenomena. This study compares survey experiments and a natural experiment on the same topic. In two survey experiments providing information about Medicare, we observe double-digit learning effects. In contrast, most respondents in our contemporaneous natural experiment show little evidence of learning. Consistent with our expectations, the only people who showed comparable levels of learning to respondents in our survey experiment were individuals exposed to Medicare facts in their media source of choice as well as people who were uncertain about the facts from the very beginning. Our conclusion is that survey experiments, at least on this topic, generate effects that are only observed among parts of the population who are likely to be exposed to treatment messages or predisposed to accept them.

1 Assistant Professor, Department of Political Science, Florida State University, jason.barabas @ fsu.
2 Assistant Professor, Department of Political Science, Florida State University, jjerit @ fsu.edu.

Randomized social science experiments permit causal inference, but a common criticism concerns the lack of external validity when, as is often the case, the subjects are university undergraduates. Selecting individuals randomly from a national pool offsets these concerns and allows generalizations to the larger population. However, even in nationally representative survey experiments we should be concerned about the external validity of the treatments—i.e., the correspondence between the stimuli and the political phenomena they seek to emulate. Unlike some disciplines (e.g., economics or psychology) where the primary goal of experiments is testing some aspect of a theory, "*Political scientists use survey experiments to identify how citizens make decisions and respond to real-world political objects, in order to enhance understanding of politics*" (Gaines, Kuklinski, and Quirk 2007, 2; emphasis original; also see Kinder and Palfrey 1993, 27). In this way, the correspondence between the experimental stimuli and real world referent is assumed. A significant treatment effect in a survey experiment is viewed as saying something about the direction, if not the rough magnitude, of effects that might be expected to occur in the real world.

This paper examines whether this is in fact the case. We do so by taking advantage of a naturally occurring event that was expected to increase media attention to Medicare. Each spring, the trustees overseeing Medicare and Social Security release their future funding estimates. We leveraged this situation by bracketing the event with survey experiments that delivered information about Medicare's financial status to randomly selected samples of the American public. As a result, this study offers a rare opportunity to compare survey experiments with the real world phenomena they seek to emulate. In the case we examine, the only people who showed comparable levels of learning to respondents in our survey experiment were individuals exposed to Medicare facts in their media source of choice as well as people who were uncertain

about the facts from the very beginning. We conclude that the typical survey experiment generates effects that are only likely to be observed among particular subgroups of the population, not necessarily the public at large.

Varieties of Experiments and Validity

Years ago social science experiments were rare and scholars implored others to use them (Kinder and Palfrey 1993). Today, however, experiments are more common and have been used in studies of everything from framing (Druckman 2004; Nelson, Clawson, and Oxley 1997) and stereotyping (Peffley, Hurwitz, and Sniderman 1997; Berinsky and Mendelberg 2005) to group conflict (Grant and Rudolph 2003) and political knowledge (Gilens 2001). The studies just cited and dozens more deliver their treatments via surveys (e.g., Burden and Klofstad 2005; Krosnick and Schuman 1988; Schuman and Bobo 1999), but experiments also come in other varieties (for reviews, see Druckman, Green, Kuklinski and Lupia 2006; McDermott 2002a; 2002b). There are laboratory experiments, typically conducted in universities with students or members of the local community as subjects (e.g., Grosser and Schram 2006; Iyengar and Kinder 1987; Kam 2007). There also are natural and quasi-experiments in which analysts look for variations in real world phenomena without manipulating them (e.g., Barabas 2004; Lassen 2006; Mondak 1995).[1]

[1] 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. A quasi-experiment is "an experiment in which units are not randomly assigned to conditions" (p. 511; also see Cook and Campbell 1979). Some scholars reject these distinctions, arguing that a study either is an experiment or it is not (e.g., King, Keohane, and Verba 1994, 7, note 1).

A key advantage of experiments, particularly those with randomized treatments, is that they have a high degree of internal validity (i.e., they tell us whether one factor causes another). Survey experimenters, who typically rely on nationally representative adult samples, often claim the mantle of external validity as well. But the issue goes beyond the representativeness of the subjects. As Shadish, Cook, and Campbell (2002) define it, "External validity concerns inferences about the extent to which a causal relationship holds over variations in persons, settings, treatments, and outcomes" (p. 83). For scholars embedding experiments in opinion surveys, this means paying attention to whether the "stimuli" (i.e., the treatments) are themselves externally valid (see Druckman 2004, 684-5 or Gilens 2002, 248). One concern, Kinder and Palfrey (1994, 27) write, is that a study's findings are "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 it is unrealistic to examine the effects of a single frame when citizens typically experience competing frames (Chong and Druckman 2008). To the extent that treatments in survey experiments are overly strong, the effects we observe may not generalize to other settings.[2]

There is a second, and related, issue. In his reflection on framing effects research, Kinder (2007, 157) worries about the manner in which treatments are received:

> …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.

[2] As others have noted, there is a tradeoff between the realism of the experimental setting and the control exerted by the researcher (Gaines et al. 2007; McDermott 2002b; Schram 2005).

Likewise, in the framework offered by Gaines, Kuklinski, and Quirk (2007, 15), there is an inflation parameter, which reflects the fact that "the artificially clean environment of the survey question makes treatment easier to receive than in real life." Thus, the natural world contains competing messages and other distractions that make exposure to the treatment probabilistic; in a survey experiment, exposure is essentially forced. On the other hand, in the natural world, there is the potential for reinforcement through repetition of information in the media and interpersonal discussion. In a single-shot survey, exposure to the treatment usually occurs only once. The important point, from our perspective, is that these differences are consequential for how we interpret the results of survey experiments—and to whom we should generalize the results. It may be the case, for example, that significant treatment effects in a survey experiment only generalize to the subset of people in the natural world who are highly attentive or those who are the most accepting of new information. We outline our expectations in greater detail below, but first we describe the news event that makes our study possible.

The Medicare Trust Fund Warning and Hypotheses

Each spring, the board of trustees overseeing Medicare releases its future funding estimates. Unlike past years, the 2007 report was likely to trigger a warning calling for special legislation to be introduced once Medicare's finances fell below a specific threshold.[3] In other

[3] The 2006 Medicare Report states: "The Medicare Modernization Act of 2003 requires that the Medicare Report include a determination of whether the difference between total Medicare outlays and dedicated financing sources (such as premiums and payroll taxes) exceeds 45 percent of total outlays…. The Act requires that an affirmative determination in *two* consecutive reports be treated as *a funding warning for Medicare that would*, in turn, require a Presidential proposal to respond…" (emphasis added).

words, a major policy event was poised to increase national attention to Medicare in 2007. We leveraged our advance knowledge of this situation by combining a natural experiment—a pre-post comparison of public knowledge before and after the 2007 trustees' report—with survey experiments that delivered information about Medicare's finances. This design allows us to compare treatment effects from Medicare information provided in the real world with treatment effects from a survey experiment providing the same key facts about Medicare.

From a substantive standpoint, our study contrasts two popular but widely misunderstood programs. 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 treatments counter what are likely to be mistaken beliefs in the minds of many Americans. People are worried but not completely knowledgeable about the relative funding strength of Medicare and Social Security (Shaw and Mysiewicz 2004). In addition to comparing treatment effects in surveys and natural settings, our study illuminates the conditions under which people correct mistaken factual beliefs (Kuklinski et al. 2000).

The treatments consisted of information about the financial status of Medicare. Of course, there are differences between the real world and survey experiment. But, this is the situation that confronts nearly all survey experimenters—and researchers continue to draw inferences about the political world. As Gaines et al. (2007) write, "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" (p. 5). Our analysis attempts to provide an empirical basis for this common practice.

Our first expectation is that there will be a difference in the size of treatment effects in the survey and natural experiment (Hypothesis 1). In particular, we expect treatment effects in

the survey experiments to be significantly larger than comparable effects from our natural experiment. We also consider two additional hypotheses about individual-level traits that influence the extent to which a person will be exposed to, and internalize, naturally occurring treatment messages in the mass media (Zaller 1992). Our second hypothesis predicts that treatment effects in the natural setting will be largest for those reporting high levels of media exposure (Hypothesis 2). 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 the highly attentive in the natural experiment. Such a pattern would provide guidance for interpreting results from survey-based experiments. It would suggest that the typical survey experiment generates effects that are likely to be observed only among the highly attentive in the real world (see Hovland 1959 for a discussion of these issues).

Alternatively, we also consider who is most likely to accept the information that it is Medicare, not Social Security, which is in worse financial shape. A variety of recent studies point to the role played by certainty in the attitude formation process (e.g., Alvarez 1997; Barabas 2004; Peterson 2004). We hypothesize that those who are least certain about their beliefs will be most likely to learn new facts and update their priors (Hypothesis 3).

To sum up, a key unanswered question is how well treatment effects in a survey experiment correspond to those occurring naturally in the real world. The conventional wisdom is that findings from survey experiments are generalizable to the population at large, but this is based on a narrow interpretation of external validity. We venture beyond the representativeness of the subject pool by drawing attention to the treatments themselves and the manner in which they are received.

Research Design, Data, and Methods

The first of two survey experiments makes use of data collected from the Time-sharing Experiments for the Social Sciences (TESS) platform, which is funded by the National Science Foundation. We divided a single TESS study with 1,622 subjects into two parts. Half of the subjects (n=805) were randomly selected to be interviewed by Knowledge Networks (KN) between March 2-10, 2007, about a month *prior* to release of the latest trust fund report. The other half (described later) was interviewed *after* the news event. Subjects in the first cross-section were randomly divided into four groups. The first (n=206, the control condition) was asked a knowledge question about the relative fiscal status of the Medicare and Social Security programs *without* being given any information about the state of the trust funds. Respondents in the remaining treatment conditions were exposed to varying amounts of policy relevant information before answering the knowledge question. 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, while those in the fourth group (n=201) received information about both Medicare and Social Security. Panel A of Figure 1 illustrates the research design for the two-wave cross-sectional study (see the Appendix for a description of the treatments, which are too lengthy to report here).

Insert Figure 1 here.

The primary outcome measure is 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?” Answer choices of “Medicare,” “Social Security,” or “Both programs will exhaust their funds within the same year” were provided in a

random order. The correct answer at the time of the first survey was unambiguously Medicare. According to widely cited future funding estimates, Medicare's trust fund was projected to be exhausted in 2018 compared to the date of 2040 for Social Security. The question asks for the relative status, not specific dates, because the main concern among policymakers is that it is Medicare, not Social Security, which faces its funding dilemma first.[4]

From April 26 to May 3, 2007, and just three days after the release of the trustees' report, the remaining 817 subjects were surveyed. Once again, respondents were randomly assigned to one of four conditions. In the first condition, respondents were asked the knowledge item without being given any information (n=202). Respondents in the second condition (n=190) were given updated information about Medicare; those in the third group (n=214) were provided information about Social Security; and finally, people in the fourth condition were exposed to information about both programs. Once again, all respondents answered the relative date knowledge item.[5]

The primary goal is to compare treatment effects from trust fund information provided in a natural experiment to the treatment effects from a survey experiment. This entails comparing the changes in programmatic knowledge for subjects bracketing the natural

[4] A second study, discussed next, adds another knowledge item on the percentage of benefits people can expect in the future if no changes are made to Medicare. The text was, "If no changes are made to the Medicare program, what do you think will happen? Please indicate the level of benefits everyone will receive if no changes are made to the Medicare program over the next few decades." There were ten answer choices in evenly spaced blocks of ten percent. The correct answer is 79% during both waves. Answers of 70 to 79% or 80 to 89% were counted as correct. We refer to this fact as the "80% figure."

[5] In the 2007 report, the exhaustion dates moved back a year (to 2019 for Medicare and to 2041 for Social Security). As we expected, the Medicare fiscal warning was triggered and received media coverage.

experiment (i.e., Condition 1 vs. Condition 5) to the survey experiment treatment effects (Conditions 1 vs. 2-4 or 5 vs. 6-8).

Our first study makes use of two nationally representative cross-sectional surveys. A superior design, from the standpoint of causal inference, is to conduct the experiment with the same individuals at two time points. Panel B of Figure 1 shows the structure of a second series of survey-based experiments employing panel data. The panel studies were conducted by Polimetrix, another internet survey firm. The first wave of the survey was administered in early-March like the previous study (the survey started on March 1 and finished on March 21). The second wave of the panel survey started on the same day as the KN survey (April 26) and finished in mid-May (May 16). In this analysis, we make use of the 1,050 respondents who completed surveys at both waves. Each of the six conditions has 151 to 202 cases.

Once again, the purpose of the panel study is to compare treatment effects from survey experiments to those resulting from naturally occurring phenomena. This entails comparisons of the control conditions (Conditions 1-3 in wave 1; Conditions 1, 4, and 5 in wave 2) to the treatment conditions (Conditions 4-6 in wave 1; Conditions 2,3, and 6 in wave 2) both within and across time points.[6] Our examination of the KN and Polimetrix surveys will allow us to contribute to the debate over the use of internet surveys (e.g., Berrens et al. 2003). Because the key treatment and outcome questions in the cross-sectional design are repeated in the panel and because the field periods were similar, it is possible to compare findings across studies.[7]

---

[6] Thirty percent of the 1,500 respondents who started at t_1 did not complete a t_2 survey. Attrition was mostly random, but Cond. 1 had more women by t_2 and Cond. 5 had fewer Republicans at t_2 than at t_1.

[7] The two firms differ in their methodology. Knowledge Networks recruits an RDD sample via telephone, and provides Internet access to those in the sample who do not already have it. Polimetrix maintains a

To recap, our quantity of interest is the difference between the treatment effects in the survey experiments and the natural experiment. If the effects are similar, one might conclude that treatments in survey experiments are good proxies for naturally occurring political events, at least on the topic we examine here. However, if the effects differ dramatically across contexts, that might cast doubt upon the external validity of treatments in survey experiments. We also will consider related topics such as the duration of effects, spillover effects from one experiment to another, and compound treatments.

The initial analyses will be conducted with simple comparisons of means and *t*-tests. However, we improve upon the precision of the estimates by constructing maximum likelihood statistical models that control for demographic factors and partisan affiliations. In those analyses we compare the first differences (i.e., the predicted probabilities of reporting the correct answer for any given experimental group compared to the baseline control group, conditional on the covariates included in the model). To provide confidence intervals on these treatment effect estimates, we make use of the *Clarify* statistical program for *Stata* (King, Tomz, and Wittenberg 2000). Missing demographic responses were imputed with the *Amelia II* routine in *R* 2.4.1 to avoid dropping the respondents who did not disclose their level of education or income (King et al. 2001), although failing to impute does not alter our results.

---

national pool of respondents and matches them to nationally representative samples (Rivers 2006). Data analyzed here from both firms were intended to be nationally representative, but the Polimetrix sample includes an oversample of adults age 55 and older. The substantive results do not change when the data are re-weighted to be nationally representative.

Empirical Results

How well do treatments in survey experiments correspond to the real world? To answer that question we must first characterize the "real world." The 2007 trust fund report triggered a funding warning, which was the first time in history this had ever happened. Nevertheless, the event generated a moderate amount of media coverage. This was due in part to the unexpected death of former Russian President Boris Yeltsin as well as the on-going coverage of a scandal involving the administration's firing of U.S. attorneys. Coverage of the trust fund report also might have been muted because the cost estimates changed little from the previous year's figures. Whatever the source of the editorial decisions concerning the newsworthiness of the story (Bennett 2006), there were 48 stories across 42 newspapers or television stations in the week surrounding the Medicare announcement.[8] Therefore, although the coverage had a fair amount of breadth, it was not an event that generated multiple stories in the same news source for several days or weeks. The important point, from our perspective, is that nearly every story led with the Medicare trust fund warning and conveyed the relative date information.[9]

---

[8] Two coders analyzed the transcripts of all major papers, TV networks, and cable sources in Lexis-Nexis Academic Universe and the NewsBank archive. The correlation in the two sets of scores was .97 (kappa = .94) for a randomly selected subset (40%) of the data. The number of stories would have been higher had we included another 48 instances in which the *Associated Press* story was reprinted, often verbatim, in local and regional newspapers.

[9] Here are some examples of the outlets that provided trust fund date information, almost always on both Medicare and Social Security: the *Atlanta Journal Constitution*, the *Austin American-Statesman*, the *Buffalo News*, *Cincinnati Post*, the *Chicago Sun-Times*, the *El Paso Times*, the *Farmington (NM) Daily Times*, *Florida Today*, the *Financial Times* (London), *Investor's Business Daily*, the *Los Angeles Times*, the *Miami-Times*, the *New York Times*, the *Seattle Times*, the *South Florida Sun-Sentinel*, the *Tampa*

Our main dependent variable is knowledge of the relative date information (this item appears in both surveys). In the second study, we also consider another important fact: the percentage of benefits that will be paid if no changes are made to Medicare (see note 4). The correct answer is roughly 80% and coverage of this information was rare. Of the television and newspaper outlets, only CNN provided the 80% figure and this appeared on the day the trustees' report came out. This fact did not appear in the newspapers from our sample.

Assuming that researchers use survey experiments to better understand political events occurring in the real world (Gaines, Kuklinski, and Quirk 2007), the implicit assumption is that treatment effects in the survey experiments will be similar—at least in terms of the direction of the effects—to those observed in the natural experiment. To see whether this is the case, we turn to the analyses, starting with the cross-sectional data from Knowledge Networks.

Study 1 – Cross-Sectional Data at Two Time Points

Figure 2 displays the mean proportion correct in the various conditions. We begin with the untreated control group, which is depicted in the figure as the line ending in open circles. Thirty-two percent of the control subjects provided the correct answer to the question regarding which program would exhaust its trust fund first at t_1 (mean=.32, standard error=.03; 95% confidence interval from .26 to .38). Subjects who received the date information for Medicare (shown in gray squares) did slightly better, at 37%, but as a group they were not statistically different from the controls given the size of the standard error (.03) and associated confidence interval (.30 to .43). The same is true of the condition given only the Social Security date information, which is depicted in the figure as the line with dark triangles. Thirty eight percent

---

Tribune, the *Wall Street Journal*, the *Washington Post*, the *Washington Times*, *USA Today*, *CNN*, *CBS*, *Fox News*, *National Public Radio*, and the *Associated Press*.

(.38, s.e.=.03) of that group knew the correct answer at t_1, which again was not statistically different from the control group. Finally, the condition that received both Medicare and Social Security dates, noted by asterisks, was significantly more likely to provide the correct answer than the control group ($p < .01$; two-tailed *t*-test). Roughly 58% knew the correct answer (mean=.58; s.e.=.03; c.i.=.51 to .65). That difference corresponds to a sizeable 26 percentage point treatment effect during t_1 of the survey experiment for the subjects who received both dates relative to the control group (mean=.26; s.e.=.05, c.i.= .16 to .35).[10]

Insert Figure 2 here.

We only observed one treatment effect, although technically there were three treatment conditions. Evidently, respondents in the other two conditions (i.e., those providing information about Medicare *or* Social Security) were not able to infer the correct answer. At any rate, the 26 percentage point treatment effect we did observe is at least ten points higher than comparable learning effects attributed to media coverage in non-experimental studies (e.g., Druckman 2005).

Figure 2 also depicts responses for comparable groups several weeks later (i.e., in wave 2). The untreated control group moved up four points to 36% correct, (mean=.36, s.e.=.03; c.i.=.29 to .42), but it was not a statistically significant difference relative to the untreated control group at t_1. In other words, there was no appreciable treatment effect in the natural experiment across the two independent cross-sections. Likewise, the percent correct for groups receiving either the Medicare or Social Security dates at t_2 was not statistically different from the percent correct at t_1. Finally, among those receiving date information for both programs at t_2, 58% provided the correct answer. This figure differs from the control group at t_2 ($p < .01$; two-tailed),

[10] Four questions separated the treatment and the dependent measures. Had we asked the knowledge questions immediately after the stimuli, we likely would have observed even larger treatments effects.

but is indistinguishable from the percent correct for subjects receiving the same information at t_1. In other words, the raw, unadjusted treatment effect at t_2 was estimated at 21 percentage points (s.e.=.05), which was similar to the effect found at t_1. We will refine the estimates using statistical models in a moment, but based upon these patterns it seems the effects in the survey experiment were roughly 20 percentage points larger than the effects in the natural experiment.

Statistical Analyses of the Cross-Sectional Data

All of these unadjusted differences in means are confirmed in the probit models reported in Table 1. There are two models and each includes socio-economic factors like education, income, age, gender, race, and partisanship in addition to dummy variable indicators for the treatment groups. The first two rows show the insignificant effects for providing either Medicare or Social Security date information relative to the control group (the omitted category). The same is true for these coefficients at t_2; they have no significant effect. However, providing the dates for both programs increases the likelihood of correct responses at t_1 (coeff.=.73; s.e.=.13; $p <$ .01) and t_2 (coeff.=.50; s.e.=.13; $p < .01$). These effects remain significant even after controlling for significant predictors of knowledge like education, age, gender at t_2, and Republican identifiers at t_1. Thus, despite compositional differences in the conditions that might have emerged during the randomization process, providing information about the relative status of both trust funds helps people learn the correct response at either time point.[11]

[11] In a randomization check, the groups are balanced across all factors except for income, gender, and partisanship in the Knowledge Networks data (Conditions 3 and 7 are wealthier, Condition 4 has more women, and Condition 2 has fewer Republicans than Condition 1). In the Polimetrix data, the exceptions were race and partisanship (Condition 3 has more black respondents and Condition 6 had fewer Republicans than Condition 1). Providing the partial date information (i.e., Medicare or Social Security

Insert Table 1 here.

Since comparison of the probit coefficients is not straightforward, Figure 3 converts the statistical estimates into predicted probabilities. This will allow us to compare predicted effect sizes for the natural experiment as well as the two significant terms in the models reported in Table 1.[12] The average effect is 5.5 percentage points for the natural experiment, but the confidence interval runs from -2.8 to 14.6, rendering it statistically indistinguishable from zero and thus a null effect. In comparison, the t_1 survey-based experimental effect is 27.8 points (c.i.= 18.4 to 37.3) and at t_2 it is 19.4 points (c.i.= 9.9 to 29.0).

Insert Figure 3 here.

Assuming that researchers use survey-based experiments to better understand political events occurring in the real world (Gaines, Kuklinski, and Quirk 2007), the treatment effects in these survey settings are 14 to 22 points "too large" (e.g., the difference of 27.8 minus 5.5 or 19.4 minus 5.5). Were we to rely solely on the conclusions of the survey-based experiment, we would be overly optimistic about the possibilities for educating the public. This has implications for the various organizations that seek to do just this, via information campaigns, adwatches, and other means. The key question is whether the patterns we have documented hold up in our next study.

---

only) served as a successful manipulation check; only subjects who received both dates learned. Finally, multinomial logit analyses confirm the learning patterns if the incorrect responses are separated.

[12] In Table 1, the control groups are omitted categories, so it is not possible to compute a predicted probability without a corresponding coefficient. In Figure 3, the natural experiment predicted probability is from a probit model with the control group at t_2 as the only experimental group dummy variable (as expected, it is insignificant; coeff. = .16; s.e.=.13).

Study 2 – Panel Data

Scholars have questioned whether survey treatment effects persist (Gaines, Kuklinski, and Quirk 2007, 5-6). They also have wondered about compound effects when individuals receive two treatments or whether survey-based treatments spillover to affect other outcomes. To investigate these and related questions, we conducted a second national study in which the same people were questioned before and after the Medicare announcement (see Panel B of Figure 1).

Figure 4 provides a visual depiction of the group means. The control group, depicted again with a hollow circle, shows that the proportion correct at t_1 is .47 (.39 to .54). Two other groups (Conditions 2 and 3) were untreated at t_1 and they also had means that were statistically indistinguishable (at .44 or .46 overall). Conditions 4 and 6 provided the same relative date information as in the first study using the KN data. Here the means for these groups were higher than the control group subjects. For example, the difference in means at t_1 between Condition 1 and Condition 4 was .17 (s.e.=.05; c.i.=.07 to .27). It was .12 (s.e.=.06; c.i.= .01 to .23) between Condition 1 and Condition 6. These correspond to treatment effects of 17 and 12 percentage points at t_1. Condition 5, which received information about what would happen after the Medicare trust fund is exhausted but no date information, had a mean correct response of .51 on the relative date question. Their group mean was higher than the mean for the controls, but not significantly so.

Insert Figure 4 here.

Several interesting patterns emerged in the second wave. First, the group that was untreated at both waves, Condition 1, had no discernable change in the proportion of correct responses. As with Study 1, the inference we draw from this is that the natural intervention of the Medicare trust fund warning and the accompanying media coverage did not affect levels of

knowledge. Once again, we observe large treatment effects in other comparisons. Condition 2, which was untreated at t_1 and treated with the date information at t_2, experienced a significant increase in knowledge. At t_2 the difference in knowledge between this group and Condition 1 (the control) was 24 percentage points. Given the panel format, that means subjects in Condition 2 improved by 26 percentage points from t_1 to t_2 ($p < .01$). Consistent with Hypothesis 1, treatment effects across contexts are not the same. This suggests that survey-based treatments may be too strong relative to the real world phenomena they seek to mimic.

Several other interesting patterns emerge. People in Condition 4, which had been given the date information at t_1, appear to have forgotten it by t_2 when they did not receive any experimental treatment. In other words, their group mean moved from a high of .64 at t_1 to .49 at t_2, which was indistinguishable from the control group (Condition 1) at t_2. On the other hand, Condition 6, which received the date information at t_1 *and* t_2 remained 12 points more informed than the control condition at the second wave, but the treatments did not compound. That is, despite reinforcement of the facts at two time points, Condition 6 subjects did not elevate their performance. Condition 3, which received a *different* type of information (i.e., the 80% figure) had an elevated level of correct responses on the relative date question (mean=.54, s.e.=.04). The difference between this group and the control was only marginally significant ($p < .07$), though we will return to this particular finding in a moment.

Learning an Alternative Medicare Fact

During the second wave, Condition 3, like Condition 5 at wave 1, was given information about what would happen if no changes were made to the Medicare trust fund (the 80% figure). While it is not depicted in Figure 4, people in both conditions were significantly more likely to provide a correct response relative to the control group. At t_1, very few respondents in the control

group knew the 80% figure. Only 12% of the subjects were able to provide the correct response, which was not much different than guessing given the number of possible answer choices on this item. Nearly 36% of those in Condition 5 provided the correct answer at t_1 (mean=.36, s.e.=.04), which was higher than the controls by 24 points (mean=.24 with an interval of .15 to .33). No other group showed increases in this type of knowledge during t_1 and none were expected.

At t_2, Condition 3 significantly increased their knowledge of the 80% figure to .29 (s.e.=.03). This is a 16 point increase relative to the control group at t_2 (mean=.16, s.e.=.04; $p <$ .01). It is also a 13 percentage point improvement for these same individuals compared to where they were at t_1. Interestingly, Condition 5 reverted back to the level of the control group. In other words, Condition 5 received enough information to increase their percent correct to 36% at t_1, but they reverted 20 points back to a level of 16% correct by wave 2, rendering their t_2 responses no different than the controls. This means the survey treatment effects dissipated a few weeks later even though it was possible to treat a different group (Cond. 3) at this same point in time.[13]

Statistical Analyses of the Panel Data

Table 2 presents the analytical results using the panel data on two forms of Medicare policy-specific knowledge. The first three columns show the results for a model in which the dependent variable is knowledge of the relative date information at t_1 (column 1), t_2 (column 2), and over time as a change score (column 3). Taking the results in that same order, the raw differences in means reported in Figure 4 are confirmed in the probit model. As expected, only the coefficients for providing the date information at t_1 were positive and significant relative to the untreated control group (coefficients of .44 and .35 with standard errors of .14 and .15

[13] In contrast, recent experimental research has documented the persistence of factually *incorrect* beliefs (e.g., Bullock 2006; Kuklinski et al. 2000; Todorov and Mandisodza 2004).

respectively; $p < .01$). Consistent with past research in knowledge, education and age are positively related to knowing the correct answer while women and blacks know less ($p < .01$).

Insert Table 2 here.

The second column of estimates (corresponding to t_2) also confirms the patterns depicted in Figure 4. The coefficient for the group that was given date information at t_2 (Condition 2) is indeed positive and statistically significant (coeff. = .66; $p < .01$) as is the group (Cond. 6) which received the date information during both waves (coeff. = .36; $p < .01$). However, the one "new" finding concerns what might be considered a spillover effect. The group that served as a control group during t_1 but was given the 80% funding information at t_2 showed a positive and significant treatment effect at t_2. It was significant at $p < .07$ in the raw means, but here the .31 coefficient is more than twice the size of the .14 standard error, putting the significance level at $p < .05$ using a two-tailed test. These subjects were not given the relative date information, but they were marginally more likely to get that item correct at t_2. Again, these effects hold even while controlling for a person's level of education, age, gender, and race.

In the analyses presented in the first and second columns, the effects were relative to the untreated control group in each wave (i.e., the omitted condition). In the third column of Table 3, we examine the change in the treatment groups across waves (relative to the change in Condition 1). We expected significant patterns in the case of Condition 2, the group that received nothing in wave 1 but was given the relative date information during wave 2. As the first coefficient in the third column suggests, there is evidence of learning with Condition 2 subjects having more knowledge about Medicare compared to where they were a few weeks prior (coeff. = .58; $p < .01$). The only other significant effect among the groups was movement in the opposite direction. Confirming the pattern we observed in Figure 4, the group that was treated with the date

information during wave 1 and given nothing at wave 2 showed a negative and significant effect. The -.27 coefficient with the .13 standard error means the 195 individuals in this group, as a whole, lost the knowledge they had acquired only a few weeks earlier ($p < .05$). In this case, none of the demographic variables are significant, which indicates that changes in knowledge were roughly equal across the demographic and partisan subgroups.

The last three columns of Table 2 report similar analyses for the second dependent variable (the level of benefits Medicare is expected to pay once the trust fund is exhausted). Confirming what was reported earlier, people receiving the 80% figure in either wave 1 or wave 2 show learning effects in a comparison with the control group (coeff.=.88 in the fourth column or coeff=.63 in the fifth column, both $p < .01$). Likewise, the change scores confirm that the group exposed to 80% figure at t_2 showed evidence of learning (coeff. = .28; $p < .05$), while those exposed to this information at t_1 again seemed to forget the correct response by the time of the second wave (coeff. = -.57; $p < .01$).

Predicted Probabilities

Figure 5 shows the effects of the natural experiment as well as the other statistically significant effects from Table 2. The first entry shows the null effect for the comparison of the control group (Condition 1) respondents over time.[14] The next four entries are the positive treatment effects from providing the relative date information at t_1 (17.3 and 14.2 point effects), at t_2 only (a 25.9 point effect), during both waves (14.4), or expressed as a change score from control at t_1 to date information during t_2 (17.1 point effect). The confidence intervals are depicted in gray shading around the black dot, which corresponds to the first difference of the

[14] To produce this estimate, the referent category was changed to Condition 6, which received treatments during both waves but did not change significantly across the waves. See note 12 for rationale.

predicted probabilities using the coefficients in Table 2. The second to last value corresponds to the spillover effect documented earlier (roughly twelve percentage points). The final value shows the decay effect. Respondents who were given dates at t_1 but not treated during t_2 reported fewer correct answers by about 5 percentage points. With the exception of the decay effect, learning effects induced by a survey experiment are dramatic. While the size of these treatment effects varies, they are all substantially larger than the learning effects occurring in the natural world.[15]

Insert Figure 5 here.

Panel B of Figure 5 shows the predicted treatment effects for the second dependent variable. Once again, the effect of the natural experiment is estimated to be zero. In contrast, the survey treatment during t_1 produces a 16 percentage point effect, while at t_2 the effect is estimated to be 13 percentage points. As a change score, the predicted treatment effect is about half as large, at 7 percentage points. There also was a -9 percentage point decay effect (s.e. = .03; c.i. = -15 to -4). No matter how we examine the data (e.g., t_1, t_2, or as a change from t_1 to t_2), the treatment effect is much larger than what really occurred based upon our analysis of the people in Condition 1. In this instance, treatment effects induced by a survey experiment are too large by almost 20 points (the average of the positive treatment effects across both panels).

Consistent with Hypothesis 1, there is a discrepancy in the direction of treatment effects across the two contexts—with positive and statistically significant effects in the survey experiment and null effects in the natural experiment. One explanation for this discrepancy is the difference in the length of time that elapses between exposure to the stimulus and measurement of the outcome variable (i.e., time until measurement). In the survey experiment, four questions

[15] Many of the survey experiment effects are also larger than the effects of gender (-11.5) or race (-12.4), but on par with education (+18) and smaller than the effects of age (+45 points from sample min to max).

separated the treatment from the dependent measure. In the natural world, time until measurement must be represented in terms of days, not minutes. If it is this difference—and not the varying strength of the treatments across contexts—that contribute to the patterns we have observed, we should see large effects in the natural world that diminish over time. We explored this possibility with the Polimetrix data. Here the wave 2 respondents were randomly assigned to take the survey either one, two, or three weeks after trustees' announcement. When we examine the size of the treatment effects in the natural setting across the three time intervals, there is no evidence of decay. The effect of the trustees' announcement was small, even in the immediate aftermath of the report's release.

Of course, one may still object that differences across the two contexts have the inevitable effect of making the stimuli different (e.g., Kinder 2007). It would be perilous to push this argument too far, however, for that challenges the rationale for doing survey experiments. It is reasonable, though, to wonder whether the particular facts we chose to study were too hard or obscure. If so, then it might be difficult to absorb the relevant information in the natural world, at least in comparison to the sterile environment of the survey experiment. There are several reasons to doubt this played a role in our study. For starters, facts about the trust funds (e.g., the exhaustion dates and funding information) often are the *only* part of the trustees' report that receives media coverage. To confirm this, we coded the news coverage of the trustees' report for discussion of *any* fact, not just the two we studied here. In a search of Lexis-Nexis and NewsBank, the exhaustion dates and funding information were mentioned at least twice as often as other facts pertaining to these two programs. Additionally, coverage of the trustees' report in 2007 was typical compared to past years. During the period from 1997 to 2006, the *New York Times* covered the spring trust fund announcement in about two stories (with a range of 1 to 3

stories). In 2007, the *New York Times* again devoted two stories to the event, indicating that coverage was not unusually low or high during the period we examined.

We believe, then, that the discrepancy in the magnitude of treatment effects across our natural and survey experiment is meaningful (as opposed to artifactual), and worthy of scholarly attention. We now turn to an idea that was raised by Hovland (1959) writing almost a half century ago. At the time, he was attempting to reconcile contradictory findings about attitude change from experimental and survey studies. He pointed to the important role played by exposure and other factors, such as a person's "commitment" to an issue.

Effects Based upon the Likelihood of Exposure and Acceptance

Our second and third hypotheses state that treatment effects in the natural world might be especially large among particular subpopulations. In particular, Hypothesis 2 directs our attention to the highly aware, and predicts that treatment effects in the natural world (i.e., Condition 1) should be largest among this subpopulation. That prediction is confirmed according to the first entry in the first column of Table 3.

Insert Table 3 here.

The variable *Key Fact Was in R's Source* is the most detailed measure of exposure we could create. Each respondent in the Polimetrix study was asked which media source they used most.[16] Drawing upon the media content analysis described earlier, we created a term that was

[16] 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. Respondents in the control group named 17 different media sources (consisting of regional and national outlets); roughly half of these provided the relative date information.

scored as one if each respondent's news source mentioned the Social Security *and* Medicare dates (i.e., they were exposed to the relative date information like the subjects in the experiment). All other responses were given a zero (and in the second column of Table 3 it is coverage of the 80% figure that receives a one, zero otherwise). As the .58 coefficient on the *Key Fact* term indicates, respondents who reported using media sources that supplied the relative date information were more likely than others in the control group to learn from t_1 to t_2 ($p < .05$). This is consistent with Hypothesis 2 and it translates into a 12 point change in the probability of learning the fact.

Our third hypothesis states that those who were least certain about the Medicare fact in question would be the most likely to take in the new information. The first column of Table 3 also supports that prediction. Individuals who reported being certain about their Medicare beliefs were *less* likely to change from an incorrect to correct response in the second wave (coeff. = -.56, s.e.= .34; $p < .10$).[17] For those who were *least* certain, the change in the predicted probability of reporting a correct response at t_2 increased by 10 percentage points. Thus, some individuals within our natural experiment learned, though the size of these learning effects remain about half the size of the average treatment effect from the survey experiment.

We also had to dig fairly deep to find these effects. The remaining entries in the first column of Table 3 show that every other variable we tried as a proxy for likely exposure or acceptance (as well the demographic and partisanship terms) is insignificant. That is, those who report following Medicare in the news, discussing Medicare, or hearing about the trust fund report are no more likely to learn. The same holds for respondents high in policy-specific knowledge as well as those who report using various sources (television, newspapers, or the

[17] Belief certainty was assessed at t_1 on a four point scale from 0 (not at all certain) to 1 (very certain).

internet).[18] One final group of people might be particularly attuned to news stories about Medicare: those who were receiving Medicare benefits at the time of our survey (which includes retirees as well as millions of disabled workers and their families). Using a question that asked respondents whether they currently were covered by Medicare, we examined this possibility. We found no evidence that being a Medicare recipient enhanced learning effects, either in the natural world (Condition 1) or in any of the other conditions (i.e., no main or interactive effects).[19]

Although our main interest is in the control group respondents (i.e., those excluded from the survey experiment), we re-estimated the model from Table 3 on the entire sample and found similar results (e.g., the coefficient for *Key Fact* was positively signed while the coefficient for *Prior Certainty* was negatively signed; $p < .01$ for both). Most importantly, we observed the same pattern of results for our five condition indicators, further underlining the robustness of the findings in Table 2. Consistent with our second and third hypotheses, the results suggest that the findings from our survey experiments generalize, but only to the population of people who were exposed to the same facts in *their* news source and to those who were not certain of their factual beliefs on the subject of Medicare.

---

[18] The endpoints of the variables were coded in the following manner: *Follows Medicare in the News* (1= very closely), *Discusses Medicare* (1=discusses Medicare with neighbors, family members, or co-workers), *Heard/Read Trust Fund Report* (1=heard or read a lot); *Policy-Specific Knowledge* (8=knows eight Medicare and Social Security facts), *Main Source* variables (1=respondents main source is television, newspaper, or the internet relative to the omitted category of radio/other). The correlation between the various exposure items is minimal (the average *r* is .03). The variables also are statistically insignificant when entered in the models in Table 3 one at a time.

[19] Additionally, respondents who had recently received a benefit statement from the Social Security Administration were neither more nor less knowledgeable than other respondents.

We did not receive support for our expectations when it came to the second dependent variable concerning the percentage of benefits Medicare recipients can expect if no changes are made (column 2). Based upon the content analysis we found that the 80% figure was mentioned by CNN but not by any other source. Perhaps because of the low level of news coverage, both the *Key Fact* term and *Prior Certainty* are insignificant in the second model. As was the case with the relative date analysis (column 1), no other predictors explain learning.

Our primary goal was to examine the correspondence between treatment effects in natural and survey experiments, but our results also have implications for theories of public opinion. When the media covered key facts from the trustees' report (e.g., the relative date information), the most attentive absorbed the information, lending some support to notions of a rational public (Page and Shapiro 1992). The fact that we measured knowledge at two points in time and administered multiple treatments to a subset of our respondents allows us to dig beneath the surface. When we do, the picture is less optimistic. As we noted earlier, respondents in the survey experiment rapidly forgot the relative date information, and knowledge did not seem to accumulate for respondents who received the treatment twice. The implications are clear: if policy-specific information is vital to the formation of well-grounded opinions (Gilens 2001), this information must appear repeatedly in multiple formats in the mass media.

Discussion

For every instance in which one might logically expect a survey-based information treatment effect, we found one. Also, those effects were much larger than corresponding estimates from a natural experiment on the same topic with the same outcome measures (in support of Hypothesis 1). This finding was robust to specification changes, types of data (cross-sectional or panel), methods of subject recruitment (Knowledge Networks versus Polimetrix),

and it even appears using alternative forms of the dependent variable (i.e., the relative dates vs. the 80% funding figure).[20] We also found support for Hypotheses 2 and 3, which suggests that the typical survey experiment generates effects that are likely to be observed among the highly attentive or those without strong prior beliefs about a policy.

Having said all that, our concentration on one issue—Medicare—raises its own concerns. We opted to study a single issue in depth, but our conclusions will carry more weight as scholars compare natural and survey experiments in other issue areas and with different outcomes (e.g., opinion questions).[21] Past research has addressed related topics such as the duration of effects (e.g., Druckman and Nelson 2003; Mutz and Reeves 2005; Norris and Sanders 2003) and scholars have employed designs using panel studies and experiments (Druckman and Holmes 2004). We are aware of no other study that explores the external validity of survey treatments.

Let us be clear: we do not argue that scholars should abandon the use of survey-based experiments. That hardly seems likely anyway, given the expansion of internet surveys and multi-investigator studies such as TESS. Survey experiments are valuable because they help researchers discern causal effects using nationally representative samples (see Kam, Wilking, and Zechmeister 2007 for a recent discussion). We also believe that the effects observed in the scores of studies using survey experiments are real. At the same time, this study shows the

[20] We did not report weighted estimates, but the results do not change when weights are employed or if, in the case of the panel data, we use a lagged dependent variable instead of the change score approach.

[21] In preliminary analyses not reported here, the effects we observe on the political knowledge outcome variables are also seen with corresponding attitudinal measures (e.g., confidence in Medicare, perceptions of whether the program is in a fiscal crisis, and policy opinions on Medicare privatization reforms). We observe sizable and statistically significant effects in the survey experiment conditions, but no corresponding changes in the natural world comparison.

importance of considering the correspondence between survey treatments and the political stimuli they seek to emulate. As one scholar recently observed, "[m]any survey experiments attempt to mimic some experience from everyday life, such as the acquisition of information from the media or a social encounter with a more or less friendly stranger. But these efforts are often crude simulacra of the real-life experiences of interest" (Gilens 2002, 248; see McDermott 2002b for a related discussion regarding lab experiments).[22]

It is true that experiments "need not be isomorphic with a naturally occurring (i.e., a real world) referent" (Druckman et al. 2006, 634). This is especially the case for laboratory experiments whose primary goal is theory testing (e.g., Quattrone and Tversky 1988; see Brader 2005, 402 or Guala and Mittone 2005 for discussion). But this has not been the sole purpose of survey experiments. As several scholars have observed, political scientists use often survey experiments to illustrate how people respond to real world political stimuli (campaign commercials, arguments for a policy, facial displays, and so on). In this context, experimental findings—and interpretations of them—take on added significance (Gaines et al. 2007). Until recently, however, researchers have not focused on how one translates significant experimental findings into statements about the effects of similar treatments in the real world.

---

[22] As McDermott (2002a, 39-40) writes, "For political scientists, questions surrounding external validity pose the greatest concern with experimentation. What can experiments tell us about real-world political phenomena? Beyond the nature of the subject pool, this concern is at least twofold. First, in the laboratory it is difficult to replicate key conditions that operate on political actors in the real world….Second, and related, many aspects of real-world complexity are difficult to simulate in the laboratory…Failure to mimic or incorporate these constraints into experiments, and difficulty in making these constraints realistic, might restrict the applicability of experimental results to the real political world." McDermott argues that the realism of an experiment is essential; mundane similarities with the real world are not.

The conclusion we draw from our examination of the Medicare trust fund announcement is that the typical survey experiment does not correspond well to news events receiving low to moderate amounts of media coverage. In contrast, we suspect that the typical treatment in a survey experiment approximates news events that receive a substantial amount of coverage—though here we can only speculate because the particular facts we examined did not receive extensive coverage. Our findings have broader implications as well. For those studying how levels of knowledge change in response to media coverage of actual news events (e.g., Jerit, Barabas, and Bolsen 2006; Nicholson 2003; Price and Zaller 1993), our results are a reminder that even seemingly important events might not lead to appreciable gains in knowledge. In that sense, previously documented learning effects that are based upon events receiving unusually high levels of media coverage (e.g., campaigns, high profile policy debates) may not be typical.

Researchers can do little to alter the fundamental tradeoff between internal validity and the realism of an experiment. Happily, the continued use of survey experiments offers the best hope of dealing with this challenge. Because no single study maps perfectly on to the real world, the accumulation of findings across multiple experiments—each sensitive to the context of the study and the interpretation one draws from it—seems like the most fruitful path to take.

Appendix

Survey Experiment Treatments

Introduction that All Treatment Conditions Received:

"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:"

Medicare Date Information Treatment:[23]

BEFORE THE 2007 REPORT: "According to the trustees who oversee the Medicare program, the trust fund will run out of money in **2018**."

AFTER THE 2007 REPORT: "According to the trustees who oversee the Medicare program, the trust fund will run out of money in **2019.**"

Social Security Date Information Treatment

BEFORE THE 2007 REPORT: "According to the trustees who oversee the Social Security program, the trust fund will run out of money in **2040**."

AFTER THE 2007 REPORT: "According to the trustees who oversee the Social Security program, the trust fund will run out of money in **2041**."

[23] We realize that terms like "run out of money" or "exhausted" or even "financial problems" may lead citizens to misleading inferences (Jerit and Barabas 2006), however these terms are used in the media and by policymakers to describe the situation.

Medicare and Social Security Date Information Treatment:[24]

> "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 **2019**. Social Security also faces financial problems, but the trust fund for Social Security is projected to be exhausted in the year **2041**."

Medicare Funding Information if No Changes are Made:[25]

> "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. What this means is that if no changes are made over the next few decades, future beneficiaries will receive roughly **80** percent of promised benefits because payroll taxes will still be available to pay for most costs."

[24] At wave 1, the estimates were 2018 and 2040 based upon the latest information available.

[25] The percentage of benefits that Medicare can pay after exhaustion remained at 79% in both waves.

References

Alvarez, R. Michael. 1997. *Information and Elections.* Ann Arbor, MI: Univ. of Michigan Press.

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

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

Berinsky, Adam J., and Tali Mendelberg. 2005. "The Indirect Effects of Discredited Stereotypes in Judgments of Jewish Leaders." *American Journal of Political Science* 49: 845-64.

Berrens, Robert P., Alok K. Bohara, Hank Jenkins-Smith, Carol Silva, and David L. Weimer. 2003. "The Advent of Internet Surveys for Political Research: A Comparison of Telephone and Internet Samples." *Political Analysis* 11 (Winter): 1-22.

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

Bullock, John. 2006. "The Enduring Importance of False Political Beliefs." Unpublished Manuscript, Stanford University.

Burden, Barry C., and Casey A. Klofstad. 2005. "Affect and Cognition in Party Identification." *Political Psychology* 26 (6): 869-86.

Chong, Dennis, and James N. Druckman. 2008. "Framing Public Opinion in Competitive Democracies." *American Political Science Review* 101 (November): 637-56.

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

Druckman, James N. 2004. "Political Preference Formation: Competition, Deliberation, and the (Ir)relevance of Framing Effects." *American Political Science Review* 98: 671-686.

Druckman, James N. 2005. "Media Matter: How Newspapers and Television News Cover Campaigns and Influence Voters." *Political Communication* 22: 463-81.

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 (Nov.): 627-35.

Druckman, James N., and Justin W. Holmes. 2004. "Does Presidential Rhetoric Matter? Priming and Presidential Approval." *Presidential Studies Quarterly* 34 (June): 755-78.

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

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

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, pp. 232-250.

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

Grant, J. Tobin, and Thomas J. Rudolph. 2003. "Value Conflict, Group Affect, and the Issue of Campaign Finance." *American Journal of Political Science* 47(July): 453-69.

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

Guala, Francesco, and Luigi Mittone. 2005. "Experiments in Economics: External Validity and the Robustness Phenomena." J*ournal of Econometric Methodology* 12 (Dec.): 495-515.

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

Iyengar, Shanto, and Donald Kinder. 1987. *News that Matters*. Chicago: Univ. of Chicago Press.

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, Jason Barabas, and Toby Bolsen. 2006. "Citizens, Knowledge, and the Information Environment." *American Journal of Political Science* (Apr.) 50: 266-82.

Kam, Cindy D. 2007. "When Duty Calls, Do Citizens Answer?" *Journal of Politics* 69: 17-29.

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.

Kinder, Donald. 2007. "Curmudgeonly Advice." *Journal of Communication* 57 (Mar.): 155-62.

Kinder, Donald R., and Thomas R. Palfrey, eds. 1993. *Experimental Foundations of Political Science*. Ann Arbor, MI: University of Michigan Press.

King, Gary James, Honaker, Anne Joseph, and Kenneth Scheve. 2001. “Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95 (Mar.): 49-69.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. ``Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44 (Apr.): 347-61.

Krosnick, Jon A., and Howard Schuman. 1988. “Attitude Intensity, Importance, and Certainty and Suceptibility to Response Effects.” *Journal of Personality and Social Psychology* 54 (6): 940-52.

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 (Aug.): 790-816.

Lassen, David Dreyer. 2005. “The Effect of Information on Voter Turnout: Evidence from a Natural Experiment.” *American Journal of Political Science* 49 (Jan.): 103-88.

McDermott, Rose. 2002a. “Experimental Methods in Political Science.” *Annual Review of Political Science* 5: 31-61.

McDermott, Rose. 2002b. “Experimental Methodology in Political Science.” *Political Analysis* 10 (Autumn): 325-42..

Marmor, Theodore R. 2000. *The Politics of Medicare*, 2nd ed. NY: Aldine de Gruyter.

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 (Feb.): 1-15.

Nelson, Thomas E., Rosalee A. Clawson, and Zoe M. Oxley. 1997. “Media Framing of a Civil Liberties Conflict and Its Effect on Tolerance.” *American Political Science Review* 91 (Sept.): 567-83.

Nicholson, Stephen P. 2003. “The Political Environment and Ballot Proposition Awareness.” *American Journal of Political Science* 41 (July): 403-10.

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.

Page, Benjamin I., and Robert Y. Shapiro. 1992. *The Rational Public: Fifty Years of Trends in Americans' Policy Preferences*. Chicago: University of Chicago.

Peffley, Mark, Jon Hurwitz, and Paul M. Sniderman. 1997. "Racial Stereotypes and Whites' Political Views of Blacks in the Context of Welfare and Crime." *American Journal of Political Science* 41 (Jan.): 30-60.

Peterson, David A. M. 2004. "Certainty or Accessibility: Attitude Strength in Candidate Evaluations." *American Journal of Political Science* 48 (July): 513-20.

Price, Vincent, and John Zaller. 1993. "Who Gets The News? Alternative Measures of News Reception and Their Implications for Research." *Public Opinion Quarterly* 57: 133-64.

Quattrone, George A., and Amos Tversky. 1988. "Contrasting Rational and Psychological Analyses of Political Choice." *American Political Science Review* 82 (Sept.): 719-36.

Rivers, Douglas. 2006. "Sample Matching: Representative Sampling from Internet Panels." Polimetrix White Paper Series. Available at: http://www.polimetrix.com/documents/ Polimetrix_ Whitepaper_ Sample_ Matching.pdf"

Schram, Arthur. 2005. "Artificiality: The Tension between Internal and External Validity in Economics Experiments." *Journal of Economic Methodology* 12 (June): 225-37.

Schuman, Howard, and Lawrence Bobo. 1988. "Survey-based Experiments on White Racial Attitudes toward Racial Segregation." *American Journal of Sociology* 94 (Sept.): 273-99.

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 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. Saris and Sniderman. Princeton, NJ: Princeton.

Todorov, Alexander, and Anesu N. Mandisodza. 2004. "Public Opinion on Foreign Policy: The Multilateral Public that Perceives Itself as Unilateral." *Public Opinion Quarterly* 68 (3): 323-348.

Zaller, John. 1992. *The Nature and Origins of Mass Opinion.* New York: Cambridge.

Table 1. Probit Analysis of Cross-Sectional Data on Medicare Knowledge

	Medicare Exhausted Before Social Security	
	Time 1	Time 2
Treatment: Medicare Exhaustion Date	.18	-.04
	(.13)	(.13)
Treatment: Social Security Exhaustion Date	.18	.15
	(.13)	(.13)
Treatment: Medicare and Social Security Dates	.73 ***	.50 ***
	(.13)	(.13)
Education	.83 ***	.93 ***
	(.24)	(.23)
Income	-.03	.33
	(.22)	(.21)
Age	.91 ***	1.28 ***
	(.21)	(.22)
Female	-.29 ***	-.19 **
	(.09)	(.09)
Black	-.22	-.03
	(.17)	(.16)
Democrat	-.26	-.04
	(.19)	(.18)
Republican	-.74 ***	-.08
	(.20)	(.18)
Constant	-.74 ***	-1.23 ***
	(.24)	(.22)
Number of cases	805	817

Note: Probit coefficients are reported in the table. Standard errors are in the parentheses. The "Medicare Exhausted Before Social Security" dependent variable is coded as 1 for the correct answer that Medicare will exhaust its trust fund before Social Security, zero otherwise. The omitted group in each wave was a control group which was not treated. All independent variables have been rescaled on a zero to one interval to facilitate interpretation with the coding for the highest category as follows: *Treatment: Medicare Exhaustion Date* (1=experimental Condition 2 at t_1 or Condition 6 at t_2), *Treatment: Social Security Exhaustion Date* (1=Condition 3 at t_1 or Condition 7 at t_2), *Treatment: Medicare and Social Security Dates* (1=Condition 4 at t_1 or Condition 8 at t_2), *Education* (1=doctoral degree), *Income* (1=$175,000 or more), *Age* (1=94 years old), *Female* (1=woman), *Black* (1=African-American), *Democrat* (1=Democratic identifier), and *Republican* (1=Republican identifier). See text for details on samples and estimation techniques.
*** $p < .01$; ** $p < .05$; * $p < .10$ (two-tailed).

Table 2. Probit Analysis of Panel Data on Medicare Knowledge

	Medicare Before Soc. Sec.			Medicare Still at 80% Funding		
	Time 1	Time 2	Change Score	Time 1	Time 2	Change Score
Control Group T_1, Dates T_2	-.08	.66 ***	.58 ***	.25	-.05	-.18
	(.14)	(.14)	(.13)	(.17)	(.17)	(.13)
Control T_1, Funding Info. T_2	.04	.31 **	.21	.28	.63 ***	.28 **
	(.14)	(.14)	(.13)	(.18)	(.16)	(.14)
Dates T_1, Control T_2	.44 ***	.09	-.27 **	.26	-.03	-.18
	(.14)	(.14)	(.13)	(.17)	(.17)	(.13)
Funding Info. T_1, Control T_2	.12	.09	-.02	.88 ***	.12	-.57 ***
	(.14)	(.14)	(.13)	(.17)	(.17)	(.14)
Dates T_1, Dates T_2	.35 ***	.36 ***	.00	.12	.20	.06
	(.15)	(.15)	(.13)	(.19)	(.18)	(.14)
Education	.46 ***	.62 ***	.11	.30 *	.04	-.17
	(.15)	(.16)	(.14)	(.18)	(.18)	(.15)
Income	-.11	-.08	.02	-.13	-.05	.06
	(.17)	(.18)	(.16)	(.20)	(.21)	(.17)
Age	1.23 ***	1.25 ***	-.01	.90 ***	1.04 ***	.00
	(.20)	(.21)	(.18)	(.24)	(.25)	(.20)
Female	-.29 ***	-.29 ***	.01	-.23 ***	-.20 **	.03
	(.08)	(.08)	(.07)	(.10)	(.10)	(.08)
Black	-.32 ***	-.58 ***	-.19	-.24	-.03	.13
	(.13)	(.14)	(.12)	(.16)	(.16)	(.13)
Democrat	.05	.05	-.01	.48 ***	.30 **	-.14
	(.11)	(.11)	(.10)	(.14)	(.14)	(.11)
Republican	-.08	-.08	.00	.34 ***	.17	-.11
	(.11)	(.11)	(.10)	(.15)	(.14)	(.11)
Constant/Cut Point 1	-.55 ***	-.67 ***	-.95 ***	-1.88 ***	-1.63 ***	-1.43 ***
	(.19)	(.20)	(.18)	(.24)	(.24)	(.19)
Constant/Cut Point 2			1.10			1.07 ***
			(.18)			(.19)
Number of Cases	1,050	1,050	1,050	1,050	1,050	1,050

Note: Probit coefficients, or ordered probit in the case of the change scores, are reported in the table. Standard errors are in the parentheses. The "Medicare Before Soc. Sec." dependent variable is coded as 1 for the correct answer that Medicare will exhaust its trust fund before Social Security, zero otherwise. The "Medicare Still at 80% Funding" dependent variable is knowledge that Medicare will be able to provide roughly 80% of benefits (includes answers between 70% to 89%) even if no changes are made, zero otherwise. The omitted group is experimental Condition 1, the control group, which was not treated during either wave of the survey-based experiment. All independent variables have been rescaled on a zero to one interval to facilitate interpretation with the coding for the highest category as follows: *Group T_1, Dates T_2* (1=experimental Condition 2), *Control Group T_1, Funding Information T_2* (1=Condition 3), *Dates T_1, Control T_1* (1=Condition 4), *Funding Information T_1, Control T_2* (1=Condition 5), *Dates T_1, Dates T_2* (1=Condition 6), *Education* (1=post-graduate), *Income* (1=$150,000 or more), *Age* (1=99 years old), *Female* (1=woman), *Black* (1=African-American), *Democrat* (1=Democratic identifer), or *Republican* (1=Republican identifier).

*** $p < .01$; ** $p < .05$ (two-tailed); * $p < .10$ (two-tailed).

Table 3. Models of Exposure to and Acceptance of Medicare Facts in the Panel Data

	Medicare Before Soc. Sec.	Medicare Still at 80% Funding
Key Fact Was in R's Source	.58 **	-.19
	(.27)	(.41)
Prior Certainty of Factual Belief	-.56 *	-.15
	(.34)	(.40)
Follows Medicare in News	.08	.15
	(.44)	(.50)
Discusses Medicare	-.21	-.21
	(.25)	(.28)
Heard or Read about Report	-.02	-.70
	(.39)	(.46)
Policy-Specific Knowledge	-.02	-.01
	(.06)	(.06)
Main Source: Television	-.24	-.08
	(.30)	(.33)
Main Source: Newspaper	-.46	.59
	(.37)	(.39)
Main Source: Internet	.28	.11
	(.27)	(.31)
Education	-.18	-.25
	(.40)	(.46)
Income	.04	.03
	(.48)	(.59)
Age	.13	.48
	(.58)	(.68)
Female	-.14	.36
	(.20)	(.24)
Black	-.06	-.15
	(.39)	(.46)
Democrat	.24	-.07
	(.27)	(.31)
Republican	.19	.29
	(.28)	(.32)
Constant/Cut Point 1	-1.26 **	-1.41 **
	(.60)	(.65)
Constant/Cut Point 2	.84	1.49 **
	(.60)	(.66)
Number of Cases	163	163

Note: The entries are ordered probit coefficients (with standard errors in the parentheses). The dependent variable is the change in learning from t_1 to t_2. *** $p < .01$; ** $p < .05$; * $p < .10$ (two-tailed).

Figure 1. A Natural Experiment Bracketed by Survey Experiments

Panel A. Two-Wave Cross-Sectional Design (TESS/Knowledge Networks Data)

Condition 1:		O		
Condition 2:	$X_{\text{Medicare Date Info. '06}}$	O		
Condition 3:	$X_{\text{Soc. Sec. Date Info. '06}}$	O		
Condition 4:	$X_{\text{Med. \& Soc. Sec. Date Info. '06}}$	O		
		2007 Medicare and Social Security Trustees' Report		
Condition 5:				O
Condition 6			$X_{\text{Medicare Date Info. '07}}$	O
Condition 7:			$X_{\text{Soc. Sec. Date Info. '07}}$	O
Condition 8:			$X_{\text{Med. \& Soc. Sec. Date Info. '07}}$	O

Panel B. Panel Design (Polimetrix Data)

Condition 1:		O		O
Condition 2:		O	$X_{\text{Medicare \& SS Date Info. '07}}$	O
Condition 3:		O	$X_{\text{Medicare Funding Info. '07}}$	O
Condition 4:	$X_{\text{Medicare \& SS Date Info. '06}}$	O		O
Condition 5:	$X_{\text{Medicare Funding Info. '06}}$	O		O
Condition 6:	$X_{\text{Medicare \& SS Date Info. '06}}$	O	$X_{\text{Medicare \& SS Date Info. '07}}$	O
		2007 Medicare and Social Security Trustees' Report		

Early-April 2007 | April 23, 2007 | Late-April to mid-May 2007

Time

X=survey experiment treatment
O=survey observation

Note: Condition 1 in both experiments and Condition 5 in Panel A are control groups that do not receive any treatments. See the text and the Appendix for more details on the content of the informational messages.

Figure 2. Two-Wave Cross-Sectional Data on Knowledge that Medicare Will Be Finacially Exhausted before Social Security

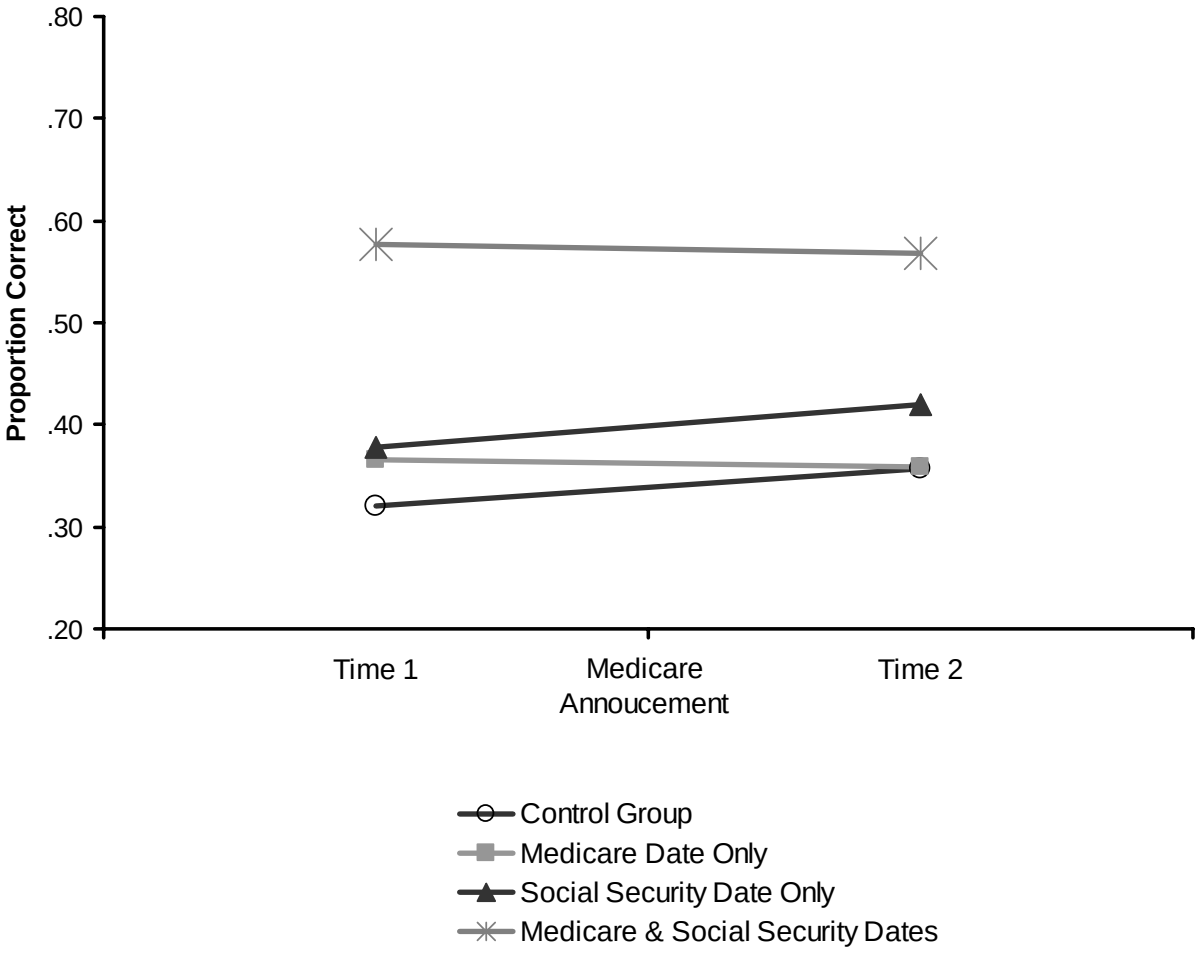


Figure 3. Natural and Survey Experiment Treatment Effects in the Knowledge Networks Cross-Sectional Data

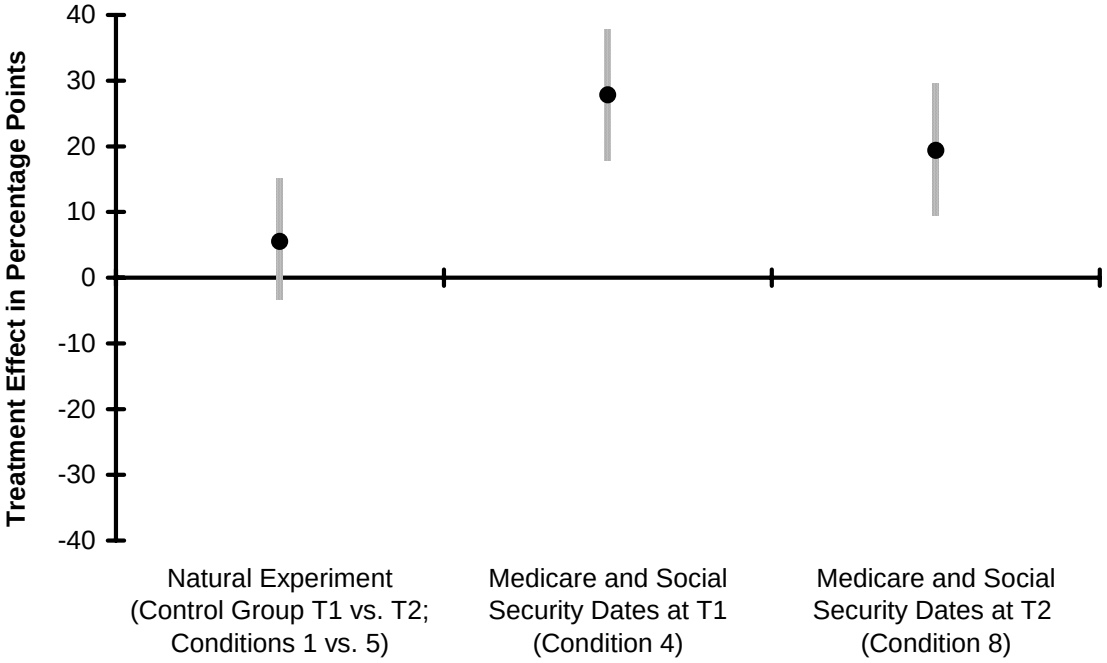
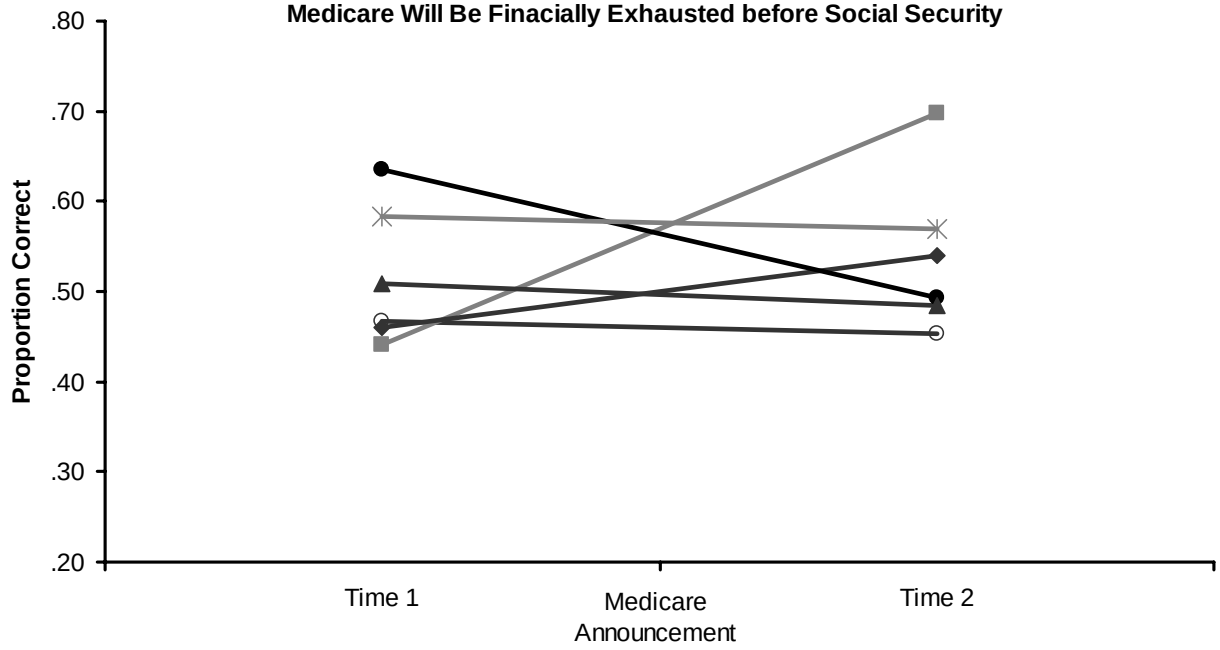


Figure 4. Panel Data on Knowledge that Medicare Will Be Finacially Exhausted before Social Security



Condition 1 (Control Group at Wave 1 and Wave 2)

Condition 2 (Wave 1 Control Group, Wave 2 Medicare and Social Security Dates)

Condition 3 (Wave 1 Control Group, Wave 2 Funding Information)

Condition 4 (Wave 1 Medicare and Social Security Dates, Wave 2 Control Group)

Condition 5 (Wave 1 Funding Information, Wave 2 Control Group)

Condition 6 (Medicare and Social Security Dates at Wave 1 and Wave 2)

Figure 5. Natural and Survey Experiment Treatment Effects in the Polimetrix Panel Data