Accounting for Noncompliance in Survey Experiments

Supplemental Appendix

Jeffrey J. Harden* Anand E. Sokhey† Katherine L. Runge‡

December 6, 2018

Abstract

This appendix contains supplementary material designed to assist researchers who wish to address noncompliance in their survey experiments. In the first section, we define key terms and discuss the methodological details of noncompliance as they relate to survey experiments. In the second, we further elaborate on our proposed methods for measuring noncompliance, list the causal estimands available with these approaches, and provide more details on our meta analysis of survey experiments in published articles in political science. In the third section, we present replications of published survey experiments using our proposed methods.

Contents

*Associate Professor, Department of Political Science, University of Notre Dame, 2055 Jenkins Nanovic Halls, Notre Dame, IN 46556, jeff.harden@nd.edu.

†Associate Professor, Department of Political Science, University of Colorado Boulder, 333 UCB, Boulder, CO 80309, anand.sokhey@colorado.edu.

‡Graduate Student, Department of Political Science, University of Colorado Boulder, 333 UCB, Boulder, CO 80309, katherine.runge@colorado.edu.

1 Noncompliance in Survey Experiments

Noncompliance can be conceptualized with the potential outcomes framework (see Imbens and Rubin 2015). Consider a survey experiment with one treatment condition and one control condition. A respondent receives treatment if he or she reads and processes the information in the vignette and control otherwise. Define Z_i as a binary variable indicating treatment assignment status for respondent i and D_i as a binary variable with two potential outcomes indicating whether respondent i received the treatment: $D_i(Z_i)$, $Z_i \in \{0,1\}$. The potential outcomes are defined as $Y_i(Z_i, D_i[Z_i])$ and the observed outcome, Y_i, is based on realizations of Z_i and D_i:

$$Y_i = Y_i(Z_i, D_i[Z_i]) = \begin{cases} Y_i(0,0), & \text{if } Z_i = 0, D_i = 0, \\ Y_i(1,0), & \text{if } Z_i = 1, D_i = 0, \\ Y_i(1,1), & \text{if } Z_i = 1, D_i = 1. \end{cases}$$

Survey respondents can be categorized as compliers or noncompliers as shown in Table A1. Compliers receive the condition to which they are assigned. Noncompliers do not follow their assigned status. They may be always-takers or never-takers, who disregard treatment assignment and always or never receive treatment, respectively. Following Angrist, Imbens, and Rubin (1996), we assume monotonicity, which states that assignment to treatment will not *dissuade* respondents from taking it (i.e., $D_i[1] - D_i[0] \geq 0$). This eliminates the possibility of "defiers" who take the opposite condition of their assignment. A researcher conducting a survey experiment can also assume no always-takers, which implies that noncompliance is one-sided. Not all respondents assigned to a condition receive that condition, but no one receives a different condition from the one they were assigned (Horiuchi, Imai, and Taniguchi 2007).[1]

[Insert Table A1 here]

Most importantly, the presence of noncompliance means that comparing the mean outcome

[1] Two-sided noncompliance may be possible in a survey experiment, athough we expect that it is rare. The presence of two-sided noncompliance would not prevent a researcher from implementing our proposed methods for measuring and addressing the problem. We recognize that one could view our conception of compliance in this note as "narrow."

among the experimental groups according to the randomized assignment, Z_i, no longer represents an estimate of the average treatment effect (ATE), but rather an estimate of the "intent-to-treat" effect (ITT, see Imbens and Rubin 2015). The ITT provides information about the *effectiveness* of a treatment. For example, it can convey whether making a drug therapy *available* improves health outcomes, regardless of whether individuals take the medicine (Imbens and Rubin 2015). It is typically a conservative estimate that is the same sign as the ATE, but with reduced magnitude because of the assumption that the effect among never-taker noncompliers is zero (Imbens and Rubin 2015).

1.1 What is the Problem with the ITT?

If researchers ignore the possibility of noncompliance—as our meta analysis demonstrates is common (see below for details)—they may unknowingly report estimates of the ITT instead of the ATE. But what makes the ITT problematic? First, we contend that typically there is no compelling reason for researchers conducting survey experiments to be primarily interested in treatment effectiveness; there is not much to gain from estimating the effect of placing a vignette of text in front of survey respondents. Rather, researchers typically want to know the effect of respondents' processing of that information on their attitudes or some other outcome of interest. Thus, the quantity of greatest substantive interest is usually the ATE. Moreover, if a researcher decides that the ITT is, in fact, the quantity of interest in a survey experiment, he or she should explain the reasoning behind this decision.

Additionally, researchers should be concerned with conceptual precision in their empirical analyses—they should be able to clearly define and justify the estimand they report. Estimating the ITT but interpreting it as the ATE (even unintentionally) is inconsistent with this goal. If the ATE is the quantity of interest, researchers should either provide positive evidence that all respondents complied with the experimental protocol or follow steps (such as those we describe below) to arrive at that quantity.

Finally, even if the ITT and ATE signs are the same, we contend that researchers should care about accurately estimating the magnitude of a treatment effect. Magnitude communicates valu-

able information about respondents' reactions to experimental stimuli. It may be useful for comparisons with other survey experiments, for power analysis, or in discussing the substantive implications of results. Reducing interpretation to a positive/negative dichotomy unnecessarily limits the substantive richness of the research. In an era when the social sciences have been criticized for lacking relevance to public discourse, researchers should aim to provide the most complete picture of their results possible.

1.2 Satisficing in Surveys and Noncompliance

In recent years, political scientists have increasingly administered survey experiments online to representative or convenience samples. Participants in these samples are often recruited online and compensated with cash or a cash equivalent rewards. These rewards are frequently modest, and researchers typically have only limited means of forcing respondents to pay attention and put forth meaningful effort during administration of the survey instrument. Compensation is nearly always fixed ahead of time, so respondents have strong incentive to complete the survey quickly. Nonetheless, the text of experimental vignettes can be long and/or complex in nature, even if only a small part of that text actually changes between conditions. For example, Bearce and Tuxhorn's (2017) experiment contains vignettes of over 300 words on monetary policy.

Thus, it is reasonable to expect some survey respondents to skip some or all of a vignette and/or to engage in satisficing behavior—exerting only enough effort to meet some minimal acceptability criterion (Simon 1956). Of course, this issue is not new; social scientists have long acknowledged the potential for such dynamics in survey research (e.g., Krosnick 1991). However, this behavior may now be more problematic than ever, as online administration removes the (potentially motivating) presence of an interviewer/observer (Chang and Krosnick 2009; Kapelner and Chandler 2010). Moreover, the satisficers in a given sample typically differ from nonsatisficing respondents on several characteristics (Berinsky, Margolis, and Sances 2014).

1.3 Measuring Noncompliance in Survey Experiments

The first requirement to properly account for noncompliance in survey experiments is to measure it. Following standard practice, we assume that compliance is a binary measure.[2] We propose two straightforward methods: (1) recording latencies for experimental vignettes, and (2) repurposing manipulation checks. These are somewhat simple measures, and are certainly not the only possibilities (see Dafoe, Zhang, and Caughey 2018). They are useful because they are easy to implement and generally effective at separating compliers from noncompliers. However, in all circumstances we strongly recommend testing the proposed strategy in a pilot study before employing it on a full sample.

1.3.1 Latency

Our first approach involves measuring latency, or the amount of time that the experimental vignette appears in front of respondents. While this method does not guarantee that a respondent did, in fact, read the vignette carefully (or at all), it does indicate which respondents could not have possibly done so. The analyst can determine some minimum acceptable time based on a pilot study of the experiment, then code compliers as those respondents whose latency meets or exceeds that time. One approximate benchmark researchers may consider using in conjunction with a pilot study comes from research on reading speed. That work suggests that "normal" adult reading speed is approximately 250 words per minute, while "speedreaders" can reach as many as 700 words per minute (Rayner 1998, 392–393; see also Dyson and Haselgrove 2001).

Our meta analysis suggests that this approach is not widespread. We found that only about 12% of articles in our sample of 130 reported any method resembling what we describe above (loosely defined). While some studies provide measures of the amount of time respondents spend on the entire survey, it is uncommon to report the time spent specifically on the experimental vignette. Thus, the practice of conducting survey experiments could be improved by adopting this approach as standard protocol. In particular, we believe it is useful because it can be implemented without the subjects' knowledge and without adding more questions to a survey instrument.

[2] Allowing compliance to be a latent, continuous variable would take us beyond the scope of this article.

1.3.2 Manipulation Checks

Another possible measure of noncompliance involves repurposing manipulation checks (for examples, see Ahler 2014; Dafoe, Zhang, and Caughey 2018). Traditionally, manipulation checks have been used to verify internal validity; they tell the researcher whether he or she actually manipulated the intended concept in respondents' minds. However, they can also be used to measure noncompliance. As Berinsky, Margolis, and Sances (2014) explain, manipulation checks "[ask] a question that could only be answered by reading the treatment..." (743). We contend that this represents a useful means of determining whether a respondent read and thought about the vignette (and thus helps address satisficing behavior). The key choices the researcher must make with this approach include question wording, the number of questions to ask, and the number of correct answers required to qualify as a complier. As a general rule, we are fairly skeptical of the survey respondents that researchers typically employ on online platforms (e.g., MTurk). Thus, we recommend setting strict standards for compliance with manipulation checks. Respondents should be required to answer multiple questions correctly and these questions should be written such that the correct answers cannot be identified by skimming the vignette.[3]

We also recommend that analysts go beyond just summarizing responses to the manipulation check question(s). Our meta analysis—which produced 29 out of 130 articles that mentioned a manipulation check—revealed that reporting summary measures only is a common practice. For example, researchers report the proportion of respondents who answered the manipulation check correctly. Observing a "high" number (such as 95%) is favorable for assessing internal validity, but by our logic it also indicates the presence of some noncompliers. With the methods described below, we show that formally accounting for noncompliance requires using a manipulation check at the individual level.

As with latency, pilot studies are important for employing manipulation checks to effectively measure compliance. We advise researchers to consider the substantive context of their experimental vignettes and design manipulation check questions that are intended to directly address

[3]We agree with a reviewer that conceptual compliance checks can open up questions of interpretation. Thus, we urge scholars to be as transparent as possible about the choices they have made.

compliance. Researchers should ask themselves: "What type of question will produce responses that tell me if the subject internalized the content in the vignette?" Answering this question will likely take some trial and error with pilot samples.

1.3.3 Additional Implementation Issues

Both of the strategies described above require the researcher to make some seemingly arbitrary decisions. While pilot testing can help with this issue, it likely will not remove it completely. Accordingly, we also recommend that researchers describe and justify their chosen strategy for measuring compliance in preanalysis plans (see Monogan 2013). Publicly committing to a particular strategy *before* collecting data holds the researcher accountable and minimizes the risk of adjusting the definition of compliance after looking at results.

Additionally, it is possible that compliance could vary across subjects. For example, reading speed likely varies from person to person. We hesitate to recommend using different latency cutoff times for different subjects unless the researcher is very confident that he or she can accurately measure subjects' individual reading speeds. Absent that possibility, one approach might be to combine latency and manipulation checks. For example, a researcher could code a subject as compliant if and only if (1) he or she passed the latency cutoff and (2) he or she answered the manipulation check correctly. Alternatively, the researcher may decide that a respondent only has to satisfy one of those two conditions to be considered compliant. This latter approach would allow speed readers who do not hit the latency cutoff to prove that they really read the vignette (and thus should be considered compliers). It would also employ the latency measure as a "back-up" to the manipulation check if the analyst was concerned that respondents did not carefully read the manipulation check question(s).[4]

Our overall recommendation is simply that researchers should think seriously about measuring noncompliance well in advance of fielding their survey experiments. The best strategy for measuring compliance in a given experiment will depend on the experimental vignettes, the sample, and the researchers' goals. We offer the two strategies described above as feasible possibilities for

[4]In fact, the researcher could even place a latency measure on the manipulation check question to assess whether respondents were taking the manipulation check seriously.

a wide range of experiments, but we do not necessarily expect that they will always represent the optimal approaches.

1.4 Causal Estimation under Noncompliance

Using methods such as those described above will give researchers a variable, C_i, in their completed dataset that indicates which respondents complied with treatment. With this variable measured, several possible estimands are available. We summarize each one here. The notation, which we adapt from Imbens and Rubin (2015, Chapter 23), is as follows:

1.4.1 Variables

- $Z_i \in [0,1]$: Random assignment to treatment $(Z_i = 1)$ or control $(Z_i = 0)$.
- $D_i \in [0,1]$: Receipt of treatment $(D_i = 1)$ or control $(D_i = 0)$.
- $Y_i(Z_i, D_i[Z_i])$: Potential outcome based on realizations of Z_i and D_i.
- Y_i: Observed outcome.
- $C_i \in [0,1]$: Indicator for compliers.

1.4.2 Sample Summaries

- N: Total number of observations in the full sample of data.
- $N_1 = \sum_{i=1}^{N} Z_i$: Total number of observations assigned to treatment.
- $N_0 = \sum_{i=1}^{N} (1 - Z_i)$: Total number of observations assigned to control.
- $N_t = \sum_{i=1}^{N} D_i$: Total number of observations receiving treatment.
- $N_c = \sum_{i=1}^{N} (1 - D_i)$: Total number of observations receiving control (i.e., failing to receive treatment).
- $N_{1t} = \sum_{i=1}^{N} Z_i \cdot D_i$: Total number of observations assigned to treatment and receiving treatment.
- $N_{0c} = \sum_{i=1}^{N} (1 - Z_i) \cdot (1 - D_i)$: Total number of observations assigned to control and receiving control (i.e., failing to receive treatment).
- $\overline{Y_1} = \frac{1}{N_1} \sum_{i=1}^{N} (Z_i \cdot Y_i)$: Average outcome for observations assigned to treatment.

- $\overline{Y_0} = \frac{1}{N_0}\sum_{i=1}^{N}([1-Z_i]\cdot Y_i)$: Average outcome for observations assigned to control.
- $\overline{Y_t} = \frac{1}{N_t}\sum_{i=1}^{N}(D_i\cdot Y_i)$: Average outcome for observations receiving treatment.
- $\overline{Y_c} = \frac{1}{N_c}\sum_{i=1}^{N}([1-D_i]\cdot Y_i)$: Average outcome for observations receiving control (i.e., failing to receive treatment).
- $\overline{Y_{1t}} = \frac{1}{N_{1t}}\sum_{i=1}^{N}(Z_i\cdot D_i\cdot Y_i)$: Average outcome for observations assigned to treatment and receiving treatment.
- $\overline{Y_{0c}} = \frac{1}{N_{0c}}\sum_{i=1}^{N}([1-Z_i]\cdot[1-D_i]\cdot Y_i)$: Average outcome for observations assigned to control and receiving control (i.e., failing to receive treatment).

1.4.3 Estimands

ITT

The ITT is defined by the random treatment assignment:

$$\text{ITT} = \frac{1}{N}\sum_{i=1}^{N}[Y_i(1, D_i\{1\}) - Y_i(0, D_i\{0\})]. \tag{A1}$$

It can be estimated as the difference in average outcomes for observations assigned to treatment and assigned to control:

$$\widehat{\text{ITT}} = \overline{Y_1} - \overline{Y_0}. \tag{A2}$$

This estimand ignores the issue of compliance. It would likely only be deliberately estimated if the researcher's central interest was in effectiveness of the treatment, or the effect of making the treatment available.

As-Treated

Alternatively, the researcher could estimate the as-treated effect. This estimand ignores the random treatment assignment and computes the difference between the average outcomes by treatment received:

$$\widehat{\text{AT}} = \overline{Y_t} - \overline{Y_c} \tag{A3}$$

The approach is generally problematic because it breaks the randomization due to the fact that respondents select into compliance. It would only identify the ATE if there were no confounders of treatment receipt, which is unlikely in most all applied settings.

Per-Protocol

The per-protocol approach involves estimating the causal effect after dropping *observed* noncompliers. In the context of field experiments, this approach usually means dropping only the noncompliers in the treatment group because the researcher cannot observe compliance in the control group. Accordingly, the estimand compares observations that received treatment (i.e., compliers who were assigned to treatment) to observations assigned to the control. The latter group is comprised of both compliers and noncompliers.

$$\widehat{\text{PP}} = \overline{Y_t} - \overline{Y_0} \tag{A4}$$

This estimand is usually not useful because the subsample assigned to treatment is unlikely to be a good counterfactual for the subsample that receives treatment. Indeed, the practice amounts to conditioning on a posttreatment variable. It is only an unbiased estimate of the ATE if all observations are compliers.

CACE

A more common (and useful) estimate is the complier average causal effect (CACE). This is a "local" ATE (LATE)—an average causal effect for the subset of respondents induced to comply with treatment via the random assignment. It can be consistently estimated via instrumental variables analysis (IV) by using Z_i as an instrument for D_i (Angrist, Imbens, and Rubin 1996). The CACE is defined as:

$$\text{CACE} = \frac{\sum_{i=1}^{N} C_i[Y_i(1, D_i\{1\}) - Y_i(0, D_i\{0\})]}{\sum_{i=1}^{N} C_i}. \tag{A5}$$

This estimand is the effect of the treatment for those who take the treatment only when assigned it. The CACE requires the exclusion restriction assumption, which states that the effect of Z_i on Y_i is zero when $C_i = 0$ (i.e., no effect for noncompliers). If we assume no always-takers (and no

defiers), this amounts to the assumption that the effect of Z_i on Y_i is zero for never-takers. From a substantive perspective, it means that the CACE will always be larger in magnitude than the ITT (Horiuchi, Imai, and Taniguchi 2007, 678).

The CACE is a local estimate, the generalizability of which is governed by the degree to which the vignette induces respondents to receive the treatment. Nonetheless, it is generally a useful estimate in the context of survey experiments. Researchers who conduct survey experiments are not likely to be interested in treatment effectiveness like a researcher considering a real-world policy intervention (i.e., the ITT). Instead, researchers conducting survey experiments are concerned with the effect of their experimental vignettes on those respondents who actually processed them (i.e., received treatment).[5]

Average Effect of Receipt for Compliers (AERC)

Another estimand may be available in the context of a survey experiment: the average effect of receipt for compliers, which we denote the AERC. In discussing the per-protocol estimand, Imbens and Rubin (2015, 537) note that "if we could, in fact, discard all noncompliers, we would be left with only compliers, and then comparing their average outcomes by treatment status would estimate the [AERC]." Depending on the compliance measurement strategy in a survey experiment, the analyst may be able to separate compliers from noncompliers in the control group if the control group views its own experimental vignette and/or completes a manipulation check. This feature contrasts with some field experiments in which only the treatment group is actually offered

[5]If the researcher is not satisfied with the CACE and instead declares the quantity of interest to be the ATE, one option to consider would be the inverse compliance score weighting (ICSW) approach of Aronow and Carnegie (2013). ICSW allows researchers to estimate the ATE under noncompliance. It involves estimation of a compliance score, which is the "conditional probability of being a complier given covariates" (693). Implementation depends on several assumptions and the nature of the noncompliance (one- or two-sided). In general, the researcher first models C_i as a function of pretreatment covariates. The estimated compliance score is the fitted probability of compliance for every respondent, given that model. The final step is to weight the CACE estimation (e.g., the IV regression) by the inverse of the compliance score. The intuition behind this approach is that the reweighting balances the distribution of covariates among the compliers to match the full population. Aronow and Carnegie (2013) show that under some assumptions, this method produces a consistent estimate of the ATE.

something as treatment.[6] The AERC can be estimated as:

$$\widehat{\text{AERC}} = \overline{Y_{1t}} - \overline{Y_{0c}}. \tag{A6}$$

A unique property of this estimand is that it is equivalent, in expectation, to the CACE under the "second exclusion restriction" assumption. This assumption has no direct empirical consequence and is typically implied. It states that the effect of *receipt* of treatment is equal to *assignment* to treatment for compliers (see Imbens and Rubin 2015, 529). In the sample, estimates of the CACE and AERC will be different because they use different subsets of the data. Given the common use of instrumental variables to estimate treatment effects under noncompliance, we recommend estimating the CACE.

2 Journal Article Meta Analysis

We assessed how political scientists address noncompliance in survey experiments through a meta analysis of published research from the last decade. Specifically, we examined all articles reporting results from a survey experiment published in five journals from 2006–2016: *American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, *Public Opinion Quarterly*, and *Political Behavior*. We first searched for articles using Google Scholar. Then, we coded the articles based on answers to a number of questions relating to their handling of noncompliance. We describe the details of these procedures below. Our central objectives were to assess (1) how often researchers think about noncompliance in analyzing survey experiments, and (2) what, if anything, they do to address the issue.

2.1 Searching for Articles

We searched https://scholar.google.com/ for "survey experiment" in each journal listed above, one journal at a time. We set the date range to 2006–2016. We downloaded all of the articles returned by these searches, checked each one to make sure it included a survey experiment, then

[6] Note that it is conceptually distinct from two-sided noncompliance because the noncompliers in the control group do not receive the treatment.

saved it for coding in the next step. We included any paper that reported a survey experiment in the main text. This produced 130 total articles from four subfields: American politics (84), comparative politics (17), international relations (8), and methodology (17). The remaining four articles covered multiple subfields. Additionally, the articles spanned all five journals: 11 from *American Political Science Review*, 32 from *American Journal of Political Science*, 25 from *Journal of Politics*, 34 from *Public Opinion Quarterly*, and 28 from *Political Behavior*.

2.2 Coding the Articles

We read each article and coded answers to the following binary response questions. The percentage of articles affirming each question is given in parentheses.

- *Does the article mention the possibility of noncompliance with the experiment? (13%).* Answers to this question give a sense of how common it is for political scientists to think about noncompliance as a potential issue in the analysis of survey experiments. We found that it is not common—only 17 of the 130 articles mention noncompliance.
- *Does the article mention a change in the estimand due to noncompliance? (6%).* We coded a positive response here if the article acknowledged that, due to noncompliance, the reported effect was an ITT.
- *Does the article report multiple estimands? (4%).* We recorded a positive response to this question if the article reported more than one estimand (e.g., ITT and CACE) due to the issue of noncompliance.
- *Does the article report an estimate of the ATE after acknowledging noncompliance? (0%).* We found no article in our sample that estimated the ATE under noncompliance.
- *Does the article report latency time? (12%).* We coded any mention of time as a positive response, including time spent on the entire survey or experimental vignette.
- *Does the article report a manipulation check? (22%).* We coded any mention of a manipulation check as a positive response here, regardless of whether the manipulation check was actually used to assess compliance.

2.3 Meta Analysis Conclusions

Overall, we conclude that considering noncompliance as a potential problem is not a common practice in the design and conduct of survey experiments. Across a decade's worth of survey experimental research in some of political science's top journals, under one-sixth of the articles examined even mention noncompliance. Furthermore, we find that those articles that do acknowledge the issue rarely do much about it. Only eight of the 130 articles in our sample adjust their causal estimand, and none of the articles followed the protocol we suggest above.

3 Replications

Here we conduct replications of published survey experiments to assess the consequences of accounting for noncompliance. Our first empirical example uses data from a survey experiment conducted by Harden (2016). Then we summarize the results of several other replications.

3.1 Replication of Harden (2016)

Harden (2016) reports results from a survey experiment administered online via Qualtrics to 1,975 American adults who were compensated with small amounts of cash or gift cards. Thus, the concerns about satisficing discussed above are germane to this sample. Importantly, Harden (2016) measured the time, in seconds, that the experimental vignette was on the screen for each respondent.

3.1.1 Experimental Details

Respondents were shown an exchange between a hypothetical constituent and legislator about the distribution of government funds for education (see below for details). The constituent asks the legislator about securing money for public education in the district. The legislator's response is randomly assigned to take one of two different forms. In the "pork barrel" condition, the legislator explains that he was able to secure funding in an unrelated transportation bill. In the "fair share" condition, the legislator's response emphasizes that the district will be getting funding from an education bill based on need. The outcome variable is a feeling thermometer rating of the legislator, scaled from 0–100. Here we consider the hypothesis that citizens prefer the fair share condition

(signaling distributing based on need) over distribution via pork barrel politics.

The pork barrel and fair share vignettes contain 156 and 157 words, respectively, which produced a median latency of 31 seconds and an interquartile range of 34 seconds. Harden (2016) does not mention the possibility of noncompliance and reports the ITT, implicitly assuming it is equivalent to the ATE. We consider four definitions of complier status based on latency cutoffs of 15, 25, 35, and 45 seconds (74%, 62%, 44%, and 28% compliers, respectively). We report this range of values to assess the sensitivity of results to the choice, but an average reading speed of 250 words per minute (Rayner 1998) would suggest that 35 or 45 seconds are the most reasonable. Assuming respondents only spent the recorded time reading, the four values we select correspond, respectively, to reading times of 628, 377, 270, and 209 words per minute.

3.1.2 Results

Figure A1 presents the estimated treatment effect of the pork barrel condition using our four latency cutoffs for four different estimands: the ITT (original result), as-treated, per-protocol, and CACE. Circles represent point estimates and lines indicate 95% confidence intervals.

[Insert Figure A1 here]

The estimates show considerable variation. The ITT is negative and statistically significant, but quite small in magnitude (3 points on the 0–100 scale). This contrasts with the as-treated effects, which are all *positive* and statistically significant estimates of 3–4 points regardless of the time cutoff. The per-protocol effects are substantively small and not significant. The CACE estimates are negative and statistically significant, and become increasingly larger in magnitude (4–10 points) as the latency time cutoff increases.

Overall, the ITT and CACE estimates both tell similar stories, but the CACEs indicate more intensity in the finding. The CACE estimates range from 35% larger than the ITT (15 second cutoff) to 260% larger (45 seconds). Of course, the CACE confidence intervals are larger too, and the CACE and ITT estimates are never statistically significantly different from each other. Nonetheless, given that there is no compelling reason to assess treatment effectiveness in this example (i.e., the ITT), the CACE would be a more reasonable estimand on which to focus. Doing

so would indicate strong support for the hypothesis that citizens prefer fair allocation of funds over pork barrel politics.

3.1.3 Vignette and Question Wording

The following text is reproduced from the appendix to Harden (2016).

Introductory Text

The next questions are based on excerpts from e-mail conversations between a constituent and a state legislator. Imagine that you are the constituent asking the question. Then based only on the information that is given, evaluate your feelings toward the legislator as if the legislator were your representative. Select your evaluation on the "feeling thermometer" provided after the description. This measure ranges from 0 to 100, with higher scores indicating a more favorable rating. If you feel neutral toward the legislator, select the score 50.

Constituent Question

The school buildings in our district have gotten terrible the last several years. Can't you get any funding for some renovations?

Pork Barrel Treatment

Dear Constituent, Things will start to improve next year. Do you remember the big transportation bill passed this past summer? I was able to add an amendment just before the final vote that specifically set money aside for renovating schools in our district's borders. It's a special allotment of bonus money, just for our district! I was able to convince the legislature that we have a real need, which means we will be getting almost twice as much as the average district for renovations. So look for better looking schools in the future! Sincerely, Rep. Vincent A. Taylor

Fair Share Treatment

Dear Constituent,

Things will start to improve next year. Do you remember the education bill passed this past summer? That bill set aside an allotment of several million dollars solely for school renovation. Dividing that money quickly became a partisan struggle. However, with the help of several others, I was able to convince the legislature that the money should be divided based on need. The area covering most of our district will be getting almost twice as much as the average district for building repairs. So look for better looking schools in the future!

Sincerely,
Rep. Vincent A. Taylor

Respondents answered the following question after viewing a message.

> Evaluate your feelings toward the legislator as if the legislator were your representative. You may choose any number from 0 to 100. 0 = Most unfavorable; 50 = Neutral; 100 = Most favorable. [0–100 Slide Bar].

3.1.4 Latency Time Distribution

Figure A2 graphs the distribution of respondent times on the vignette screen. Recall that the median time was 31 seconds with an interquartile range of 34 (25^{th} percentile: 14 seconds, 75^{th} percentile: 48 seconds).

[Insert Figure A2 here]

3.2 Additional Replications

We conducted additional replications of published work to assess the broader implications of accounting for noncompliance in survey experiments. We went back to our meta analysis and identified all of the articles that report using latency or a manipulation check with experimental vignettes. Then we searched for the replication data of all of these studies and attempted to replicate as many as possible. Among the 29 papers that mention a compliance measure, we found complete replication data publicly available for seven. We successfully replicated five of these articles.[7] In

[7]Those articles are Horowitz and Levendusky (2011), Ahler (2014), Grose, Malhotra, and Van Houweling (2015), Corbacho et al. (2016), and Stephens-Dougan (2016).

this set of replications we recorded the ITT and CACE for every experimental treatment effect reported in the articles. Most of them presented several such estimates due to multiple experimental conditions, multiple outcomes, or both. This exercise yielded a total of 51 estimates across the five articles we could replicate, as well as the replication of Harden (2016) described previously. [8]

After recording the ITT and CACE estimates and standard errors, we compared the change in magnitudes moving from the former estimand to the latter. Figure A3 reports the results. The median change in the treatment effect (panel a) is a 28% increase in magnitude, although in one case the increase is over 200%. The interquartile range of those changes is (5%, 71%). The change in standard error size is even larger (median: 59%, IQR: [47%, 84%]). In short, estimating the CACE increases the magnitude of the estimated effect, which is expected because by assumption the effect is zero for noncompliers. Additionally, employing it increases researchers' uncertainty around the estimated effect.

[Insert Figure A3 here]

Whether or not a change to the CACE is substantively consequential will vary across studies. However, we contend that a median increase of 28% in magnitude in the small sample of studies we replicate here is notable. The CACE should always point in the same direction as the ITT, but in our view the sign of a treatment effect is usually not enough to evaluate support for a hypothesis. It is also necessary to consider the intensity of that effect. In doing so, researchers may find that the effect using our recommended estimand (CACE) may tell a different story than the effectiveness of treatment (ITT).

[8] We present this analysis as a proof of concept. The manipulation checks in these studies were not necessarily designed to measure compliance, so we should exercise some caution in interpreting the results.

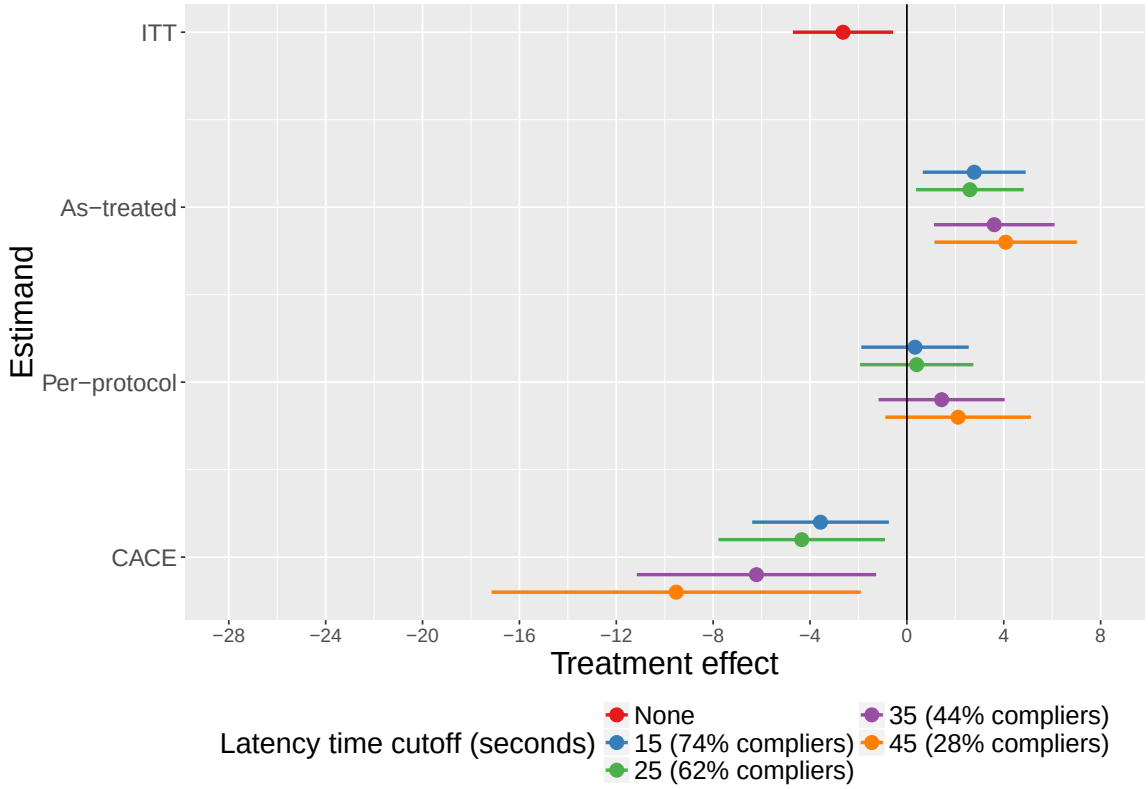
Table A1: Classification of Compliance Status in a Survey-Experimental Design

	$D_i(0) = 1$	$D_i(0) = 0$
$D_i(1) = 1$	Noncomplier (always-taker)*	Complier
$D_i(1) = 0$	Noncomplier (defier)*	Noncomplier (never-taker)

Note: Cell entries report the complier type by the potential values of $D_i(Z_i)$, $Z_i \in [0, 1]$.

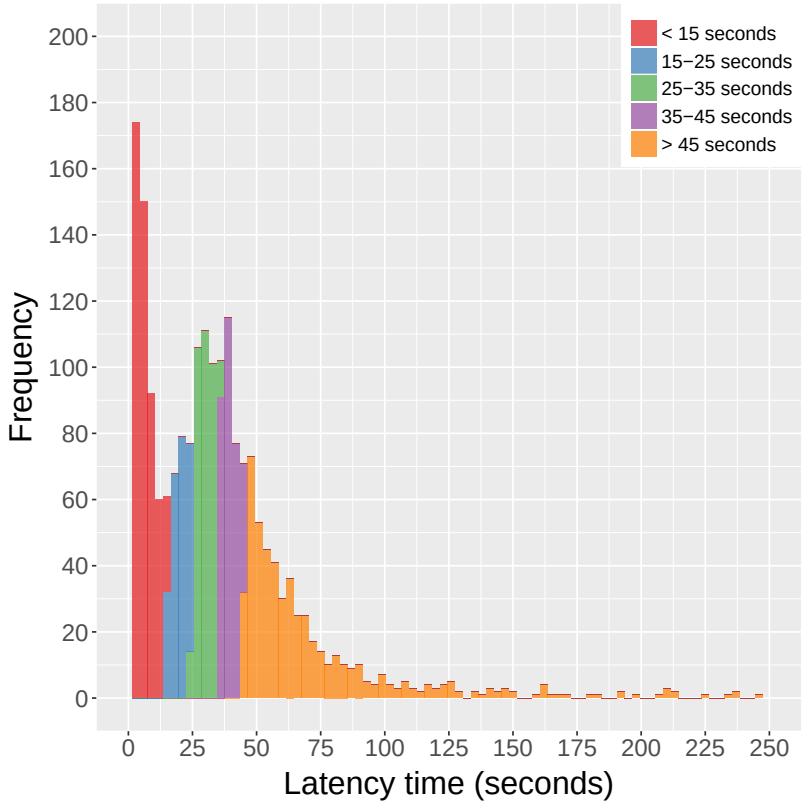
*Noncomplier types that do not exist in the context of a survey experiment.

Figure A1: Treatment Effects with Different Estimands in the Replication of Harden (2016)



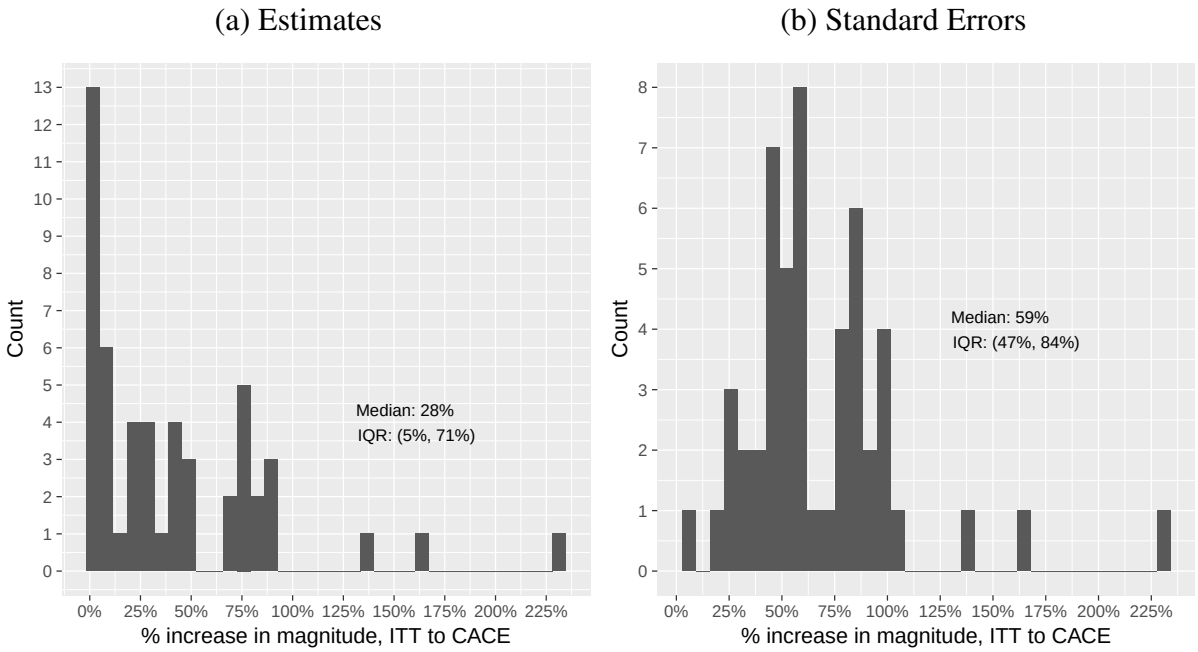
Note: The graph presents point estimates and 95% confidence intervals for several different estimands representing the effect of the pork barrel treatment in the survey experiment described in Harden (2016). Colors denote the cutoff criteria for the time denoting compliers.

Figure A2: Distribution of Respondent Latency Time in the Harden (2016) Survey Experiment



Note: The graph presents the distribution of respondent latency time in the Harden (2016) survey experiment. Colors denote several cutoff criteria for the time of compliers.

Figure A3: Distribution of ITT to CACE Change in Magnitude in Replicated Survey Experiments



Note: The graphs present the distribution of percentage change in magnitude for treatment effect estimates (panel a) and standard errors (panel b) moving from the ITT to the CACE in the replication studies.

References

Ahler, Douglas J. 2014. "Self-Fulfilling Misperceptions of Public Polarization." *Journal of Politics* 76(3): 607–620.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91(434): 444–455.

Aronow, Peter M., and Allison Carnegie. 2013. "Beyond LATE: Estimation of the Average Treatment Effect with an Instrumental Variable." *Political Analysis* 21(4): 492–506.

Bearce, David H., and Kim-Lee Tuxhorn. 2017. "When Are Monetary Policy Preferences Egocentric? Evidence from American Surveys and an Experiment." *American Journal of Political Science* 61(1): 178–193.

Berinsky, Adam J., Michele F. Margolis, and Michael W. Sances. 2014. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys." *American Journal of Political Science* 58(3): 739–753.

Chang, Linchiat, and Jon A Krosnick. 2009. "National Surveys via RDD Telephone Interviewing Versus the Internet Comparing Sample Representativeness and Response Quality." *Public Opinion Quarterly* 73(4): 641–678.

Corbacho, Ana, Daniel W. Gingerich, Virginia Oliveros, and Mauricio Ruiz-Vega. 2016. "Corruption as a Self-Fulfilling Prophecy: Evidence from a Survey Experiment in Costa Rica." *American Journal of Political Science* 60(4): 1077–1092.

Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2018. "Information Equivalence in Survey Experiments." Forthcoming, *Political Analysis*. http://www.allandafoe.com/ie.

Dyson, Mary C., and Mark Haselgrove. 2001. "The Influence of Reading Speed and Line Length on the Effectiveness of Reading from Screen." *International Journal of Human-Computer Studies* 54(4): 585–612.

Grose, Christian R., Neil Malhotra, and Robert P. Van Houweling. 2015. "Explaining Explanations: How Legislators Explain their Policy Positions and How Citizens React." *American Journal of Political Science* 59(3): 724–743.

Harden, Jeffrey J. 2016. *Multidimensional Democracy: A Supply and Demand Theory of Representation in American Legislatures*. New York: Cambridge University Press.

Horiuchi, Yusaku, Kosuke Imai, and Naoko Taniguchi. 2007. "Designing and Analyzing Randomized Experiments: Application to a Japanese Election Survey Experiment." *American Journal of Political Science* 51(3): 669–687.

Horowitz, Michael C., and Matthew S. Levendusky. 2011. "Drafting Support for War: Conscription and Mass Support for Warfare." *Journal of Politics* 73(2): 524–534.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge University Press.

Kapelner, Adam, and Dana Chandler. 2010. Preventing Satisficing in Online Surveys: A 'Kapcha' to Ensure Higher Quality Data. In *Proceedings of CrowdConf 2010*.

Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5(3): 213–236.

Monogan, James E. 2013. "A Case for Registering Studies of Political Outcomes: An Application

in the 2010 House Elections." *Political Analysis* 21(1): 21–37.

Rayner, Keith. 1998. "Eye Movements in Reading and Information Processing: 20 Years of Research." *Psychological Bulletin* 124(3): 372–422.

Simon, Herbert A. 1956. "Rational Choice and the Structure of the Environment." *Psychological Review* 63(2): 129–138.

Stephens-Dougan, LaFleur. 2016. "Priming Racial Resentment without Stereotypic Cues." *Journal of Politics* 78(3): 687–704.