Confounding in Survey Experiments: Diagnostics and Solutions

Allan Dafoe\textsuperscript{1}, Baobao Zhang\textsuperscript{1}, and Devin Caughey\textsuperscript{2}

\textsuperscript{1}Department of Political Science, Yale University
\textsuperscript{2}Department of Political Science, Massachusetts Institute of Technology

This draft: December 10, 2015

For printing, consider skipping the lengthy appendix beginning on page 32.

Abstract

Despite random assignment, scenario-based survey experiments suffer from threats to internal validity that resemble confounding bias in observational studies. Manipulating specific features of a scenario generally changes subjects’ beliefs about background features of the scenario, possibly confounding the effects of the beliefs of interest. Describing a hypothetical country as “a democracy”, for example, makes respondents more likely to think it is wealthy, European, predominantly Christian and white, and interdependent and allied with the United States. We show how to theorize about confounding in survey experiments, and how to diagnose it using placebo tests. We also evaluate three potential solutions. Encouraging subjects not to think about specific countries does not reduce confounding. Controlling for potential confounders in the scenario reduces confounding only on the controlled and closely related attributes. The best solution is embedding a natural experiment in the scenario, a new technique that mitigates all sources of confounding.

Endnotes are labeled using capital letters, and hyperlink to the Online Appendix. Our main studies reported in this paper have been pre-registered and pre-analysis plans have been posted. At the time of publication complete replication files will be posted. The most recent version of this paper, as well as our pre-analysis plans and other related materials, can be found at allandafoe.com/confounding. For helpful comments, we would like to thank Peter Aronow, Cameron Ballard-Rosa, Adam Berinsky, David Broockman, Alex Debs, Chris Fariss, Alan Gerber, Donald Green, Sophia Hatz, Dan Hopkins, Susan Hyde, Josh Kalla, Gary King, Audrey Latura, Jason Lyall, Elizabeth Menninga, Nuno Monteiro, Brendan Nyhan, Jonathan Renshon, Bruce Russett, Cyrus Samii, Maya Sen, Robert Trager, Mike Tomz, Jessica Weeks, Teppei Yamamoto, Sean Zeigler, Thomas Zeitzoff, and participants of the University of North Carolina Research Series, the Yale Institute for Social and Policy Studies Experiments Workshop, the Yale International Relations Workshop, the University of Konstanz Communication, Networks and Contention Workshop, the Polmeth 2014 and 2015 Summer Methods Meetings, and the Survey Experiments in Peace Science Workshop. For support, we acknowledge the MacMillan Institute at Yale University, and the National Science Foundation Graduate Research Fellowship Program.
Confounding in Survey Experiments: Diagnostics and Solutions

December 10, 2015

Abstract

Despite random assignment, scenario-based survey experiments suffer from threats to internal validity that resemble confounding bias in observational studies. Manipulating specific features of a scenario generally changes subjects’ beliefs about background features of the scenario, possibly confounding the effects of the beliefs of interest. Describing a hypothetical country as “a democracy”, for example, makes respondents more likely to think it is wealthy, European, predominantly Christian and white, and interdependent and allied with the United States. We show how to theorize about confounding in survey experiments, and how to diagnose it using placebo tests. We also evaluate three potential solutions. Encouraging subjects not to think about specific countries does not reduce confounding. Controlling for potential confounders in the scenario reduces confounding only on the controlled and closely related attributes. The best solution is embedding a natural experiment in the scenario, a new technique that mitigates all sources of confounding.
1 Introduction

Many important social-scientific questions concern the effects of some real-world attribute or event on people’s opinions, preferences, and behaviors:

- Does representation by a co-ethnic increase citizens’ political participation (Broockman, 2013)?
- How do candidates’ party brands and issue positions affect the decisions of voters (Butler and Powell, 2014; Tomz and Van Houweling, 2008)?
- How does the skill level of potential immigrants influence citizens’ support for immigration (Hainmueller and Hiscox, 2010; Malhotra, Margalit and Mo, 2013; Tai and Truex, 2015)?
- Do leaders’ rhetoric and behavior in interstate conflicts affect their standing in the public (Levendusky and Horowitz, 2012; Tomz, 2007; Trager and Vavreck, 2011)?
- How do voters react to foreign intervention in their country’s elections (Marinov, 2013)?
- Is public support for using military force abroad influenced by whether the target country is democratic (Johns and Davies, 2012; Mintz and Geva, 1993; Tomz and Weeks, 2013)?

Studying such questions in the real world is often difficult because the effects of the cause of interest (representatives’ ethnicity, immigrants’ skills, etc.) are typically confounded by other causes of the outcome. Moreover, strategies for reducing confounding, such as controlling for covariates or exploiting randomization, are often infeasible. For these reasons, the works cited above all adopt an alternative research strategy: fielding opinion surveys that present respondents with hypothetical scenarios whose content is randomly manipulated. Through randomization, scenario-based survey experiments ensure that the manipulation’s effects are not confounded by other causes of subjects’ survey responses. The power of this approach has led to a recent explosion in political scientists’ use of survey experiments (Figure 1).

The promise of survey experiments is illustrated by Tomz and Weeks (2013), whose exemplary study of the democratic peace we use as a running example through this paper. Tomz and Weeks examine the longstanding claim that democratic publics resist going to war with other democracies (e.g., Doyle, 1986, 1161). Many scholars (e.g., Maoz and Russett, 1993) have examined this proposition using real-world dispute data. As is well understood, however, observational studies are vulnerable to confounding, and even after decades of work we still cannot confidently conclude that the democratic peace is entirely due to democracy itself, as opposed to correlated characteristics such as religion, American hegemony, or

1 Confounding refers to a situation in which exposure to the cause of interest is not independent of units’ potential outcomes.

2 We define a survey experiment as the deliberate (typically random) manipulation of the form or content of a survey. Our focus is on scenario-based survey experiments, which are survey experiments that involve the depiction of a scenario and questions that pertain in part to the scenario. The general problem we discuss applies to all experiments, including framing and list experiments, but the specific problem and solutions that are our focus are distinct to scenario-based survey experiments.
capitalism. These concerns with confounding in observational studies motivated Tomz and Weeks’s experimental approach. Tomz and Weeks presented American and British survey respondents with hypothetical scenarios involving a country developing nuclear weapons in a threatening manner; the country was described as either democratic or non-democratic. As predicted, respondents are less supportive of using military force against countries described as democratic. Tomz and Weeks conclude that this “experimental approach allows us to conclude with confidence that the effect of democracy is genuinely causal” (2013, 862).

As we demonstrate, however, scenario-based survey experiments such as Tomz and Weeks (2013) are susceptible to bias. Further, these are likely to be similar to the confounding biases endemic to observational studies. To see why, it is helpful to remember that the purpose of the studies described above is to infer from subjects’ responses something about causal relationships in the real world (Gaines, Kuklinski and Quirk, 2007, 2). Ultimately, the researcher is interested in outcome differences between two hypothetical versions of the real world, one in which the causal factor is present and one in which it is absent, with other factors held constant. In order to make inferences about these real-world effects, survey experimenters seek to conjure up analogous hypothetical worlds in the minds of survey respondents. That is, by manipulating some aspect of the survey, the experimenter seeks to set different respondents’ beliefs about some aspect of the scenario to particular values, without affecting other beliefs. If the experiment succeeds in doing so, then the observed comparison provides an unbiased estimate of the causal effect of the beliefs in question, which informs our understanding of its real-world analogue. In the case of the democratic peace, for example, we hope to learn about the effects of a country being democratic from the experimental effects of respondents believing that a country in a scenario is democratic, induced by experimentally manipulating the description of the country as a democracy.

But randomizing the description of a feature, even if it succeeds in manipulating respondents’ beliefs about the feature, is not sufficient to identify the effect of these beliefs. This is because the manipulation will generally affect subjects’ beliefs about other features of the scenario as well. In fact, if respondents update their beliefs based on their knowledge of the
real world, then the associations between beliefs about the main causal factor and beliefs about other factors will mirror those in the analogous observational study. A respondent told that a country in a scenario is a “democracy,” for example, will likely infer that it is more likely to be wealthy, Christian, European, allied with the U.S., and so on. If people’s beliefs about any of these other features affect the outcome of interest, then the effect of believing a country is a democracy in the survey experiment will be confounded in the same way that observational studies are. We label this the problem of confounding in survey experiments.

To assess the extent to which others are aware of this problem we reviewed all survey experiments published in *American Journal of Political Science*, the *American Political Science Review*, and *International Organization* since 2002. We found that most studies (70%) do not acknowledge any limits to the internal validity of their causal inference; only five studies (15% of scenario-based survey experiments) acknowledged any possibility of confounding (that respondents’ beliefs about other features of the scenario may be affected by the experimental manipulation). Tomz and Weeks (2013, 853) are in fact among the most explicit, noting that labeling a country democratic may cause “information leakage” (Sher and McKenzie, 2006) that leads respondents to infer “that the country was also an ally, a major trading partner, or a powerful adversary.” Hainmueller and Hopkins (2014, 533) express worry about the “possibility of confounding if the experimentally manipulated attributes are correlated with other influential attributes” and Hainmueller, Hopkins and Yamamoto (2014, 5) express the concern that previous studies’ “findings for certain attributes are masking the effects of others.” Thus, it is clear from our literature review that the possibility of confounding in scenario-based survey experiments is not widely recognized. Further, there has been close to no systematic examination of its existence or character, nor about the efficacy of potential solutions. Our paper does this.

We contribute to scholarly understanding of this problem in several respects: we offer a theory that clarifies how it works, we offer diagnostics for the presence of confounding, we evaluate two existing responses to it and find them limited, and we introduce a new solution that overcomes confounding. Our first contribution is to formalize confounding in survey experiments in the language of instrumental variables (IV), where the survey manipulation is the instrument and respondents’ beliefs are the treatment. The IV framework makes explicit the assumptions that are sufficient to make inferences about the treatment effect from the estimated effect of the manipulation. Without these assumptions or others like them, the effect of the manipulation bears no necessary relationship to the effect of treatment, and may even have the opposite sign. Since IV analysis is a special case of mediation analysis with a single mediator, these results can also be understood from the perspective of mediation analysis.

Second, we provide a framework for theorizing about confounding *ex ante* and diagnosing it *ex post*. We argue that respondents (a) have prior beliefs about attributes’ associations that roughly reflect their real-world associations and (b) update their beliefs in a roughly Bayesian manner. Telling American respondents that a hypothetical immigrant is from India rather than Mexico, for example, should increase respondents’ estimate of the immigrant’s education level because real-world immigrants from India are better-educated than...
immigrants from Mexico. This realistic Bayesian model yields precise predictions about respondent beliefs, and our results are consistent with these predictions. Other models of respondent belief updating can also be employed, such as models of inference relying on stereotypes or heuristics.

Third, we show how to use the realistic Bayesian model to identify likely confounders and to diagnose confounding using placebo tests: tests of effects that must be zero under the identification assumptions. A key assumption of IV estimation is that the instrument affects the outcome only through the treatment; if this exclusion restriction is violated, the estimated effect of treatment will generally be biased. We therefore recommend that researchers assess whether the experimental manipulation influenced the subject’s beliefs about other background features of the scenario that are likely to be relevant to the subject’s response. In studies of the democratic peace, for example, describing a country as democratic or not democratic should not influence whether respondents think the country is in the Middle East. Imbalance between survey conditions in subjects’ beliefs provides evidence that the treatment effect is confounded.

Fourth, we identify and evaluate three potential solutions to the problem of confounding in survey experiments: abstract encouragement, covariate control, and embedded natural experiments, the last of which is our own innovation. An abstract encouragement design explicitly asks respondents to consider the scenario in the abstract rather than thinking of real-world examples. A covariate control design attempts to reduce or eliminate confounding by specifying further details about the scenario, thus hopefully fixing respondent beliefs in a way that reduces confounding. The scenario designed by Tomz and Weeks (2013), for example, specifies the target country’s military capacities and level of trade with the United States. A particularly elaborate form of covariate control design is conjoint analysis, which presents subjects with hypothetical profiles with different combinations of many attributes (Hainmueller, Hopkins and Yamamoto, 2014).

We also introduce a novel design: the embedded natural experiment (ENE). Whereas covariate control tries to eliminate confounding by controlling for specific characteristics in the scenario, ENEs induce respondents to perceive that treatment assignment is independent of all pretreatment causes of the outcome. That is, the ENE design embeds in the survey scenario a hypothetical natural experiment in which respondents perceive the attribute of interest to be as good as randomly assigned. To study the mass basis of the democratic peace, for example, we construct an ENE in which the leader of a fragile authoritarian regime is subject to an assassination attempt. If the assassination succeeds, the regime collapses and transitions to democracy; if it fails, the country remains authoritarian. As long as respondents perceive the assassination outcome and ensuing regime transition to be as good as random and they respond like realistic Bayesians, the experimental manipulation should not affect any of their beliefs about the background characteristics of the country—only whether they perceive it to be democratic.

Across several example applications, we find that randomization of the vignette alone is not sufficient to prevent confounding in survey experiments. Absent strategies to reduce confounding, imbalance on placebo variables is pervasive, in a manner consistent with the

---

4 A natural experiment is an observational setting in which causes are assigned haphazardly, and ideally in a manner that is as good as random (Dunning, 2012; Sekhon and Titiunik, 2012).
realistic Bayesian model of respondent updating. We find that abstract encouragement is not effective at reducing imbalance, and that covariate control is effective only on variables that are explicitly or implicitly controlled. As in observational studies, covariate control provides no guarantees with respect to uncontrolled characteristics, and can even exacerbate confounding under certain conditions. The ENE design is most effective at reducing imbalance, even on those covariates explicitly controlled in the covariate control design.

Overall, we conclude that while scenario-based survey experiments are powerful tools, they face identification challenges similar to observational studies. In particular, they face the risk that the apparent effects of (subjects’ beliefs about) the factor of interest are instead the consequence of (their beliefs about) other characteristics. Crucially, this is a problem of internal validity—whether the observed covariation between treatment and outcome reflects a causal relationship.\(^5\) The methods we propose can ameliorate the problem of confounding in survey experiments, but as we discuss they have limitations of their own. Random assignment in scenario-based survey experiments does not free us from having to think carefully and creatively about causal identification.

2 Why Random Assignment of Vignettes Is Not Enough

2.1 The Survey Manipulation as an Instrumental Variable

Although the issue of confounding in survey experiments has been implicitly recognized before, to our knowledge it has never been explicitly formalized. To do so, we use the framework of instrumental variables, with the survey manipulation as the instrument ($Z$), respondents’ beliefs about a feature of the scenario as the treatment ($D$), and their survey responses as the dependent variable ($Y$). Although we find IV the most natural way of conceptualizing the problem, we note that it can also be conceptualized using the framework of mediation analysis, with the experimental manipulation as the treatment and beliefs as a mediator of its effect (Imai et al., 2011).\(^6\)

In most scenario-based survey experiments, the causal question of interest is how respondents’ beliefs about a feature of the scenario $D$ affect their survey responses $Y$. In some survey experiments, however, scholars are additionally, or even primarily, interested in the effect of the manipulation itself ($Z$) on $Y$; this is known as the “intention-to-treat” (ITT) effect. For example, studies of racial cues or media framing are sometimes interested in the effects of presenting information in a certain way, regardless of the mechanism for these effects. The ITT effect is also the primary if not exclusive focus of methodologically oriented survey experiments, such as Berinsky, Margolis and Sances’s (2013) comparison of methods for maintaining respondents’ attention. Although focusing on the ITT effect has the advantage of avoiding the problem of confounding (assuming that the survey manipulation is randomized), in many cases this focus is likely to be unsatisfying. In Tomz and

\(^5\)Shadish, Cook and Campbell (2002, 38). External validity raises issues of its own (e.g., Barabas and Jerit, 2010).

\(^6\)We caution, however, that the typical approach to mediation analysis does not offer a solution, as it invokes much stronger assumptions than is necessary or plausible for this setting, namely ignorable variation in $D$, conditional on $Z$ and other controls. This is equivalent to assuming one has already overcome the problem of confounding.
Weeks (2013), for example, the real-world phenomenon of interest is not citizens’ reaction to learning a target country’s regime-type (which could be correlated with its wealth, religion, etc.), but rather their reaction to the country’s regime-type itself (independent of its wealth, religion, etc.).

Even when ultimately interested in the effects of respondents’ beliefs, survey experimenters usually estimate only the ITT, under the implicit logic that the ITT provides information about the effects of beliefs. Typically, the unstated premise is that the ITT has the same sign as the treatment effect. An IV framework helps clarify the assumptions required for this logic to hold (i.e., for Z to be a valid instrument for D). For ease of exposition, we assume that Z and D are both dichotomous. Define the complier average causal effect (CACE) as the average effect of D on Y among those respondents whose value of D is affected by Z. As is well known, the CACE is nonparametrically identified under the following assumptions (e.g., Angrist, Imbens and Rubin, 1996):

- **A1 Consistency and No-Interference:** There is no variation in versions of treatment, and each unit’s potential outcomes are not affected by other units’ values of Z and D.

- **A2 Independence of the Instrument:** Z is independent of potential treatment assignments and potential outcomes.

- **A3 Strong Monotonicity:** The sign of the effect of Z on D for every respondent is known, and for some respondents it is non-zero.

- **A4 Exclusion Restriction:** Z affects Y only through D and not through other causal pathways

Under assumptions A1–A4, the ITT lies between the CACE and zero, and a significantly positive (negative) ITT estimate provides evidence that the CACE is positive (negative). If D can be measured, the CACE itself can be estimated using an IV estimator (see Section 5.3 for further discussion).

Assumption A1, also known as the Stable Unit Treatment Value Assumption (SUTVA), is usually a prerequisite for estimating causal effects in any context. Assumption A2 is easily satisfied in survey experiments as long as the survey content is randomly assigned. Assumption A3, though not guaranteed, is typically very plausible. It would be violated if the manipulation did not influence respondents’ beliefs about the causal factor, which in many cases can be assessed using a manipulation check (e.g., Mutz, 2011, 82). Strong monotonicity would also be violated if, for example, labeling a country “democratic” caused some respondents to believe that it is undemocratic.

Our key concern is A4, the exclusion restriction. A4’s plausibility hinges on how subjects react to the survey manipulation. Specifically, it stipulates that the manipulation does not affect the survey responses except by changing subjects’ beliefs about the key feature of the scenario D. This would be violated if the manipulation altered subjects’ beliefs about background features of the scenario that mattered to their response (Y), where “background feature” is defined as a feature of the scenario that the subject believes is causally prior to the

---

7 One reason for the typical focus on the ITT is that the treatment variable (respondents’ beliefs) is often difficult to measure.
treatment $D$ in the scenario. For example, in the democratic peace example, the exclusion restriction stipulates that the label “democracy” may affect whether subjects believe the target country is a democracy, as well as their beliefs about country attributes affected by regime type such as whether the country has freedom of the press. But subjects should not update their beliefs about background attributes not influenced by regime-type, such as the country’s geographic location, unless those attributes have no effect on the subjects public support for war.

Figure 2 illustrates why violation of the exclusion restriction leads to confounding.\textsuperscript{8} In the left panel, the random survey manipulation $Z$ affects only $D$, making it a valid instrument that can be used to estimate $D$’s effect on $Y$. Even if $D$ is not observed, the ITT effect of $Z$ on $Y$ is guaranteed to have the same sign as the effect of $D$ on $Y$ (given binary $Z$ and $D$ and assumptions $A_1$–$A_3$). By contrast, in the right panel $Z$ also affects $e$: respondents’ beliefs about background features of the scenario that influence their response $Y$. Since $e$ affects $Y$ and shares a common cause ($Z$) with $D$, it confounds the causal effect of $D$ on $Y$. Due to this, $Z$ is not a valid instrument for $D$ and the ITT effect of $Z$ on $Y$ bears no necessary connection to the effect of $D$ on $Y$. A positive ITT, for example, could indicate a positive effect of $D$, but it could also be the result of a positive effect of $e$ that swamps a negative effect of $D$. In short, randomization in survey experiments identifies the effect of $D$ only if the survey manipulation does not influence respondents’ beliefs about background features of the scenario that are relevant to the outcome.

\textsuperscript{8}The causal diagrams in Figure 2 presume that assumptions $A_1$–$A_3$ are satisfied. For simplicity they omit common causes of $D$ and $Y$, the existence of which is what motivates an experimental approach in the first place.
2.2 Models of Respondent Beliefs

Whether the exclusion restriction is violated, and thus whether the effect of $D$ is confounded, depends on how respondents update their beliefs in response to the survey manipulation. When they see the label “democracy,” for example, do respondents’ update their beliefs about other attributes associated with (but not caused by) democracy? To formalize, let $p(e|Z, X)$ be a probability measure that represents subjects’ beliefs about background causes of the outcome, given their assigned value of the instrument ($Z$) and any other information provided in the vignette ($X$). The exclusion restriction hinges on whether $p(e|Z = 0, X)$ differs from $p(e|Z = 1, X)$. Different models of respondent behavior suggest different answers to this question. Below, we consider two such models: the no-confounding model implicitly invoked by most scenario-based survey experiments, and a realistic Bayesian model in which respondents’ update their beliefs based on real-world associations.

2.2.1 No-Confounding Model

Most analyses of scenario-based survey experiments are implicitly predicated on what we call a “no-confounding” model. Under this model, subjects do update their beliefs in response to new information, but they do so in a limited way, updating only beliefs about the attribute actually named (and its causal descendants). Under this model, respondents told that a country is “a democracy” will change their beliefs about regime-type accordingly, but they will not use this information to update their beliefs about other background attributes. An implication of the no-confounding model is that whether a respondent is assigned to be treated ($Z = 1$) or not ($Z = 0$) should have no effect on their beliefs about $e$ (background causes of $Y$). That is, the model implies $p(e|Z = 1, X) = p(e|Z = 0, X)$. As our example applications demonstrate, the no-confounding model appears to be generally implausible.

2.2.2 Realistic Bayesian Model

As an alternative to the no-confounding model, we propose what we call a “realistic Bayesian” model. This model has two components. First, it holds that the relevant prior beliefs of survey respondents are realistic, in that they reflect the relationships among different attributes in the real world. For example, absent additional information, respondents should believe that a country described as a “democracy” is more likely to be in Europe than one described as “not a democracy”, for the simple reason that democracy and European location are positively correlated in the real world. Formally, let $p(e, D)$ denote subjects’ prior beliefs about the joint distribution of background attributes and the factor of interest, and let $f(e^*, D^*)$ denote the empirical joint distribution of their real-world analogues. The realistic Bayesian model stipulates that $p(e, D) = f(e^*, D^*)$.

The second component of this model is that survey respondents are Bayesian updaters—that is, given their priors, they respond to new information by updating their beliefs according to the laws of conditional probability. A realistic Bayesian model predicts that respondents will in general react to survey manipulations by updating their beliefs about any attribute that in the real world is correlated with the information provided in the survey manipulation. The model thus in general implies that $p(e|Z = 1, X) \neq p(e|Z = 0, X)$. For example, upon being told that a country is a democracy, a respondent will also update
their beliefs about attributes associated with democracy, such as being located in Europe or majority-Christian. Only if they perceive the country’s regime-type to be independent of (and thus to convey no information about) its background attributes will respondents not update their beliefs about those attributes.

The advantage of the realistic Bayesian model is that it predicts not only when confounding will occur, but also what form it will take. Specifically, it predicts that confounding in survey experiments should look a lot like confounding in observational studies, since the same associations that confound observational studies determine how respondents update their beliefs. This specificity is valuable because it enables scholars to formulate precise predictions about the probable form of confounding, and design their survey experiment so as to diagnose and ameliorate it. We acknowledge, however, that the realistic Bayesian model is only an approximation to how actual respondents process information (for various perspectives, see, e.g., El-Gamal and Grether, 1995; Holyoak and Cheng, 2011).E We also note that alternative informational models, such as ones based on heuristics (e.g., Kahneman and Tversky, 1973) or stereotypes (e.g., Gilliam and Iyengar, 2000), might be more appropriate in some contexts. Models of this sort would also predict confounding, though of a different form. F We hope that future research explores these questions further. We now turn to a discussion of how to diagnose and address confounding.

3 Diagnosing Confounding

Thus far we have shown that random assignment of survey content justifies inferences about the effects of respondents’ beliefs only insofar as the survey manipulation is a valid instrument for those beliefs. The manipulation’s validity as an instrument hinges on whether the exclusion restriction is violated: whether the manipulation affects background causes of the outcome (e.g., location) that are not themselves consequences of the causal factor of interest (e.g., democracy). Fortunately, the exclusion restriction assumption has observable implications that can be tested empirically. In this section, we explain how placebo tests can be used to diagnose confounding and then describe three techniques for ameliorating it.

3.1 Placebo Tests

A placebo test is a test of an effect that is known to be zero under the assumptions of the analysis (e.g., Rosenbaum, 2002, 214–21). We will devise placebo tests for the exclusion restriction $A_4$. Specifically, we will ask about aspects of the scenario that are candidate background features of the scenario $(e)$, beliefs about which could plausibly influence the subjects’ response $Y$. A good placebo attribute satisfies three criteria:

- $C_1$: The respondent does not believe that the placebo attribute is affected by the factor of interest. If the respondent believes that the placebo attribute is influenced by the factor of interest, then rejection of the placebo test does not necessarily indicate violation of the exclusion restriction. Rather, it could reflect the fact that the placebo attribute is part of the causal pathway of interest. The easiest way to satisfy this criterion is for the placebo attribute to be “pre-treatment” in scenario time since we assume that respondents know that causality only flows forward in time.
• $C_2$: If confounding is present, the manipulation $Z$ will affect respondent beliefs about the placebo attribute. $C_2$ states that if the experimental manipulation affects the survey responses $Y$ through causal pathways that don’t include $D$, then $Z$ should affect the placebo. The easiest way to satisfy this criterion is to find attributes on the confounding causal pathways. Under our model of Realistic Bayesian beliefs, these will be factors that are correlated with the factor of interest in the real world. Specifically, we screened our candidate placebo characteristics based on whether they are correlated with regime-type in the real world (see Table 3).

• $C_3$: Beliefs about the placebo attribute affect the subject’s response $Y$. The exclusion restriction stipulates that the instrument affects the outcome only through the causal factor of interest. A placebo attribute that does not affect relevant outcomes cannot confound the treatment effect, even if beliefs about the attribute are indeed influenced by the manipulation, not through treatment. Absent an alternative causal pathway between $Z$ and $Y$, $Z$ remains a valid instrument for $D$ (given assumptions $A_1$–$A_3$).

A placebo test based on an attribute that satisfies criterion $C_1$ will be statistically valid: when there is no confounding it will not reject more than the size of the test. A test based on an attribute that satisfies criteria $C_2$ and $C_3$ will be statistically powerful: when there is confounding it will be likely to reject. The most informative placebo tests will be both valid and powerful. While satisfying all three criteria is ideal, in practice placebo tests will satisfy some criteria more than others. For example, many of the most salient real-world confounders of the democratic peace, such as countries’ wealth, alliances, and trade relationships, are at least potentially affected by regime-type ($C_1$). On the other hand, some attributes that are clearly not influenced by regime-type, such as geographic region, may not have a strong effect on public support for war ($C_3$). And some causes of support for war, such as the target country’s military spending, are not strongly associated with democracy ($C_2$). Negotiating the trade-offs between these criteria requires careful ex ante theorizing as well as suitable caution in interpreting the results. Nevertheless, to the extent that they satisfy these conditions, placebo tests are a powerful tool for diagnosing whether the exclusion restriction is violated and thus inferences are confounded.

4 Ameliorating Confounding

While it is important to diagnose confounding if it exists, it is better to prevent it to begin with. Here, we discuss three strategies for ameliorating confounding. The first two strategies, abstract encouragement and covariate control, have already been employed in a number of survey experiments. The final strategy, an embedded natural experiment, is our own invention. After describing these strategies in this section, we then move to an example in which we compare their performance.
4.1 Abstract Encouragement

Perhaps the simplest strategy for ameliorating confounding in survey experiments is what we label *abstract encouragement*. Abstract encouragement asks respondents to consider the scenario or vignette in abstract terms, using a statement along the lines of, “For scientific validity the situation is general, and is not about a specific country in the news today” (Kreps and Maxey, 2015; Tingley and Tomz, 2012; Tomz and Weeks, 2012, 2013). The primary argument in favor of abstract designs has been that they can yield more externally valid or generalizable results (Mutz, 2011, 158; Tomz and Weeks, 2013, 860; though others have argued that more concrete detailed scenarios yield more externally valid results). However, a scholar might also argue that abstract designs reduce the problem of confounding by encouraging respondents to not use real-world data to inform their beliefs about the scenario. Consistent with the realistic Bayesian model, we predict that abstract encouragement will not reduce confounding.

4.2 Covariate Control

The second strategy for ameliorating confounding is what we call *covariate control*, which is both more common and more explicitly aimed at confounding than abstract encouragement. Indeed, to the extent that survey-experimental studies have recognized the problem of confounding, they have mainly addressed it through this strategy. In a covariate control design, the survey vignette includes additional details designed to fix respondents’ beliefs about covariates that might confound the causal effect of interest. For example, Tomz and Weeks (2013) seek to rule out salient potential confounders by providing details about the target country’s military capabilities and alliance and trade relationships with the United States.

In some studies, the additional details are identical across experimental conditions, but in others the main survey manipulation is crossed with variation in the controls. A particularly elaborate form of the latter version of covariate control is conjoint analysis, a high-dimensional factorial experiment that varies many attributes of the vignette across experimental conditions (Hainmueller, Hopkins and Yamamoto, 2014). Hainmueller and Hopkins (2014), for example, use conjoint analysis to examine how hypothetical immigrants’ country of origin, education, profession, and language skills influence the native citizens’ immigration attitudes.

Based on our realistic Bayesian model, we anticipate that covariate control designs will operate in a manner similar to covariate control in observational studies: they will reduce or eliminate confounding on the controlled variables and perhaps on related variables, but it will not address confounding on characteristics not correlated with the controls. In fact, it can even amplify bias, for example if one controls for a characteristic affected by treatment. In short, we anticipate that covariate control will typically provide only a partial solution to confounding in survey experiments.
4.3 Embedded Natural Experiments

Finally, we also introduce a third strategy of our own invention: the *embedded natural experiment* (ENE). This strategy is motivated by the realistic Bayesian model, which predicts that the survey manipulation will influence respondents’ beliefs about background attributes *unless* they perceive that the content of the manipulation is statistically independent of—and thus conveys no information about—those attributes. In short, a Bayesian will not update their beliefs only if they believe the treatment was as good as randomly assigned.

The survey manipulation itself is, of course, random, but the crucial question is whether respondents perceive the assignment of the causal factor in the scenario to be (as-if) random. In the absence of information indicating that it was, a realistic Bayesian respondent will rely on their prior knowledge of how the treatment in question is usually assigned in the real world—which in nearly every context is likely to be non-random. The crux of the ENE design is giving respondents additional information that leads them to believe that treatment exposure in the scenario was as good as random. The design does so by embedding in the scenario a description of a natural experiment in which treatment assignment is as-if random (cf. Dunning, 2012; Sekhon and Titiunik, 2012).

The most straightforward ENEs involve a lottery or other form of transparent random process. Consider, for example, a survey experiment that examines whether subsidizing childcare increases employees' willingness to take a time-consuming promotion (Latura, 2015). Simply manipulating whether a hypothetical firm is described as subsidizing childcare will probably not isolate the effect of interest because respondents know that some kinds of firms (e.g., ones with a family-friendly culture) are more likely to offer this policy, and these inferences may affect their decision whether to accept the promotion. In the ENE version of this experiment, which we discuss in more detail later, the firm is described as having a limited number of subsidized childcare slots that are assigned by a random lottery; the survey manipulation is whether the respondent wins the lottery. Assuming respondents perceive the lottery outcome to be truly random, they should not update their inferences about the background attributes of the firm.

More generally, ENEs may involve any treatment assignment mechanism that is at least approximately independent of background attributes. In many cases, these will involve incidents or phenomena that, if not strictly random, are at least accidental. Examples include the outcome of an assassination attempt (cf. Jones and Olken, 2009), or an episode in which two fighter jets either collide or barely miss each other. ENEs based on other quasi-experimental designs, such as regression discontinuity, are also possible. In practice, ENEs will fall somewhere on a spectrum of as-if randomness, just as observational natural experiments do (Dunning, 2012).

The critical criterion for evaluating ENE designs is not whether the ENE is strictly random, but whether respondents’ perceive it to be independent of background attributes and update their beliefs accordingly. As we have described, this criterion can be tested empirically using placebo tests. In general, we expect that well-designed ENEs will exhibit less evidence of confounding than abstract encouragement or covariate control designs. Unlike covariate control designs, which should be expected to balance only explicitly controlled attributes and their close relatives, ENEs should balance beliefs about all background attributes, regardless of whether they are explicitly controlled. This, of course, is the essential advantage of design-
based causal inference over “selection-on-observables” identification strategies. Nevertheless, it should be noted that ENE designs are not always easy or even possible to construct, and they often raise questions of generalizability and interpretation. We discuss these limitations later in the paper. We now turn to empirical examples in which we apply the tools and designs we have described thus far.

5 An Application to the Democratic Peace

We evaluate confounding, and the effectiveness of the various responses to it, using several applications. The first and most developed is a replication and extension of Tomz and Weeks’s (2013) survey experiment on the mass basis for the democratic peace (cf. Johns and Davies, 2012; Mintz and Geva, 1993). Using placebo tests, we show that randomly manipulating whether a target country is described as democratic is not sufficient to prevent respondents from updating their beliefs about the background attributes of the country, potentially confounding the causal estimate. We further demonstrate that abstract encouragement does little to mitigate this risk of confounding, and that covariate control does so only on attributes explicitly or indirectly controlled in the vignette. An embedded natural experiment design is most effective at reducing confounding. We also demonstrate how an IV estimator can be used to estimate the local average treatment effect, and discuss relevant assumptions and issues of interpretation.

5.1 Survey Design

On July 1–3, 2015, we used the Qualtrics survey platform to survey 3,080 Americans recruited through Amazon’s Mechanical Turk (MTurk).\(^9\) The basic setup of our survey experiment hews closely to Tomz and Weeks (2013). We presented respondents with a scenario in which a country is developing nuclear weapons, randomly manipulated whether the country is described as a democracy, and asked whether respondents supported using military force against the country (among other questions). In addition to the main manipulation (democracy/non-democracy), we also varied experimental conditions on two other dimensions designed to assess the effectiveness of different strategies for ameliorating confounding. The first dimension was whether respondents were assigned to receive abstract encouragement. The second dimension consisted of three versions of the vignette: a basic vignette that provided respondents with little information about the country besides the democracy manipulation; a covariate control vignette that included details about the target country; and an embedded natural experiment that described an assassination attempt as a source of as-if random variation in regime-type.

In the basic design, respondents first read the scenario background:

(S1) A country is developing nuclear weapons and will have its first nuclear bomb within six months. The country could then use its missiles to launch nuclear attacks against any country in the world.

\(^9\)See Appendix D for a complete description of the survey design and Appendix E for the full summary of our analysis. Our study pre-registration and pre-analysis plan can be found at EGAP.
Respondents then read a description of the country’s regime-type, randomly manipulated to be democratic or non-democratic:

\[ Z_{\text{basic}} \] [The country is not a democracy and shows no sign of becoming a democracy. / The country is a democracy and shows every sign that it will remain a democracy.]

Finally, respondents read the conclusion of the scenario:

(S2) The country’s motives remain unclear, but if it builds nuclear weapons, it will have the power to blackmail or destroy other countries. The country had refused all requests to stop its nuclear weapons program.

The covariate control design was identical to the basic design, except that after \( Z_{\text{basic}} \) respondents read information about the country’s military capabilities, trade, and alliances. The text of these controls was taken from Tomz and Weeks (2013), and like them we randomly varied the values of these details.

The embedded natural experiment design began with a description of an ENE in which the regime type was manipulated as follows:

\[ Z_{\text{ENE}} \] Five years ago a country, Country A, was a fragile democracy. It had a democratically elected government, headed by a popular president. At the time, a well-researched U.S. State Department report concluded that without this president, there was a very high probability that the country’s military would overthrow the government to set up a dictatorship.

Two years ago at a public event, a disgruntled military officer shot at the president of Country A. [The president was hit in the head and did not survive the attack. In the political vacuum that followed the president’s death, the country’s military overthrew the democratically elected government. Today, Country A is a military dictatorship. / The president was hit in the shoulder and survived the attack. The country’s democratically elected government survived the political turmoil. Today, Country A is still a democracy.]

After reading the vignette, respondents were asked about their support for using force against the target country, as well as demographic questions and questions related to the placebos, potential mechanisms, and the treatment. The order of all these questions was randomized; we did not find any question order effects on the outcome of interest (see Appendix D and E.8).

In order to conduct placebo tests, we asked respondents about their beliefs regarding the following background attributes of the target country: region, GDP, religion, race, oil reserves, alliance with the U.S., trade with the U.S., joint military exercise with the U.S., FDI in the U.S., and military spending. All of these variables except the last were selected based on criteria \( C_1 - C_3 \): all are at least partly pre-treatment, are correlated with regime-type in the real world, and plausibly affect public support for military action. The exception is military spending, which is not significantly related to regime-type in the real-world; we tested this variable because the vignettes in Tomz and Weeks (2013) control for nonnuclear military capabilities. To minimize the risk of placebo attributes’ being affected by treatment, the questions asked respondents what they thought the attributes were ten years in the past.
5.2 Placebo Tests

Figure 3 summarizes the main results for the placebo tests (for more details, see Appendix E). Overall, the results are inconsistent with the no confounding model, and consistent with the realistic Bayesian model. The basic design exhibits evidence of pervasive confounding: for every placebo variable, mean equality between the two experimental conditions can be rejected at the 5% level. Moreover, the direction of the imbalance is consistent with the realistic Bayesian model. Respondents told that the country is a democracy are more likely to perceive it as having the characteristics associated with democracies in the real world, such as being more likely to have higher GDP per capita, to have populations that are majority Christian and white, to not have large oil reserves, to have an alliance with the U.S. and have conducted a joint military exercise with the U.S., and to trade with and invest in the U.S. Across all vignette versions, respondents assigned to receive abstract encouragement exhibited similar imbalance as those not assigned, suggesting that as implemented abstract encouragement is ineffective at reducing confounding.
Hollow points indicate the standardized average difference between the democratic and non-democratic treatment conditions. In this coefficient plot and all following ones, we report the 95% and 99% confidence intervals estimated using heteroscedasticity-robust standard errors. Background attributes that were explicitly mentioned in the covariate control design are indicated with *, and ** indicates attributes implicitly controlled.
Like the basic design, the covariate control design exhibits large imbalances on placebo attributes that were not controlled (region, GDP, religion, race, and oil reserves). On attributes that were explicitly (alliance and trade) or indirectly (joint military exercise and FDI) controlled, the imbalance is less extreme, but it was almost never completely eliminated. The covariate control design did succeed in eliminating imbalance on military spending, but even in the basic design this was the least-imbalanced attribute, probably because democracy has no clear real-world relationship with military spending.

The ENE design was clearly most effective at reducing imbalance on placebo attributes. Only three out of ten placebo variables—region, religion, and race—are significantly imbalanced at the 5% level, and in all three cases the imbalance is far less severe than for the other two designs. Strikingly, even attributes that the covariate control vignette explicitly mentioned were more balanced in the ENE design. The ENE’s limited impact on placebo attributes is not a symptom of a weak manipulation overall. The ENE manipulation’s effect on perceived regime type was nearly as large as the other designs’ (Figure 4), and its effect on support for war was if anything larger than those of the other two (Figure 5). Overall, the results suggest that just as natural experiments tend to yield more plausible causal inferences in observational studies, so too are embedded natural experiments the most effective strategy for reducing confounding in survey experiments.

5.3 Local Average Treatment Effects

The three vignette manipulations had broadly similar effects on support for the use of force: in all three, respondents told that a country was democratic were 10 to 20 percentage points less likely to support a military attack on the target country (Figure 5, top panel). Under the IV assumptions, we know that this ITT estimate has the same sign as the local average treatment effect among those whose beliefs were influenced by the survey manipulation, a.k.a., the complier average causal effect (CACE). Setting aside the foregoing concerns about the exclusion restriction, we now show how to estimate the CACE directly.

Unbiased estimation of the CACE requires that the treatment be measured without error, a critical and often non-trivial assumption. Since democracy is a complex and nuanced concept, we used a rich set of questions to measure respondents’ beliefs about it. One subset of these questions asked respondents how likely it was that the target country fell into each of five Polity scale categories, ranging from fully non-democratic to fully democratic. We summarize these questions in two ways: with a dichotomous indicator $D_D$ for whether the respondent assigned greater probability to democratic than non-democratic categories, and with a continuous measure $D_C$ consisting of the probability-weighted average of the mean Polity score of countries in each discrete category. We also included a second battery of questions asking whether respondents thought there was “more than a 50 percent chance” that the country exhibits various components or indicators of democracy, such as an elected government, a free press, or legal opposition parties (Appendix D.8, E.5.2).
Regardless of how $D$ is measured, $Z$ had a strong “first stage” effect (Figure 4 and Appendix E.5.2). This effect, however, is substantially smaller than a literal interpretation of the regime-type prompt would suggest, largely because respondents in the “democracy” condition did not perceive the target country to be especially democratic. On average, respondents in this condition gave the country an imputed Polity score of around +3 on a −10 to +10 scale, a Polity score representative of countries like Russia and Iraq. It is likely the case that respondents perceived the “democracy” to be not very democratic because their beliefs were influenced by the other information in the scenario, such as the fact that the country was developing nuclear weapons in a threatening manner. As a consequence, stating that the target “is a democracy and shows every sign that it will remain a democracy” increased the probability of $D_D = 1$ by only around 50 percentage points and $D_C$ by around 5 points on the 21-point Polity scale, in both cases likely much less than the intended counterfactual.
As a consequence of respondents’ incomplete “compliance” with their assigned regime-type, the estimated (complier average) causal effect (CACE) is about twice as large as the ITT for all three designs. The CACE for the ENE design is largest: among those affected by the manipulation, believing the country was democratic \((D_D = 1)\) decreased respondents’ probability of supporting use of force by an estimated 40 percentage points; we estimate a similarly large CACE for a change in the perceived democracy level by 10 points on the Polity scale. The treatment effect estimates for the basic and covariate control designs are significantly smaller, about 20 percentage points, but it is unclear whether this reflects true differences in the LATE or the contaminating influence of confounding.

6 Extensions to Other Studies

We have extended these methods to several other studies, two of which we summarize here. The first survey experiment, on the effects of employer-provided daycare, illustrates the
usefulness of an ENE design in a setting very different than the democratic peace example. The second experiment, a replication of Desante’s (2013) study of racial discrimination, considers the issues that arise when the real-world causal estimand is not well-defined and thus it is impossible to devise a plausible ENE design.

6.1 Effects of Subsidized Childcare

Latura (2015) examines whether people are more likely to accept a time-consuming promotion if their firm provides subsidized high-quality extended hours childcare. We performed an experiment embedded in Latura’s survey experiment (fielded April 2015 on MTurk; see Appendix G). In the basic design \( (n = 771) \), after reading about other aspects of their situation and the firm, some respondents were informed that “the company you work at subsidizes the cost of high-quality, extended-hours childcare for employees.” The ENE design \( (n = 1003) \) informed all respondents that their firm operates an “on-site, high-quality, extended-hours day-care center open from 6:00 AM to 10:00 PM on weekdays. The center is free for employees, but slots are allocated via random lottery.” The control group was then told that they did not win a day-care slot; the treatment group that they did. \(^1\)

We asked respondents three placebo questions: (1) Does the company offer employee benefits other than childcare; (2) does the company expect employees to answer work-related email on the weekends; (3) does the company help employees to balance family-work issues. As Figure 6 shows, all the placebo variables are imbalanced in the basic design. The ENE design reduces imbalance relative to the basic one, but it does not fully eliminate it. The
imbalance in the ENE design suggests either that respondents updated their beliefs in non-Bayesian fashion, or that they did not fully believe that the lottery was random with the same probability. One subtle possibility is that since we didn’t specify the probability of winning the lottery, a respondent in control could reasonably infer that there were only a few spots allocated by lottery (i.e., the lottery was a public relations stunt), whereas a respondent in treatment could infer that many spots were allocated. If this was the problem, then it reveals how careful one must be in constructing an ENE to make sure the respondents not only perceive treatment to be as-if random, but as-if random with the same probability across conditions.

6.2 Why is Latoya Discriminated Against?

Desante (2013) uses a survey experiment to study whether Americans are more willing to support welfare for people who are white than black, and why this is the case. The survey manipulation is the name of the welfare applicant (e.g., Emily vs. Latoya). The number and age of the applicant’s children are held constant. The experiment also manipulates a “Worker Quality Assessment” as being either “Poor” or “Excellent.” In so doing, this design hopes to rule out “principled conservative” reasons for discrimination, leaving only “racial animus” as the basis for discrimination (on confounding after manipulating names, see Simonsohn, 2015).

For placebo questions, we looked for characteristics related to principled conservative motivations. This is implied by DeSante’s design, since if beliefs about principled conservative characteristics are affected by the name manipulation then we cannot rule out that observed discrimination is not due to principled conservative reasons. Borrowing questions used by the North Carolina welfare agency, we asked respondents whether they thought the applicant has completed high school, has worked recently, has a criminal record, and has good parenting skills.

We include two additional questions: whether the applicant grew up in a low socioeconomic status (SES) family and whether the applicant is likely to have another child in the next two years.

In May 2014, we fielded a survey experiment using respondents recruited from MTurk (full details in Appendix F). We evaluated two designs: a basic design \( n = 156 \) that included only the applicant’s household information and welfare history, and a covariate control design that also included a worker quality assessment \( n = 312 \). In the basic design, four out of six placebo attributes were significantly imbalanced (only high-school degree and parenting skills were not). In the covariate control design, the number of imbalanced attributes was reduced to two: low SES and having another child.

These results suggest that DeSante’s control strategy reduced imbalance on most characteristics that a “principled conservative” might discriminate on (prior work experience, criminal conviction), but not on all characteristics that a person could argue is relevant to a principled allocation of welfare (low SES, intention to have children). Thus, while the results in Desante (2013) do provide insight into the reasons for racial discrimination, caution is still required when interpreting this as evidence of racial animus.
Studies of racial cues raise subtle issues about the causal estimand (cf. Sen and Wasow, 2016). This is revealed by asking what would an ENE design look like if we want to manipulate respondents’ perception of someone’s race? It’s hard to think of a process that as-if randomly assigns race, independent of “other background characteristics”, in large part because race is not a clearly defined phenomenon, in addition to it being not readily manipulated. Race is certainly more than skin-pigment, but if it is thought to include things like education, work-ethic, and intended family size then it becomes inseparable from the other bases for “principled conservative” discrimination.

One approach adopted by many scholars for studies of racial and ethnic cues is to think of them as studying the effect of the cue itself, rather than of belief about the race of the person in the scenario. The scholar then examines, through (implicit) mediation analysis, why the racial cue has the effect that it does: racial animus or for other reasons. Given the difficulty of conceptualizing race as a causal factor, this is a productive approach. Nevertheless, it confronts similar problems of inference, simply under a different vocabulary: one’s “placebo measures” are now “potential mediators” of the cue. If the racial cue affects the other mediators then we become less confident that the ITT effect is due to the mediator of interest (racial animus). Covariate control designs then would work best when they reduce other mediating causal paths and isolate the mediator of interest.

7 Limitations of Different Designs

The evidence from the preceding studies suggests that ENE designs, when feasible, are generally the best strategy for reducing confounding. Nevertheless, it is also clear that
neither the covariate control nor the ENE design is a silver bullet. We discuss the limitations of each design in turn, beginning with covariate control.

### 7.1 Limits of Covariate Control designs

While covariate control did not eliminate confounding in our example applications, it did reduce imbalance on those variables that were specified. Is the solution then simply to devise extremely detailed scenarios that specify every possible confounder? Unfortunately, such a strategy is unlikely to work, for four reasons.

The first is **respondent exhaustion**: the longer and more complex a scenario is, the more likely respondents are to satisfice and read less closely (Krosnick, 1999). Hainmueller, Hopkins and Yamamoto (2015), for example, find that in conjoint experiments satisficing is substantial with as few as four characteristics. While researchers can always emphasize the manipulation of greatest interest, this may simply redirect attention away from other details.

A deeper limit to elaborate covariate control is the **plausibility constraint**: as the number of controls increases, so too does the probability of a counterfactual comparison that is implausible to respondents. As in observational studies, the more variables we control for, the more likely it is for a counterfactual to require extrapolation beyond the support of the data (King and Zeng, 2006). There is, for example, simply no empirical referent for a Western European country that uses Sharia law for criminal proceedings. Researchers can prune away implausible vignettes (Hainmueller, Hopkins and Yamamoto, 2014, 20), but as the number of control variables increases the subset of plausible combinations will likely become smaller — until it is empty.

A different strategy is to use a **proper noun vignette**: specifying a real-world referent in the scenario, thus implicitly controlling for an almost infinite number of background attributes. For example, in another survey experiment we provided selective information about a country’s past foreign-policy behavior; in one version the country was identified as Iran. We found that naming the country reduced imbalance on background attributes such as the country’s regime type, but it did not eliminate it. This is likely to always be the case since the vignette is a hypothetical, allowing for the possibility of other unspecified changes in background covariates. Any confounding bias will also likely be more severe the less respondents know about the real-world referent, since then the respondents can infer more from the vignette.

A variant of this strategy that avoids hypotheticals, which we label a **real-history design**, entails describing a real historical episode and selectively informing or reminding respondents of certain facts about that episode. Such a strategy has promise, but it is limited by the kinds of scenarios generated by the real world. Further, to the extent that respondents are not perfectly informed about the scenario, they can still reasonably draw inferences or be reminded about background features of the real scenario in a manner that could induce confounding.

Finally, we note that covariate control can create or amplify bias as well as reduce it. The most obvious way it can do so is by controlling for consequences of the real-world factor of interest, which can induce post-treatment bias. In many survey experiments, this bias is difficult to fully avoid because the outcome is often premised on the occurrence of a particular scenario (e.g., a country seeking nuclear weapons) that may be influenced by
treatment (e.g., whether a country is democratic). Bias can be created in more subtle ways as well, including involving controls for background characteristics. Consider a version of the democratic peace experiment that controls for geographic region by setting the scenario in the Middle East. Doing so would exacerbate confounding due to religion because the negative correlation between being democratic and being majority-Muslim is even stronger in the Middle East than in the world as a whole. Again because of the close mapping between confounding in survey experiments and confounding in observational studies, for guidance on how to select controls in survey experiments researchers can turn to advice for doing so in observational studies.

In summary, covariate control is a promising technique for reducing confounding in survey experiments, but it has substantial limitations. Trying to control for many characteristics could exhaust the respondent, lead to implausible counterfactuals, and could actually induce or amplify biases if care is not taken in the selection of control details. Most importantly, the covariate control design does not reduce confounding on characteristics dissimilar to the controls and does not even necessarily fully balance beliefs on those details that are specified.

7.2 Limits of Embedded Natural Experiments

Embedded Natural Experiments also face limits in their application. While our theory and empirics suggest that, when well constructed, they can overcome all kinds of confounding, they (1) are often hard to construct and (2) shape the causal estimand in consequential ways.

In creating these surveys, we brainstormed many possible hypothetical natural experiments. We rejected almost all of them. They were: not plausibly as-if random, too subtle or complicated, distractingly uncommon, or unlikely to generate large enough effects on regime-type (recall that IV bias is larger for weak instruments). We were simply unable to think of a plausible strong natural experiment for which the “democracy” level would be a country like Belgium and the “non-democracy” level a country like Egypt (let alone North Korea), because the real world has not produced such natural experiments. Just as with observational data it is not possible to estimate the effect of making Belgium a dictatorship without strong assumptions, it may not be possible to get at this implausible counterfactual with a plausible ENE. Thus, the plausibility constraint may bind as much on ENEs as on covariate control designs.

An additional concern about ENE designs is that they only allow us to estimate the local causal effect for the kinds of countries that fit the ENE scenario. By contrast, abstract covariate control designs appear to allow us to estimate more general causal effects. We believe this concern is largely misguided. It is correct that the ENE design only estimates a local causal effect; in our case, for example, the respondents perceive the ENE scenario to be much more typical of the Middle East and North Africa than Western Europe (Figure 11). We would thus want to note that our estimate of the average treatment effect of regime-type is more applicable to changes in regime-type in the Middle East, than to changes in Europe. However, as Figure 11 demonstrates, respondents in the basic and covariate control designs also found the scenario more typical of the Middle East and North Africa than Western Europe.

In general, a satisfactory covariate control design is likely to give us a similarly local causal effect, though it may be less obvious to the researcher. If we pursue an abstract scenario,
then after controlling for sufficient details and retaining only plausible combinations, we will be restricted to a limited region of the covariate space, namely the space for which there is variation in the treatment, holding the controls fixed. This is often a narrow domain. The same problem arises if we use proper nouns ("Iran") in our covariate control design, since we will be restricted to those proper nouns for which both counterfactuals are somewhat plausible. In short, when an ENE is plausible, it is usually preferable to alternative designs.

8 Conclusion and Recommendations

A well-implemented experiment allows us to identify the causal effect of that which was randomly assigned. But we usually want to go beyond that to identifying the effect of some specific causal factor: the active drug in a medicine, not a placebo effect induced by the pill shape (Morton and Williams, 2010, §3.2; Gerber and Green, 2012, §2.7.1). To do so we must assume that the experimental effect only operates through the intended causal channel. However, assumptions can and should be tested. The results of these placebo tests can then be used to improve experimental design.

In this paper we did this for scenario-based survey experiments: articulating the necessary assumptions, theorizing how they are likely to be violated, examining their testable implications, and evaluating the performance of several experimental designs. We found that the threat to internal validity is substantial in survey experiments studying the effects of respondent beliefs about a scenario. Further, we show how the potential biases have a very specific structure; we can use this knowledge to anticipate problems and improve our experiments. Specifically, the nature of the problem and solutions bears a close similarity to the problem of and solutions for confounding in observational studies. Best practice for survey experiments accordingly looks similar to best practice for observational studies. We articulate this through the following recommendations:

A. **Theorize confounding.** Think about the kinds of background characteristics that could be systematically correlated with treatment and cause the outcome.

B. **Measure your causal factor.** This can be used to evaluate the assumption of a monotonic (or known) first stage, to estimate (local) average treatment effects, and to understand the kinds of variation in $D$ that are informing your estimates.

C. **Employ a credible design.** Find a credible hypothetical natural experiment that you can embed into your scenario, and for which the resulting causal effect is relevant.

D. **Control for confounds.** If you can’t employ an embedded natural experiment, employ Covariate Control designs to reduce the risk of the worst kinds of confounding.

E. **Diagnose confounding.** Employ placebo tests to evaluate whether confounding still seems to be present, and if so, what it looks like.

F. **Theorize the bias from confounding.** Think through, informally or formally, the direction and size of biases likely to come from any remaining confounding. A causal estimate will be more compelling if you can persuasively argue that the bias is likely to be small or in the opposite direction as your prediction.
G. **Qualify your inferences.** Acknowledge the remaining risk of confounding biases.

Recognize that your estimated causal effects are local to the kinds of scenarios that you presented and the respondents’ inferences about the context of the scenario.

Scenario-based survey experiments are extremely valuable tools for social science. They allow us to study important causal questions that are otherwise elusive. But they are not a panacea. To best use survey experiments, we need greater awareness of their problems and the full range of solutions.
References


Simonsohn, Uri. 2015. “How to Study Discrimination (Or Anything) with Names; If You Must.”
URL: http://datacolada.org/2015/04/23/36-how-to-study-discrimination-or-anything-with-names-if-you-must/


# A Table of Contents

1 Introduction 2

2 Why Random Assignment of Vignettes Is Not Enough 6
   2.1 The Survey Manipulation as an Instrumental Variable 6
   2.2 Models of Respondent Beliefs 9
      2.2.1 No-Confounding Model 9
      2.2.2 Realistic Bayesian Model 9

3 Diagnosing Confounding 10
   3.1 Placebo Tests 10

4 Ameliorating Confounding 11
   4.1 Abstract Encouragement 12
   4.2 Covariate Control 12
   4.3 Embedded Natural Experiments 13

5 An Application to the Democratic Peace 14
   5.1 Survey Design 14
   5.2 Placebo Tests 16
   5.3 Local Average Treatment Effects 18

6 Extensions to Other Studies 20
   6.1 Effects of Subsidized Childcare 21
   6.2 Why is Latoya Discriminated Against? 22

7 Limitations of Different Designs 23
   7.1 Limits of Covariate Control designs 24
   7.2 Limits of Embedded Natural Experiments 25

8 Conclusion and Recommendations 26

A Table of Contents 32

B Endnotes 34

C Literature Review 37
   C.1 Review of Articles in Top Journals 37
      C.1.1 Classifying the Articles 37
      C.1.2 Articles Reviewed 39
      C.1.3 Summary Statistics 43
D “Democratic Peace” Survey Experiment Details

D.1 Outline of the Survey .......................................................... 43
D.2 Three Vignette Types .......................................................... 44
  D.2.1 Basic .............................................................................. 44
  D.2.2 Covariate Control ......................................................... 44
  D.2.3 Embedded Natural Experiment ...................................... 45
D.3 Support for Force and Mediation Questions ......................... 46
D.4 Together-Placebos Design and Separated-Placebos Design ...... 46
D.5 Survey Questions ............................................................... 47
  D.5.1 Justifications for Placebo Test Questions ....................... 47
D.6 Correlation Between Democracy and Percent Muslim .......... 51
D.7 Placebo Test Questions ....................................................... 52
  D.7.1 Notes on Placebo Test Questions ................................... 52
  D.7.2 Text of Placebo Test Questions ..................................... 52
D.8 Treatment Measures .......................................................... 55
D.9 Support for Military Action .................................................. 57
D.10 Mediation Questions .......................................................... 57
  D.10.1 If the U.S. attacked..................................................... 58
  D.10.2 If the U.S. did not attack ............................................... 58
  D.10.3 Morality of Using Force .............................................. 58
D.11 Demographics Questions ................................................... 59
  D.11.1 Education ................................................................. 59
  D.11.2 Political Party ........................................................... 59
  D.11.3 Age ............................................................................. 59
  D.11.4 Sex ............................................................................. 60
  D.11.5 Political Ideology ....................................................... 60

E Full Summary of “Democratic Peace” Survey Results ............ 60

E.1 Survey Procedure ............................................................... 60
E.2 Balance Tests .................................................................... 60
E.3 Coding Placebo Test Results .............................................. 63
E.4 Placebo Test Questions ....................................................... 65
  E.4.1 Analysis of Placebo Test Responses ............................... 65
  E.4.2 Placebo Test Results .................................................... 65
  E.4.3 Joint Test of Placebo Outcomes Using NPC ................... 80
E.5 Treatment Measures .......................................................... 81
  E.5.1 Coding and Analysis of Treatment Measure Results ....... 81
  E.5.2 Treatment Measures Results ......................................... 82
E.6 Substantive Outcome: Support for War ................................ 86
  E.6.1 ITT and IV Analyses ..................................................... 86
  E.6.2 ITT and IV Results ....................................................... 87
E.7 Abstract Encouragement Design ......................................... 90
E.8 Question Block Order Does Not Affect Substantive Outcome Measure .......... 93
  E.8.1 Question Block Order in Together-Placebos Design ........ 93
  E.8.2 Placebo Test Questions Do Not Affect Substantive Outcome Measure 94
E.9 Substantive Outcome Question Does Not Affect Severity of Imbalance in Placebo Outcomes .................................................. 98
E.10 Attention Checks ................................................................ 100

F Replication and Expansion of DeSante (2013) ....................... 103
F.1 Survey Overview .............................................................. 103
F.2 Text of the Survey ............................................................ 103
   F.2.1 Age ........................................................................... 103
   F.2.2 Introduction .............................................................. 103
   F.2.3 Welfare Application .................................................. 103
   F.2.4 Allocation of Money to Applicants ................................ 105
   F.2.5 Pop-up Window .......................................................... 106
   F.2.6 Placebo Test Questions ............................................. 106
   F.2.7 Demographics Questions ........................................... 107
   F.2.8 Racial Resentment Measure ....................................... 108
F.3 Results of Placebo Test Questions ...................................... 109
   F.3.1 Coding and Analyzing the Placebo Outcomes ............... 109
   F.3.2 Placebo Test Results ................................................. 109

G Latura’s (2015) Survey Experiment ....................................... 111
G.1 Overview of Experiment ................................................. 111
G.2 Text of the Survey .......................................................... 112
   G.2.1 Demographics Questions .......................................... 112
   G.2.2 Directions ............................................................... 113
   G.2.3 Experimental Vignettes ............................................ 113
   G.2.4 Substantive Outcome Question ................................. 114
   G.2.5 Placebo Test Questions ............................................ 114
   G.2.6 Mechanisms ............................................................ 114
   G.2.7 More Demographics Questions ................................. 114
G.3 Results of Placebo Test Questions ...................................... 116
   G.3.1 Coding and Analyzing the Placebo Outcomes ............... 116
   G.3.2 Placebo Test Results ................................................. 116

B Endnotes

Notes

A “As republics emerge (the first source) and as culture progresses, an understanding of the legitimate rights of all citizens and of all republics comes into play; and this, now that caution characterizes policy, sets up the moral foundations for the liberal peace.” and “Domestically just republics, which rest on consent, then presume foreign republics also to be consensual, just, and therefore deserving of accommodation.”

B Exactly which others factors are held constant depends on the definition of the quantity of interest. In the Neyman-Rubin causal framework, the ceteris paribus condition
pertains exclusively to factors not affected by treatment. In other cases, especially with treatments such as race that are not the result of a well-defined manipulation or intervention, the researcher may prefer that the counterfactual comparison hold constant factors possibly affected by the attribute of interest (e.g., in case of race, educational attainment).

C We employ this phrase because, in our experience communicating these issues to others and especially non-methodologists, it best evokes the nature of the problem. It draws attention to the near perfect mapping to the methodological problem of confounding in analogous observational studies, and to the similar solutions of controlling for confounds and finding a natural experiment.

D Formally our model of Bayesian respondents implies:

\[
p(D = d_k, e = e_k | Z = z) = \frac{p(Z = z | D = d_k, e = e_k)p(D = d_k, e = e_k)}{p(Z = z)}
\]

\(p(Z = z | D = d_k, e = e_k)\) is the probability that the individual believes that a communicator would describe a scenario with \(D = d_k\) and \(e = e_k\) as \(Z = z\). For example, \(p(Z = \text{"democracy"} | D = \text{democracy}, e = \text{European})\) is the subjective probability that an individual assigns to the event that a communicator would call a country in a scenario a "democracy", given that the country is a democracy and is in Europe. Most individuals are thus likely to assign a high value to the preceding quantity, and a very low value to \(p(Z = \text{"democracy"} | D = \text{dictatorship}, e = \text{Middle Eastern})\).

E Psychologists have argued that Bayesian inference serves a good first approximation for how humans learn about causal relationships (Holyoak and Cheng, 2011; Perfors et al., 2011). Many legitimate criticisms have been raised about whether humans have realistic beliefs and do in fact revise according to conditional probability. Experiments have shown that some subjects ignore priors and make decisions based solely on the likelihood ratio, or that they give too much weight to priors (El-Gamal and Grether, 1995). However, there does not yet exist a model of human belief updating that, in our view, offers as good a first approximation as the Bayesian model. Any such alternative model can be empirically evaluated against the Bayesian model using the empirical strategy we use in this paper.

F Scholars might want to consider other “non-realistic” models in which respondents have non-realistic beliefs, perhaps reflecting the portrayal of the world by media or their prejudices, but still update in a Bayesian manner. For example, Gilliam and Iyengar (2000) found that respondents “fill in” beliefs about the race of a suspect in a new story: when no racial information was provided about the suspect of a violent crime, 44% of the time respondents recalled the suspect being black, and 19% of the time white (Table 2); this contrasts with the actual distribution of the race of perpetrators in Los Angeles television coverage which involves black individuals 29% of the time and white individuals 41% of the time (Table 1). Non-realistic Bayesian models will also yield precise predictions so long as we specify ex-ante the respondents’ beliefs about the world.

G Conjoint analysis typically manipulates all of the described attributes, but this is not necessary to control beliefs. The principle of the covariate control design is the same
whether the controls are held fixed for every respondent or manipulated. What matters for causal identification is that the vignette provides text that fixes the respondent’s beliefs about the characteristic \((e)\). However, these different approaches will change the causal estimand. Fixing a control to one level (say \(e = e_1\)) will estimate the LATE when \(e = e_1\), whereas manipulating the control (say sometimes \(e = e_1\) and sometimes \(e = e_2\)) will allow the researcher to estimate the LATEs when \(e = e_1\) and \(e = e_2\), or to average across them.

Our Embedded Natural Experiments depart from the ideal in one subtle way. The ideal embedded natural experiment would not provide any information about events subsequent to the natural experiment because this could lead to “post-treatment bias.” The vignette would end after the as-if random outcome of the assassination attempt. We opted to clarify what happened with the regime so as to prevent respondents from becoming confused, since the narrative otherwise feels unresolved. In our pilot surveys, we tested two alternative versions of the ENE design. The first alternative version refers to a similar narrative, but without the assassination attempt. This allowed us to investigate how much work the natural experiment, per se, was doing. The second alternative version refers to a similar narrative that ends abruptly with the assassination attempt. This second alternative circumvents the post-treatment bias problem we describe earlier, but has the disadvantage of a narrative that feels unresolved. The results for the three versions of the ENE design were similar. To minimize any bias that including post-treatment information could induce, we make the consequences of assassination on regime type as deterministic as possible by stating that “a well researched U.S. State Department report” concluded that without the president or the dictator, the country’s regime would become a military dictatorship or a democracy, respectively. The more deterministic the relationship between assassination and regime change, the less information about other features of the scenario is provided to a Bayesian respondent from reading that the probable outcome was realized.

The “ENE design” also has a covariate control component since it includes “your firm has been designated by Forbes magazine as one of the “100 best companies to work for.”’ Thus part of the reduction in imbalance could be due to this control. This was included to conform with the design of a previous wave of Latura’s study.


We included these questions related to inter-generational poverty because previous research shows that respondents’ support for welfare is driven by whether they think welfare recipients are stuck in cycles of poverty (Gamson and Lasch, 1983; Henry, Reyna and Weiner, 2004). Within this literature, the cycle of poverty framing could move respondents in one of two directions. It might make people think that welfare helps perpetuate the cycle of poverty or it might make them think that welfare is a ladder out of poverty.

In our thinking about this study we came up with several potential natural experiments for skin-pigment. For example, a person applying for welfare could be described as having a rare mutation making them slightly darker/lighter than their identical twin; we
would show pictures of both. But we realized that the results of such a study would not speak much to racial discrimination in America, because the context is so odd. This also highlights an advantage of ENEs: they focus the researchers mind on thinking about specific manipulations of the causal factor of interest, which is helpful for clarifying the counterfactual being estimated.

C Literature Review

C.1 Review of Articles in Top Journals

C.1.1 Classifying the Articles

We reviewed articles published in the *American Journal of Political Science*, the *American Political Science Review*, and *International Organization* between 2002 and 2015. Candidate articles were selected through the following search methodology: search for the terms “survey” and “experiment” in each journal issue within the aforementioned time period\(^\text{10}\); check the methodology section (including any appendices) of each article returned by this search to determine whether the article employed a survey experiment; if so, include the article. We used the following definition for survey experiments:

A survey experiment systematically varies one or more elements of a survey across subjects and assesses the effect of that variation on one or more measured outcomes. Typically, subjects (respondents) are randomly assigned to either a treatment group (of which there may be more than one) or a control group. The crucial, defining element of a survey experiment is that it manipulates some aspect of the survey protocol (Marsden and Wright, 2010, 838).

These could include vignette/scenario-based experiments, framing experiments, list experiments, or other types of survey experiments. The following attributes of selected articles were then recorded in a master spreadsheet: title, author(s), year of publication, journal, survey experiment type (scenario, framing, list, other), hypotheses tested, the experimental manipulation, and the sample size and subject pool (including survey service used).

We then examined all articles coded as employing scenario-based survey experiments (labeled as “Vignette”), and performed several additional coding tasks:

A We determined whether the authors argued that scenario-based survey experiments overcame the problem of confounding and allowed scholars to cleanly identify causal effects (of beliefs about a feature of the scenario), and recorded relevant quotes.

B We determined whether the authors explicitly acknowledged that survey experiments could have problems with internal validity, and collected relevant quotes.

---

\(^{10}\)We used JSTOR for articles published before 2014. For articles published after 2013, which are not available in JSTOR, we searched each issue of the three journals and supplemented our manual inspection with Google Scholar search.
C We determined whether the authors explicitly acknowledged that survey experiments could have problems of confounding as defined in our paper.

D We then recorded three new binary variables, based on the above:

(a) Confident of Internal Validity: If the author expressed confidence that survey experiments overcome confounding as described in part A, the article was coded 1. If not, 0.

(b) Limits to Internal Validity: If the author expressed any concerns about the internal validity of survey experiments as described in part B, the article was coded 1. If not, 0.

(c) Survey Confounding: If the author expressed specific concerns about the possibility of confounding in scenario-covariates\textsuperscript{11} as described in part C, the article was coded 1. If not, 0.\textsuperscript{12}

E Potentially Experience Confounding: We coded for whether survey experiments in the article might experience the problem of confounding as outlined in our paper. We considered all survey experiment types when applicable.

F For scenario-based experiments ("vignette"), we created categorical variables to classify the types of claims researchers made and what types of experiments they were conducting:

(a) Type of Causal Claim: The author could claim that her scenario-based experiment examined the effect of X or the effect of being described as X, both, or used the experiment as a measurement tool. For instance, if the author stated that her experiment looked at the effect of regime type on support for war, we coded it as “the effect of X.” If the author stated that her experiment studied the effect of a candidate being described using stereotypes on respondents’ support for the candidate, we coded it as “the effect of being described as X.” We created a separate category called “measurement” for when researchers used experimental vignettes to measure political attitudes.

(b) Type of Experiment: In scenario-based experiments, the experimenter could vary some characteristics of the scenario, presentation of information, or both. For instance, if the author varied the party of candidates or the regime type of an aggressor country, she manipulated characteristics of the scenario. If the author varied the language describing a person, country, idea, or object, she manipulated the presentation of information. One example of the latter manipulation would be the researcher describing a candidate as “a shady businessman” versus “a businessman who engages in unethical practices.”

\textsuperscript{11}The authors did not need to use the language of “confounding.” All that was required was that they acknowledged that respondents’ beliefs about the scenario may be affected by the manipulation in undesired ways.

\textsuperscript{12}Articles could be coded as 1 for (a) and 1 for (c), as some authors made competing statements about the validity of their survey experiments.
C.1.2 Articles Reviewed

We reviewed survey experiments published in the *American Journal of Political Science* (AJPS), the *American Political Science Review* (APSR), and *International Organization* (IO) between 2002 and 2015. The following table contains information about the 78 articles we reviewed.

<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Year</th>
<th>Journal</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender Stereotypes and Vote Choice</td>
<td>Kira Sanbonmatsu</td>
<td>2002</td>
<td>AJPS</td>
<td>Vignette</td>
</tr>
<tr>
<td>Stereotype Threat and Race of Interviewer</td>
<td>Darren W. Davis and Brian D. Silver</td>
<td>2003</td>
<td>AJPS</td>
<td>Other</td>
</tr>
<tr>
<td>Effects in a Survey on Political Knowledge</td>
<td>Christopher M. Federico</td>
<td>2004</td>
<td>AJPS</td>
<td>Vignette</td>
</tr>
<tr>
<td>When Do Welfare Attitudes Become Racialized? The Paradoxical Effects of Education</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Certainty or Accessibility: Attitude Strength in Candidate Evaluations</td>
<td>David A. M. Peterson</td>
<td>2004</td>
<td>AJPS</td>
<td>Vignette</td>
</tr>
<tr>
<td>Predisposing Factors and Situational Triggers: Exclusionary Reactions to Immigrant Minorities</td>
<td>Paul M. Sniderman, Louk Hagendoorn, and Markus Prior</td>
<td>2004</td>
<td>APSR</td>
<td>Vignette</td>
</tr>
<tr>
<td>Racial Resentment and White Opposition to Race-Conscious Programs: Principles or Prejudice?</td>
<td>Stanley Feldman and Leonie Huddy</td>
<td>2005</td>
<td>AJPS</td>
<td>Vignette</td>
</tr>
<tr>
<td>The Indirect Effects of Discredited Stereotypes in Judgments of Jewish Leaders</td>
<td>Adam J. Berinsky and Tali Mendelberg</td>
<td>2005</td>
<td>AJPS</td>
<td>Vignette</td>
</tr>
<tr>
<td>The “Race Card” Revisited: Assessing Racial Priming in Policy Contests</td>
<td>Gregory A. Huber and John S. Lapinski</td>
<td>2006</td>
<td>AJPS</td>
<td>Framing</td>
</tr>
<tr>
<td>Through a Glass and Darkly: Attitudes Towards International Trade and the Curious Effects of Issue Framing</td>
<td>Michael J. Hiscox</td>
<td>2006</td>
<td>IO</td>
<td>Framing</td>
</tr>
<tr>
<td>Beyond Negativity: The Effects of Incivility on the Electorate</td>
<td>Deborah Jordan Brooks and John G. Geer</td>
<td>2007</td>
<td>AJPS</td>
<td>Framing</td>
</tr>
<tr>
<td>Issue Definition, Information Processing, and the Politics of Global Warming</td>
<td>B. Dan Wood and Arnold Vedlitz</td>
<td>2007</td>
<td>AJPS</td>
<td>Framing</td>
</tr>
<tr>
<td>Designing and Analyzing Randomized Experiments: Application to a Japanese Election Survey Experiment</td>
<td>Yusaku Horiuchi, Kosuke Imai and Naoko Taniguchi</td>
<td>2007</td>
<td>AJPS</td>
<td>Other</td>
</tr>
<tr>
<td>When Race Matters and When It Doesn’t: Racial Group Differences in Response to Racial Cues</td>
<td>Ismail K. White</td>
<td>2007</td>
<td>APSR</td>
<td>Framing</td>
</tr>
<tr>
<td>Domestic Audience Costs in International Relations: An Experimental Approach Opinion Taking within Friendship Networks</td>
<td>Michael Tomz</td>
<td>2007</td>
<td>IO</td>
<td>Vignette</td>
</tr>
<tr>
<td>Attributing Blame: The Public’s Response to Hurricane Katrina</td>
<td>Suzanne L. Parker, Glenn R. Parker and James A. McCann Neil Malhotra and Alexander G. Kuo</td>
<td>2008</td>
<td>AJPS</td>
<td>Framing</td>
</tr>
<tr>
<td>Title</td>
<td>Authors</td>
<td>Year</td>
<td>Journal</td>
<td>Type</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>----------------------------------------------------------------</td>
<td>------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td>What Triggers Public Opposition to Immigration? Anxiety, Group Cues,</td>
<td>Ted Brader, Nicholas A. Valentino and Elizabeth Suhay</td>
<td>2008</td>
<td>AJPS</td>
<td>Vignette /</td>
</tr>
<tr>
<td>and Immigration Threat</td>
<td>Dennis Chong and James N. Druckman</td>
<td></td>
<td>APSR</td>
<td>Framing</td>
</tr>
<tr>
<td>Framing Public Opinion in Competitive Democracies</td>
<td>James L. Gibson</td>
<td>2008</td>
<td>APSR</td>
<td>Vignette</td>
</tr>
<tr>
<td>Challenges to the Impartiality of State Supreme Courts: Legitimacy</td>
<td>Michael Tomz and Robert P. Van Houweling</td>
<td>2008</td>
<td>APSR</td>
<td>Vignette</td>
</tr>
<tr>
<td>Theory and “New-Style” Judicial Campaigns</td>
<td>Scott Sigmund Gartner</td>
<td>2008</td>
<td>APSR</td>
<td>Vignette</td>
</tr>
<tr>
<td>Candidate Positioning and Voter Choice</td>
<td>Jennifer Jerit</td>
<td>2009</td>
<td>AJPS</td>
<td>Framing</td>
</tr>
<tr>
<td>The Multiple Effects of Casualties on Public Support for War: An</td>
<td>Paul Goren, Christopher M. Federico and Miki Caul Kittilson</td>
<td>2009</td>
<td>AJPS</td>
<td>Framing</td>
</tr>
<tr>
<td>Experimental Approach</td>
<td>Michael Tomz and Robert P. Van Houweling</td>
<td>2009</td>
<td>APSR</td>
<td>Vignette</td>
</tr>
<tr>
<td>How Predictive Appeals Affect Policy Opinions</td>
<td>Dennis Chong and James N. Druckman</td>
<td>2010</td>
<td>APSR</td>
<td>Other</td>
</tr>
<tr>
<td>Source Cues, Partisan Identities, and Political Value Expression</td>
<td>Jens Hainmueller and Michael J. Hiscox</td>
<td>2010</td>
<td>APSR</td>
<td>Vignette</td>
</tr>
<tr>
<td>The Electoral Implications of Candidate Ambiguity</td>
<td>Laurel Harbridge and Neil Malhotra</td>
<td>2011</td>
<td>AJPS</td>
<td>Framing</td>
</tr>
<tr>
<td>Dynamic Public Opinion: Communication Effects Over Time</td>
<td>Robert F. Trager and Lynn Vavreck</td>
<td>2011</td>
<td>AJPS</td>
<td>Vignette</td>
</tr>
<tr>
<td>Attitudes toward Highly Skilled and Low-skilled Immigration: Evidence</td>
<td>Michael Tomz and Robert P. Van Houweling</td>
<td>2010</td>
<td>APSR</td>
<td>Vignette</td>
</tr>
<tr>
<td>from a Survey Experiment</td>
<td>Dennis Chong and James N. Druckman</td>
<td></td>
<td>APSR</td>
<td>Other</td>
</tr>
<tr>
<td>Emotional Substrates of White Racial Attitudes</td>
<td>Megumi Naoi and Ikuo Kume</td>
<td>2011</td>
<td>IO</td>
<td>Other</td>
</tr>
<tr>
<td>Cognitive Biases and the Strength of Political Arguments</td>
<td>Antoine J. Banks and Nicholas A. Valentino Kevin Arceneaux</td>
<td>2012</td>
<td>AJPS</td>
<td>Framing</td>
</tr>
<tr>
<td>Polarizing Cues</td>
<td>Stephen P. Nicholson and Nikolay Marinov</td>
<td>2012</td>
<td>AJPS</td>
<td>Framing</td>
</tr>
<tr>
<td>Taking Sides in Other People’s Elections: The Polarizing Effect of</td>
<td>Daniel Corstange and Michael Bang Petersersen</td>
<td></td>
<td>AJPS</td>
<td>Vignette</td>
</tr>
<tr>
<td>Foreign Intervention</td>
<td></td>
<td></td>
<td>APSR</td>
<td>Framing</td>
</tr>
<tr>
<td>Social Welfare as Small-Scale Help: Evolutionary Psychology and the</td>
<td></td>
<td></td>
<td>APSR</td>
<td>Other</td>
</tr>
<tr>
<td>Deservingness Heuristic</td>
<td>Justin Grimmer, Solomon Messing, and Sean J. Westwood</td>
<td>2012</td>
<td>Other</td>
<td></td>
</tr>
<tr>
<td>How Words and Money Cultivate a Personal Vote: The Effect of</td>
<td>James N. Druckman, Jordan Fein, and Thomas J. Leeper</td>
<td>2012</td>
<td>APSR</td>
<td>Framing</td>
</tr>
<tr>
<td>Legislator Credit Claiming on Constituent Credit Allocation A</td>
<td></td>
<td></td>
<td>APSR</td>
<td>Framing</td>
</tr>
<tr>
<td>Source of Bias in Public Opinion Stability</td>
<td></td>
<td></td>
<td>APSR</td>
<td>Other</td>
</tr>
<tr>
<td>Politics in the Mind’s Eye: Imagination as a Link between Social and</td>
<td>Michael Bang Petersen and Lene Aarøe</td>
<td>2012</td>
<td>APSR</td>
<td>Framing</td>
</tr>
<tr>
<td>Political Cognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

40
<table>
<thead>
<tr>
<th>Title</th>
<th>Author(s)</th>
<th>Year</th>
<th>Journal</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economic Explanations for Opposition to Immigration: Distinguishing between Prevalence and Conditional Impact</td>
<td>Neil Malhotra, Yotam Margalit and Cecilia Hyunjung Mo Christopher D. DeSante</td>
<td>2013</td>
<td>AJPS</td>
<td>Vignette</td>
</tr>
<tr>
<td>Working Twice as Hard to Get Half as Far: Race, Work Ethic, and America’s Deserving Poor</td>
<td>James N. Druckman, Erik Peterson, and Rune Slothuus</td>
<td>2013</td>
<td>APSR</td>
<td>Framing</td>
</tr>
<tr>
<td>How Elite Partisan Polarization Affects Public Opinion Formation</td>
<td>Michael R. Tomz and Jessica L. P. Weeks</td>
<td>2013</td>
<td>APSR</td>
<td>Vignette</td>
</tr>
<tr>
<td>Atomic Aversion: Experimental Evidence on Taboos, Traditions, and the Non-use of Nuclear Weapons</td>
<td>Jason Lyall, Graeme Blair, and Kosuke Imai</td>
<td>2013</td>
<td>APSR</td>
<td>Other</td>
</tr>
<tr>
<td>Explaining Social Policy Preferences: Evidence from the Great Recession</td>
<td>Thad Dunning and Janhavi Nilekani</td>
<td>2013</td>
<td>APSR</td>
<td>Vignette</td>
</tr>
<tr>
<td>Explaining Support for Combatants During Wartime: A Survey Experiment in Afghanistan</td>
<td>Martin Ardanaz, M. Victoria Murillo, and Pablo M. Pinto Geoffrey P.R. Wallace</td>
<td>2013</td>
<td>IO</td>
<td>Framing</td>
</tr>
<tr>
<td>Sensitivity to Issue Framing on Trade Policy Preferences: Evidence from a Survey Experiment</td>
<td>Andrew Healy and Gabriel S. Lenz</td>
<td>2014</td>
<td>APSR</td>
<td>Vignette</td>
</tr>
<tr>
<td>International Law and Public Attitudes Toward Torture: An Experimental Study</td>
<td>David E. Broockman</td>
<td>2014</td>
<td>APSR</td>
<td>Other</td>
</tr>
<tr>
<td>Partisans in Robes: Party Cues and Public Acceptance of Supreme Court Decisions</td>
<td>Adam J. Berinsky, Michele F. Margolis, and Michael W. Sances Samara Klar</td>
<td>2014</td>
<td>APSR</td>
<td>Vignette</td>
</tr>
<tr>
<td>The Power of Partisanship in Brazil: Evidence from Survey Experiments</td>
<td>Margaret E. Roberts et al.</td>
<td>2014</td>
<td>APSR</td>
<td>Framing</td>
</tr>
<tr>
<td>The Conditionality of Vote Buying Norms: Experimental Evidence from Latin America</td>
<td>Graeme Blair, Kosuke Imai, and Jason Lyall</td>
<td>2014</td>
<td>APSR</td>
<td>Other</td>
</tr>
</tbody>
</table>

41
Preferences for International Redistribution: The Divide over the Eurozone Bailouts
Michael M. Bechtel, Jens Hainmueller, and Yotam Margalit
2014 AJPS Vignette

The Political Mobilization of Ethnic and Religious Identities in Africa
John F. McCauley
2014 APSR Framing

False Commitments: Local Misrepresentation and the International Norms Against Female Genital Mutilation and Early Marriage
Karisa Cloward
2014 IO Other

Promises or Policies? An Experimental Analysis of International Agreements and Audience Reactions
Stephen Chaudoin
2014 IO Vignette

Decision Maker Preferences for International Legal Cooperation
Emilie M. Hafner-Burton, Brad L. LeVeck, David G. Victor and James H. Fowler
2014 IO Vignette

Attacks without Consequence? Candidates, Parties, Groups and the Changing Face of Negative Advertising
Conor M. Dowling and Amber Wichtowsky
2015 AJPS Framing

Monopoly Money: Foreign Investment and Bribery in Vietnam, a Survey Experiment
Edmund J. Malesky, Dimitar D. Gueorguiev, and Nathan M. Jensen
2015 AJPS Other

Chief Justice Roberts’s Health Care Decision Disrobed: The Microfoundations of the Supreme Court’s Legitimacy
Dino P. Christenson and David M. Glick
2015 AJPS Framing

Responsibility Attribution for Collective Decision Makers
Raymond Duch, Wojtek Przepiorka, and Randolph Stevenson
2015 AJPS Vignette

Explaining Explanations: How Legislators Explain their Policy Decisions and How Citizens React
Christian R. Grose, Neil Malhotra, and Robert Parks Van Houweling
2015 AJPS Framing

Fear and Loathing Across Party Lines: New Evidence on Group Polarization
Shanto Iyengar and Sean J. Westwood
2015 AJPS Vignette

Xenophobic Rhetoric and its Political Effects on Immigrants and their Co-Ethnic Efrén O. Pérez
2015 AJPS Framing

The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes Towards Immigrants
Jens Hainmueller and Daniel J. Hopkins
2015 AJPS Vignette

Decomposing Audience Costs: Bringing the Audience Back Into Audience Cost Theory
Joshua D. Kertzer and Ryan Brutger
2015 AJPS Vignette

Expressive Partisanship: Campaign Involvement, Political Emotion, and Partisan Identity
Leonie Huddy, Lilliana Mason, and Lene Aarøe
2015 APSR Framing

Human Rights Organizations as Agents of Change: An Experimental Examination of Framing and Micromobilization
Kayla Jo McEntire, Michele Leiby, and Matthew Krain
2015 APSR Framing

Race, Paternalism, and Foreign Aid: Evidence from U.S. Public Opinion
Andy Baker
2015 APSR Vignette

Religious Social Identity, Religious Belief, and Anti-Immigration Sentiment
Pazit Ben-Nun Bloom, Gizem Arikan, and Marie Courtemanche
2015 APSR Framing

International Knowledge and Domestic Evaluations in a Changing Society: The Case of China
Haifeng Huang
2015 APSR Other
C.1.3 Summary Statistics

We present some summary statistics of the articles we reviewed. Because we do not want to critique individual researchers or teams of researchers, we provide aggregate level data. The data from our literature review suggest that most researchers who used scenario-based survey experiments were not concerned about limits to internal validity. Furthermore, only five out of 35 mentioned the possibility of confounding in their survey experiments.

Table 2: Summary Statistics from Literature Review

<table>
<thead>
<tr>
<th>Type of Experiment</th>
<th>Framing</th>
<th>Vignette</th>
<th>Framing/Vignette</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Articles</td>
<td>31</td>
<td>32</td>
<td>2</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Not Applicable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potentially Experience Confounding</td>
<td>16</td>
<td>60</td>
<td>2</td>
</tr>
<tr>
<td>Express Confidence of Internal Validity</td>
<td>16</td>
<td>19</td>
<td>43</td>
</tr>
<tr>
<td>Express Limits to Internal Validity</td>
<td>10</td>
<td>25</td>
<td>43</td>
</tr>
<tr>
<td>Recognize Possibility of Confounding</td>
<td>5</td>
<td>30</td>
<td>43</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of Causal Claim</th>
<th>Number of Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect of X</td>
<td>19</td>
</tr>
<tr>
<td>Effect of Being Described as X</td>
<td>1</td>
</tr>
<tr>
<td>Measurement</td>
<td>3</td>
</tr>
<tr>
<td>Effect of X / Effect of Being Described as X</td>
<td>5</td>
</tr>
<tr>
<td>Effect of X / Measurement</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of Experiment</th>
<th>Number of Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristics of Scenarios</td>
<td>21</td>
</tr>
<tr>
<td>Presentation of Information</td>
<td>2</td>
</tr>
<tr>
<td>Both</td>
<td>10</td>
</tr>
</tbody>
</table>

D “Democratic Peace” Survey Experiment Details

D.1 Outline of the Survey

First, we outline the structure of the survey. Next, we describe each section of the survey in detail.

All questions in the survey are contained in sections. The order of the section is as follows:

- IRB Consent Form
- Instructions
- Experimental Vignette
- Survey Questions (contains five blocks)
• Attention Check  
• Demographic Variables  
• Debrief  

We experimentally vary the order of the five blocks in the Survey Questions section:  
A Placebo Test: Open-ended response  
B Placebo Tests: Multiple choice  
C Treatment Measure  
D Plausibility Check  
E Support for Using Force, Mediation Questions  

Each respondent had an equal probability of being assigned to each of the 120 ordering permutations possible. Any boldface or capitalization in the text below appeared in the survey. We employed Bernoulli randomization in all of our randomization procedures.  

D.2 Three Vignette Types  
Each subject had 1/3 probability of being randomly assigned to one of three vignette types. Within each vignette type, each subject had an equal chance of being assigned to one of the two experimental conditions. In the treatment condition, respondents were told the country in the scenario is a democracy. In the control condition, respondents were told the country is a non-democracy. The texts of the vignettes appear below: 

D.2.1 Basic  
A country is developing nuclear weapons and will have its first nuclear bomb within six months. The country could then use its missiles to launch nuclear attacks against any country in the world.  

[The country is not a democracy and shows no sign of becoming a democracy./The country is a democracy and shows every sign that it will remain a democracy.]  
The country’s motives remain unclear, but if it builds nuclear weapons, it will have the power to blackmail or destroy other countries.  
The country has refused all requests to stop its nuclear weapons program.  

D.2.2 Covariate Control  
A country is developing nuclear weapons and will have its first nuclear bomb within six months. The country could then use its missiles to launch nuclear attacks against any country in the world.  

• The country is not a democracy and shows no sign of becoming a democracy./is a democracy and shows every sign that it will remain a democracy.]
• The country [has not/has] signed a military alliance with the U.S.

• The country has [low/high] levels of trade with the U.S.

• The country’s nonnuclear military forces are half as strong as the U.S.’s nonnuclear forces.

The country’s motives remain unclear, but if it builds nuclear weapons, it will have the power to blackmail or destroy other countries.

The country has refused all requests to stop its nuclear weapons program.

D.2.3 Embedded Natural Experiment

Embedded Natural Experiment Fragile Democracy (ENEd)

Five years ago a country, Country A, was a fragile democracy. It had a democratically elected government, headed by a popular president. At the time, a well-researched U.S. State Department report concluded that without this president, there was a very high probability that the country’s military would overthrow the government to set up a dictatorship.

Two years ago at a public event, a disgruntled military officer shot at the president of Country A. [The president was hit in the head and did not survive the attack. In the political vacuum that followed the president’s death, the country’s military overthrew the democratically elected government. Today, Country A is a military dictatorship./The president was hit in the shoulder and survived the attack. The country’s democratically elected government survived the political turmoil. Today, Country A is still a democracy.]

• Currently, Country A is developing nuclear weapons and will have its first nuclear bomb within six months. Country A could then use its missiles to launch nuclear attacks against any country in the world.

• Country A’s motives remain unclear, but if it builds nuclear weapons, it will have the power to blackmail or destroy other countries.

• Country A has refused all requests to stop its nuclear weapons program.

Embedded Natural Experiment Fragile Non-democracy (ENEn)

Five years ago a country, Country A, was a dictatorship. At the time, a well-researched U.S. State Department report concluded that if the dictator were to die, the country had a very high likelihood of becoming a democracy.

Two years ago at a public event, a pro-democracy rebel shot at the dictator of Country A. [The dictator was hit in the head and did not survive the attack. In the political vacuum that followed, pro-democracy protestors took to the streets and forced those in the former dictator’s government to resign. Soon after Country A held national elections and it is still a democracy today./The dictator was hit in the shoulder and survived the attack. The dictator’s regime survived the political turmoil. Today, Country A is still a dictatorship.]
• Currently, Country A is developing nuclear weapons and will have its first nuclear bomb within six months. Country A could then use its missiles to launch nuclear attacks against any country in the world.

• Country A’s motives remain unclear, but if it builds nuclear weapons, it will have the power to blackmail or destroy other countries.

• Country A has refused all requests to stop its nuclear weapons program.

D.3 Support for Force and Mediation Questions

The substantive outcome of interest in Tomz and Weeks (2013) was support of using force against the aggressor country. Furthermore, the researchers used mediation questions to understand why regime type of the aggressor country affects respondents' force for using force. We deployed the same set of substantive questions that Tomz and Weeks used as a block of questions; we randomized the order the questions appeared to respondents within that block.

A Support for Using Military Force

B Mechanisms 1: consequences if military action is taken

C Mechanisms 2: consequences if military action is not taken

D Mechanism 3: the morality of military action

D.4 Together-Placebos Design and Separated-Placebos Design

One concern researchers might have with placebo test questions is whether they affect the substantive outcome of interest and vice versa. To address this concern, we designed our survey in the following way. Each subject had 1/2 probability of being randomly assigned to one survey flow of the two types: the Together-Placebos design or the Separated-Placebos design. In the Together-Placebos design, all the multi-choice placebo test questions were presented together in one block. In the Separated-Placebos design, some multiple-choice placebo tests questions were presented before the support for force question and others are presented after that question. We used the Separated-Placebos design to test whether individual placebo test questions affect responses to the support for force question. In the Together-Placebos design, we were only able to test if the placebo tests, in aggregate, affect respondents’ support for using force. It is possible that the placebo tests have cross-cutting effects that cancel each other out. As a result, we used the Separated-Placebos design: we presented all the placebo test questions in the survey, but we inserted the support for force question before the first, second, third, or fourth placebo test questions.

The placebo tests that we isolated with the Separated-Placebos Design were Placebo Tests C (Regions), Placebo D (GDP per Capita), and Placebo E (Religion). The eight possible combinations of the ordering were as follows:
D.5 Survey Questions

The survey questions consisted of the placebo test questions, the treatment measures, the support for force and mediation questions, the attention check, and the demographics questions.

D.5.1 Justifications for Placebo Test Questions

We selected placebo test variables by identifying real-world variables that show large and significant imbalances across regime types (see Table 3). For our analysis, we used data from the Quality of Government (GOG) Basic dataset, the Correlates of War (COW) formal alliance dataset, the COW trade dataset, the COW National Material Capabilities dataset, the CIA World Factbook Ethnic Group dataset, Vito D’Orazio’s Joint Military Exercises dataset.

In our previous waves we selected placebo variables informally based on our intuitions. However, following the helpful comments of Cyrus Samii on this point, for this wave we opted to select our placebos more formally by identifying real-world variables that show large and significant imbalances across regime types. This new more formal placebo selection process led us to remove placebo test questions regarding whether the country was English-speaking (insufficient imbalance) and whether the country had fought alongside the U.S. in the Iraq War, which we feared was too idiosyncratic. It also led us to include placebo test questions regarding the country’s oil reserves, racial makeup, and joint military exercise with the U.S. which were sufficiently imbalanced; oil reserves and racial makeup are unlikely to be affected by regime-type; joint military exercise is included as characteristic related but not identical to military alliance.

First, we showed that geographic regions should be included as a placebo question because democracies and non-democracies are distributed differently across regions. Figure 11 displays the percent of countries that are democracies in the ten regions of the world. We defined democracy using the variable $chga\_demo$ from QOG, which is a binary coding of democracy/non-democracy from the Cheibub et al. 2010 dataset.\textsuperscript{21}

Figure 8: Democracies in Regions of the World

In the analysis of our survey experiment, we focused on four regions that exhibit the most imbalance between regime types: Western Europe, North America, Sub-Saharan Africa, and North Africa & the Middle East. The first two have the largest percentage of countries that are democracies and the last two have the smallest percentage of countries that are democracies.\textsuperscript{22}

To select the rest of the placebo test variables, we analyzed 114 characteristics of countries in 1998 from all the datasets mentioned in the introduction. We tried to identify variables that are the most imbalanced across regime types.\textsuperscript{23} We selected data from 1998 so that our potential placebo variables are lagged behind the democracy variable by 10 years (the most

\textsuperscript{19}http://vitodorazio.weebly.com/data.html


\textsuperscript{22}We also include East Asia and Central Asia among our answer choices in the placebo test question because those were popular answers in our pilot studies.

\textsuperscript{23}The CIA World Factbook Ethnic Group dataset contains too many ethnic groups. Instead, we code the variable $majority\_white$ using the dataset. For each country, $majority\_white$ is coded 1 if the country’s population is greater than 50 percent white (Causasian) and 0 otherwise. Note the data is from 2000 and not 1998; however, we think whether a country was majority white is unlikely to have changed between 1998 and 2000.
For each potential placebo variable $P_k$ for $k \in \{1, 2, ..., 114\}$, we standardized them to create $S_{i,k}$ such that for country $i$:

$$S_{i,k} = \frac{P_{i,k}}{\text{Var}(P_k)}$$  \hspace{1cm} (1)

For each country, let $D_i = 1$ if country $i$ is a democracy in 2008 according to chga_demo and 0 otherwise. We estimated $E(S_{k,i}|D_i = 1) - E(S_{k,i}|D_i = 0)$ using $\hat{\gamma}_k$ from the following regression:

$$E(S_{k,i}|D_i) = \eta_k + \gamma_k D_i$$  \hspace{1cm} (2)

We can interpret $\hat{\gamma}_k$ as the estimated difference in means for standardized variable $S_{i,k}$ between democracies and non-democracies. Table 3 presents the coefficient estimates and robust standard errors for the 25 variables that exhibit the greatest imbalance (in absolute value) across regime types.\footnote{We also report the percentage of countries that are missing from each of the variables in the datasets.}

First, we constructed a placebo variable measuring how likely it is that the country in the scenario had large oil reserves. High fuel exports were highly correlated with being a non-democracy while high net energy imports were highly correlated with being a democracy. However, rather than ask about fuel exports/imports, our placebo question asked about oil reserves because it is relatively more exogenous to regime type.

Second, we created a placebo variable measuring how likely it is that the country in the scenario was majority Christian. As Table 3 shows, democracies had a low percentage of Muslims and a high percentage of Catholics in 1980. Since religion is slow-changing over time, we regarded it as an especially valid placebo variable (it is unlikely to be affected by regime-type on a short time scale).

Third, we created a placebo variable measuring GDP per capita. Many of the highly imbalanced variables in Table 3 are related to levels of economic development. These variables include employment in agriculture as a percentage of total employment, employment in services as a percentage of total employment, gross enrollment ratio in pre-primary schools, health expenditure as percent of GDP, and mortality rate of children under five. In selecting a placebo question we had several considerations to balance: we wanted to only ask one question to avoid burdening the respondent with multiple redundant questions; we wanted to choose a question that captures much of the common variance to these characteristics; we wanted to ask about a factor that is most likely to influence the outcome (support for using force); we wanted to ask a question that is easy to understand. These considerations led us
to ask about GDP per capita. GDP per capita, itself, is 0.4 standard deviations greater for
democracies than non-democracies in 1998 ($p < 0.001$).

Finally, we asked about the racial makeup of the country’s population. As Table 3
shows, democracies were more likely to be majority white compared with non-democracies
($p < 0.001$).

Table 3: Top 25 Variables Most Imbalanced Across Regime Types

<table>
<thead>
<tr>
<th>Variables (Standardized)</th>
<th>Coef</th>
<th>SE</th>
<th>% Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fuel exports (% of merchandise exports)</td>
<td>$-0.959$</td>
<td>0.239</td>
<td>37</td>
</tr>
<tr>
<td>Muslims as percentage of population in 1980</td>
<td>$-0.953$</td>
<td>0.152</td>
<td>11</td>
</tr>
<tr>
<td>Employment in agriculture (% of total employment)</td>
<td>$-0.941$</td>
<td>0.335</td>
<td>56</td>
</tr>
<tr>
<td>Population ages 65 and above (% of total)</td>
<td>0.922</td>
<td>0.121</td>
<td>11</td>
</tr>
<tr>
<td>Heritage Foundation Economic Freedom Index: property rights</td>
<td>0.912</td>
<td>0.147</td>
<td>21</td>
</tr>
<tr>
<td>Heritage Foundation Economic Freedom Index</td>
<td>0.877</td>
<td>0.158</td>
<td>21</td>
</tr>
<tr>
<td>Employment in services (% of total employment)</td>
<td>0.859</td>
<td>0.295</td>
<td>56</td>
</tr>
<tr>
<td>Number of military treaties</td>
<td>0.822</td>
<td>0.109</td>
<td>0</td>
</tr>
<tr>
<td>Number of treaties: defense</td>
<td>0.810</td>
<td>0.109</td>
<td>0</td>
</tr>
<tr>
<td>Gross enrollment ratio, pre-primary schools, total</td>
<td>0.807</td>
<td>0.182</td>
<td>43</td>
</tr>
<tr>
<td>Number of treaties: entente</td>
<td>0.784</td>
<td>0.110</td>
<td>0</td>
</tr>
<tr>
<td>Population ages 0-14 (% of total)</td>
<td>$-0.774$</td>
<td>0.135</td>
<td>11</td>
</tr>
<tr>
<td>Number of treaties: non-aggression</td>
<td>0.758</td>
<td>0.111</td>
<td>0</td>
</tr>
<tr>
<td>Catholics as percentage of population in 1980</td>
<td>0.740</td>
<td>0.129</td>
<td>11</td>
</tr>
<tr>
<td>Social Globalization Index</td>
<td>0.732</td>
<td>0.142</td>
<td>11</td>
</tr>
<tr>
<td>Heritage Foundation Economic Freedom Index: trade freedom</td>
<td>0.723</td>
<td>0.154</td>
<td>21</td>
</tr>
<tr>
<td>Alternative and nuclear energy (% of total energy use)</td>
<td>0.698</td>
<td>0.149</td>
<td>35</td>
</tr>
<tr>
<td>Energy imports, net (% of energy use)</td>
<td>0.686</td>
<td>0.199</td>
<td>35</td>
</tr>
<tr>
<td>Country’s population was majority white</td>
<td>0.675</td>
<td>0.120</td>
<td>0</td>
</tr>
<tr>
<td>Services, etc., value added (% of GDP)</td>
<td>0.675</td>
<td>0.147</td>
<td>16</td>
</tr>
<tr>
<td>Health expenditure, total (% of GDP)</td>
<td>0.666</td>
<td>0.135</td>
<td>10</td>
</tr>
<tr>
<td>Employment in industry (% of total employment)</td>
<td>0.655</td>
<td>0.336</td>
<td>56</td>
</tr>
<tr>
<td>Mortality rate, under-5 (per 1,000 live births)</td>
<td>$-0.654$</td>
<td>0.147</td>
<td>8</td>
</tr>
<tr>
<td>Average value of ethnolinguistic fractionalization</td>
<td>$-0.650$</td>
<td>0.199</td>
<td>47</td>
</tr>
<tr>
<td>Armed forces personnel (% of total labor force)</td>
<td>$-0.624$</td>
<td>0.168</td>
<td>17</td>
</tr>
</tbody>
</table>

We also examined variables that are related to military capability, alliance, and trade,
three variables that were included as details in the Tomz and Weeks’s survey experiment
design. Potential placebo variables include those that were explicitly controlled for by the
Tomz and Weeks’s vignettes (i.e., non-nuclear military capability, military treaties, and
volume of import/export) and those that are highly correlated with alliance and trade (i.e.,
iron and steel production, energy consumption, population, joint military exercises, and
foreign direct investment). We estimated $\gamma_k$ for these variables using the same model as
described in the previous section. Table 4 contains our coefficient estimates and robust
standard errors.\textsuperscript{26} We found that none of the variables that describe military capability

\textsuperscript{26}Again, we reported the percentage of countries that are missing in each variable.
is statistically significant at $\alpha = 0.05$. On the other hand, variables related to trade and military alliance were all statistically different between regime types at $\alpha = 0.05$.

Table 4: Variables Related to Military Capability, Alliance, and Trade with the U.S.

<table>
<thead>
<tr>
<th>Variables (Standardized)</th>
<th>Coef</th>
<th>SE</th>
<th>% Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Military Capability Variables</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iron and steel production (thousands of tons)</td>
<td>0.137</td>
<td>0.159</td>
<td>8</td>
</tr>
<tr>
<td>Military expenditures (thousands of $)</td>
<td>0.105</td>
<td>0.155</td>
<td>8</td>
</tr>
<tr>
<td>Military personnel (thousands)</td>
<td>−0.172</td>
<td>0.173</td>
<td>8</td>
</tr>
<tr>
<td>Energy consumption (thousands of coal-ton equivalents)</td>
<td>0.108</td>
<td>0.155</td>
<td>8</td>
</tr>
<tr>
<td>Total population (thousands)</td>
<td>−0.057</td>
<td>0.166</td>
<td>8</td>
</tr>
<tr>
<td>Urban population (thousands)</td>
<td>−0.089</td>
<td>0.181</td>
<td>8</td>
</tr>
<tr>
<td>Composite Index of National Capability Score</td>
<td>−0.010</td>
<td>0.172</td>
<td>8</td>
</tr>
<tr>
<td><strong>Alliance Variables</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of treaties: defense</td>
<td>0.810</td>
<td>0.109</td>
<td>0</td>
</tr>
<tr>
<td>Number of treaties: non-aggression</td>
<td>0.758</td>
<td>0.111</td>
<td>0</td>
</tr>
<tr>
<td>Number of treaties: entente</td>
<td>0.784</td>
<td>0.110</td>
<td>0</td>
</tr>
<tr>
<td>Number of military treaties (all types)</td>
<td>0.822</td>
<td>0.109</td>
<td>0</td>
</tr>
<tr>
<td>Number of joint military exercises</td>
<td>0.398</td>
<td>0.119</td>
<td>0</td>
</tr>
<tr>
<td><strong>Trade Variables</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Volume of imports</td>
<td>0.259</td>
<td>0.129</td>
<td>8</td>
</tr>
<tr>
<td>Volume of exports</td>
<td>0.323</td>
<td>0.122</td>
<td>8</td>
</tr>
<tr>
<td>Total volume of trade (imports + exports)</td>
<td>0.291</td>
<td>0.125</td>
<td>8</td>
</tr>
<tr>
<td>FDI: position on a historical-cost basis</td>
<td>0.325</td>
<td>0.138</td>
<td>46</td>
</tr>
<tr>
<td>FDI: net financial transactions</td>
<td>0.403</td>
<td>0.139</td>
<td>44</td>
</tr>
<tr>
<td>FDI: net income</td>
<td>0.386</td>
<td>0.128</td>
<td>41</td>
</tr>
</tbody>
</table>

Based on our analysis, we asked placebo test questions regarding geographic region, GDP per capita, religion, oil reserves, race, military spending, military alliance, trade, joint military exercise, and foreign direct investment.

D.6 Correlation Between Democracy and Percent Muslim

For each region of the world, we calculated the correlation (Pearson’s $r$) between a country being a democracy in 2008 and the percent of its population who were Muslims in 1980. The data came from the Quality of Government Dataset. We used `chga_demo` for our binary measure of democracy and `lp_muslim80` as our measure of the percentage Muslim in each country. Note that we could not calculate the correlation for North America because all countries in North America are democracies.
Table 5: Correlation Between Democracy and Percent of Population that is Muslim by Region

<table>
<thead>
<tr>
<th>Region</th>
<th>Pearson’s r</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Countries</td>
<td>-0.469</td>
</tr>
<tr>
<td>Eastern Europe and post Soviet Union</td>
<td>-0.681</td>
</tr>
<tr>
<td>Latin America</td>
<td>0.087</td>
</tr>
<tr>
<td>North Africa &amp; the Middle East</td>
<td>-0.656</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>0.048</td>
</tr>
<tr>
<td>Western Europe and North America</td>
<td>NA</td>
</tr>
<tr>
<td>East Asia</td>
<td>-0.379</td>
</tr>
<tr>
<td>South-East Asia</td>
<td>-0.054</td>
</tr>
<tr>
<td>South Asia</td>
<td>-0.537</td>
</tr>
<tr>
<td>The Pacific</td>
<td>-0.488</td>
</tr>
<tr>
<td>The Caribbean</td>
<td>-0.461</td>
</tr>
</tbody>
</table>

D.7 Placebo Test Questions

D.7.1 Notes on Placebo Test Questions

For Questions D through L (the multiple-choice questions), we provided subjects with the following instructions:

The following nine questions will ask you about what you think the country described in the scenario was like in the past (specifically, 10 years ago). Please tell us your best guess of what the country was like in the past.

Note that the instructions asked about country in the past. This was because we wanted to minimize the risk that subjects would think about characteristics that could be caused by a recent change in the regime type of the country, which would make these questions less valid placebos.

Before each question in D through K, we also added the following sentence:

Tell us your best guess of what the country was like 10 years ago.

For the multiple choice questions, we randomized whether the answer choices were presented in ascending (smallest value to largest value) or descending order (largest value to smallest value). Each respondent had 1/2 probability of seeing the answer choices for all questions in ascending order and 1/2 probability of seeing the answer choices for all questions in descending order.

D.7.2 Text of Placebo Test Questions

**A** Please list some countries, from the real-world, that you think are most likely to fit the scenario.

[Textbox]
B Think about the scenario you read. Write down what you think the country in the scenario is like. Write down at least five things that come to your mind.

[Textbox]

C What region of the world do you think the country is in? What regions of the world do you think the country is not in?

Please drag your two best guesses of which region the country is in to the top box. Please drag your two best guesses of which regions the country is not in to the bottom box.

D How wealthy do you think the country was in terms of GDP per capita? (GDP per capita is often considered an indicator of a country’s standard of living.)

We provide you with two example countries in each category.

- Less than $500 (Ex: Democratic Republic of the Congo, El Salvador)
- $501-$1,000 (Ex: Rwanda, Haiti)
- $1,001-$5,000 (Ex: India, Cuba)
- $5,001-$10,000 (Ex: Brazil, China)
- $10,001-$20,000 (Ex: Mexico, Russia)
- $20,001-$40,000 (Ex: Canada, Singapore)
- More than $40,000 (Ex: Kuwait, Norway)

E How likely do you think it is that the country’s population was majority Christian?

- Very Unlikely (0-20% chance)
- Unlikely (21-40% chance)
- Chances About Even (41-60% chance)
• Likely (61-80% chance)
• Very Likely (81-100% chance)

F How likely do you think it is that the country had large oil reserves?
• Very Unlikely (0-20% chance)
• Unlikely (21-40% chance)
• Chances About Even (41-60% chance)
• Likely (61-80% chance)
• Very Likely (81-100% chance)

G How likely do you think it is that the majority of the country’s population was white (Caucasian)?
• Very Unlikely (0-20% chance)
• Unlikely (21-40% chance)
• Chances About Even (41-60% chance)
• Likely (61-80% chance)
• Very Likely (81-100% chance)

H How much do you think the country spent annually on its military?27
• Very Little (less than $30 million)
• A Little ($30 to $120 million)
• About Average ($120 million to $600 million)
• A Large Amount ($600 million to $3.5 billion)
• A Very Large Amount (greater than $3.5 billion)

I How likely do you think it is that the country had been a U.S. military ally since World War II?
• Very Unlikely (0-20% chance)
• Unlikely (21-40% chance)
• Chances About Even (41-60% chance)
• Likely (61-80% chance)
• Very Likely (81-100% chance)

J What do you think was the total volume of import and export between the country and the U.S.?28

27The intervals are based on quintiles of countries’s military expenditure in 2005.
28The intervals are based on quintiles of total volume of trade between the U.S. and other countries in 2005.
• A Very Small Amount (less than $100 million)
• A Small Amount ($100 million to $350 million)
• An Average Amount ($350 million to $1.5 billion)
• A Large Amount ($1.5 billion to $10 billion)
• A Very Large Amount (greater than $10 billion)

K How likely do you think it is that the country had carried out a joint military exercise with the U.S.?
• Very Unlikely (0-20% chance)
• Unlikely (21-40% chance)
• Chances About Even (41-60% chance)
• Likely (61-80% chance)
• Very Likely (81-100% chance)

L Do you think the country had high levels or low levels of investment in U.S. businesses?
• Very high levels of investment in U.S. businesses
• High levels of investment in U.S. businesses
• Medium levels of investment in U.S. businesses
• Low levels of investment in U.S. businesses
• Very low levels of investment in U.S. businesses

D.8 Treatment Measures

We used two questions to measure how much the democracy condition affected subjects’ beliefs about the target country. We called these questions treatment measures because they measured the value of the treatment variable.

**Treatment Measure 1: Probability of Being in Each Regime Type**

Think about the country described in the scenario. We would like to know how you would characterize its government. How likely do you think it is that the country has the following types of government?

For each government type, we provide you with two reference countries.\(^{29}\)

\(^{29}\)Each respondent input her answers using one of the three following matrices randomly assigned to him or her. We do this to make sure responses are not driven by the example countries we provide.
Treatment Measure 2: Characteristics of Democracies

Think about the country described in the scenario.

For each of the following characteristics, please indicate if you think that there is more than a 50 percent chance that the country described in the scenario has the characteristic. (Select all that apply.)

(You can select none, one, or more than one.)

- The country has a freely elected head of government and legislative representatives that determine national policy.
- The country allows opposition parties that could realistically gain power through election.
- The country has free and independent media.
- The country allows people to openly practice their religion.
• The country has limitations on the executive authority through a legislature and an independent court system.

• The country allows for assembly, demonstration, and open public discussion.

D.9 Support for Military Action

The main outcome measure in the Tomz and Weeks survey experiment was whether respondents support the U.S. using military force against the country in the scenario. We asked the same question in our survey.

Support for Using Force Question

Think about the scenario you read.

By attacking the country’s nuclear development sites now, the U.S. could prevent the country from making any nuclear weapons.

Do you favor or oppose the U.S. using its armed forces to attack the country’s nuclear development sites?

• Favor strongly
• Favor somewhat
• Neither favor nor oppose
• Oppose somewhat
• Oppose strongly
• I don’t know

D.10 Mediation Questions

The mediation questions were used to determine the reasons why subjects supported or opposed use of force against the target country. We asked an open-ended mediation outcome along with the same questions Tomz and Weeks asked in their survey.

Open-ended Mediation Question

Why did you select that answer choice in the previous question?30

30This question was not be asked in the Separated-Placebos Design because we did not want to increase the complexity of an already complex design.
D.10.1 If the U.S. attacked...

Think about the country in the scenario you read. Suppose the U.S. uses armed forces to attack the country’s nuclear development sites.

Which of the following events do you think will have more than a 50% chance of happening? (Check all that apply.)

- The country will attack the U.S. or a U.S. ally.
- The U.S. military will suffer many casualties.
- The U.S. economy will suffer.
- The U.S.'s relations with other countries will suffer.
- The attack will prevent the country from making nuclear weapons in the short term.
- The attack will prevent the country from making nuclear weapons in the long term.

D.10.2 If the U.S. did not attack...

Think about the scenario you read. Suppose the U.S. does not use armed forces to attack the country’s nuclear development sites.

Which of the following events do you think will have more than a 50% chance of happening? (Check all that apply.)

- The country will build nuclear weapons.
- The country will threaten to use nuclear weapons against another country.
- The country will threaten to use nuclear weapons against the U.S. or a U.S. ally.
- The country will launch a nuclear attack against another country.
- The country will launch a nuclear attack against the U.S. or a U.S. ally.

D.10.3 Morality of Using Force

Think about the scenario you read. Do you think it is morally wrong for the U.S. military to attack the country’s nuclear development sites?

- It is morally wrong.
- It is not morally wrong.
- I don’t know.
D.11 Demographics Questions

We asked the demographics questions at the end of the survey. We did not want these questions to prime subjects and affect how they answer the previous questions. Because the demographics questions asked about identities that are fairly immutable, we did not think the previous questions affected how subjects answered them.

D.11.1 Education

What is the highest level of education you have completed?

- Less than high school
- High school
- Associate’s/Junior College
- Bachelor’s
- Graduate’s (Master’s, MBA, PhD, MD)
- I don’t know

D.11.2 Political Party

Generally speaking, do you usually think of yourself as a Republican, Democrat, Independent, or what?

- Strong Democrat
- Weak Democrat
- Independent, leaning Democrat
- Independent
- Independent, leaning Republican
- Weak Republican
- Strong Republican
- Other

D.11.3 Age

What is your age?

[Drop-down menu: 18 to older than 100]
D.11.4 Sex
What is your sex?

- Female
- Male
- Other

D.11.5 Political Ideology
On the scale below, 1 means extremely liberal and 7 means extremely conservative.
Where would you place yourself on the 7-point scale?
[7-point scale]

E Full Summary of “Democratic Peace” Survey Results

We use heteroscedasticity-robust standard errors in our OLS regressions. In coefficient plots, we report the point estimates along with the 95 percent (thick line) and 99 percent (thin line) confidence intervals.

E.1 Survey Procedure

We conducted our survey experiment in July 2015 using American respondents on Amazon.com’s Mechanical Turk. We used Qualtrics to administer our survey.

E.2 Balance Tests

We use balance tests to determine if the randomization procedure was carried out correctly. The results of the balance tests show that respondents assigned to treatment versus control had statistically indistinguishable background characteristics.

<table>
<thead>
<tr>
<th>Treatment Assignment</th>
<th>Vignette Type</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-democracy</td>
<td>Basic</td>
<td>513</td>
</tr>
<tr>
<td>Democracy</td>
<td>Basic</td>
<td>517</td>
</tr>
<tr>
<td>Non-democracy</td>
<td>Covariate Control</td>
<td>512</td>
</tr>
<tr>
<td>Democracy</td>
<td>Covariate Control</td>
<td>513</td>
</tr>
<tr>
<td>Non-democracy</td>
<td>ENE</td>
<td>516</td>
</tr>
<tr>
<td>Democracy</td>
<td>ENE</td>
<td>509</td>
</tr>
</tbody>
</table>
Figure 9: Demographic Variables by Experimental Condition

- Mean Age
- Percent Male
- Percent with College Degree
- Percent Democrat
- Mean Ideology Score (1 = Extremely Liberal; 7 = Extremely Conservative)

Basic Covariate Control ENE
Vignette Type
Value
Treatment Assignment Regime Type
Non-democracy Democracy
Figure 10: Balance Tests on Demographic Variables

Vignette Type

Mean Age

Percent Male

Percent with College Degree

Percent Democrat

Mean Ideology Score

(1 = Extremely Liberal; 7 = Extremely Conservative)

Difference of Means (Dem – NonDem)
Table 7: Balance Test: Results from Joint $F$-test Using All Five Demographics Variables to Predict Treatment Assignment

<table>
<thead>
<tr>
<th>Vignette Type</th>
<th>$F$-statistic</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>$F(5,965) = 0.19$</td>
<td>0.968</td>
</tr>
<tr>
<td>Covariate Control</td>
<td>$F(5,942) = 0.69$</td>
<td>0.632</td>
</tr>
<tr>
<td>ENE</td>
<td>$F(5,964) = 1.12$</td>
<td>0.349</td>
</tr>
</tbody>
</table>

E.3 Coding Placebo Test Results

In this subsection, we explain how we code responses to each of the placebo test questions.

A: Regions of the World

We reduce the regions to a single dimension $Y_{i,A}^N$: the sum of scores assigned to each region subject $i$ mentioned as one of the two most likely regions. North America or Western Europe have a score of 1, Central Asia and East Asia have a score of 0, and the Middle East & North Africa and Sub-Saharan Africa have a score of -1.

B: GDP per Capita

We define $Y_{i,B}^N$ as subjects’ response to the GDP per capita placebo test question. We scale the responses such that $Y_{i,B}^N$ equals the real-world median of the GDP per capita interval subject $i$ selects. For instance, in 2005, there were nine countries in the “More than $40,000” interval; the median GDP per capita among them was $58411.59. This would mean $Y_{i,B}^N = 58411.59$ if subject $i$ selects “More than $40,000.” As a robustness check, we also scale the responses ordinally so that $Y_{i,B}^N = 0$ when subject $i$ selects “Less than $500” and $Y_{i,B}^N = 4$ when she selects “More than $40,000”.

C: Religion

We define $Y_{i,C}^N$ as subjects’ response to the religion placebo test question; we will scale the responses so that $Y_{i,C}^N$ equals the mean of the probability interval subject $i$ selects.

D: Oil Reserves

We define $Y_{i,D}^N$ as one minus the subjects’ response to the oil reserves placebo test question: $Y_{i,D}^N$ equals one minus the mean of the probability interval subject $i$ selects.\(^\text{31}\)

\(^\text{31}\)Because we hypothesize that subjects think the democratic country is less likely to have had large oil reserves, we invert the responses so the direction of the confounding is the same as the direction in the other placebo tests.
E: Race
We define $Y^N_E$ as subjects’ response to the race placebo test question; we scale the responses so that $Y^N_{i,E}$ equals the mean of the probability interval subject $i$ selects.

F: Military Alliance
We define $Y^N_F$ as subjects’ response to the military alliance placebo test question; we scale the responses so that $Y^N_{i,F}$ equals the mean of the probability interval subject $i$ selects.

G: Trade with the U.S.
We define $Y^N_G$ as subjects’ response to the level of trade placebo test question. We scale the responses so that $Y^N_{i,G}$ equals the real-world median of the trade volume interval subject $i$ selects. For instance, in 2005, there are 38 countries in the “A Very Large Amount (greater than $10$ billion)” interval; the median volume of trade between these countries and the U.S. was $30.114$ billion. This would mean $Y^N_{i,G} = 30114000000$ if subject $i$ selects “A Very Large Amount.” As a robustness check, we also scale the responses ordinally so that $Y^N_{i,G} = 0$ when subject $i$ selects “A Very Small Amount” and $Y^N_{i,G} = 4$ when she selects “A Very Large Amount.”

H: Joint Military Exercise
We define $Y^N_H$ as subjects’ response to the joint military exercise placebo test question; we scale the responses so that $Y^N_{i,H}$ equals the mean of the probability interval subject $i$ selects.

I: Foreign Direct Investment
We define $Y^N_I$ as subjects’ response to the FDI test question; we scale the responses so that $Y^N_{i,H}$ corresponds to an ordinal scale with “very high levels of investment” being a 4 and “very low levels of investment” being a 0.

J: Military Capability
We define $Y^N_J$ as subjects’ response to the military capability placebo test question; we scale the responses so that $Y^N_{i,J}$ equals the real-world median of the military expenditure interval subject $i$ selects. For instance, in 2005, there are 36 countries in the “A Very Large Amount (greater than $3.5$ billion)” interval; the median value among them was $9.1815$ billion. This means that $Y^N_{i,J} = 91815000000$ when subject $i$ selects “greater than $3.5$ billion.” As a robustness check, we also scale the responses ordinally so that $Y^N_{i,J} = 0$ when subject $i$ selects “Very Little” and $Y^N_{i,J} = 4$ when she selects “A Very Large Amount.”

We do not include this placebo variable in our main results because we do not have good theoretical justification for using it. As reported in D.5.1, none of the military capability variables were statistically significant between democracies and non-democracies.
E.4 Placebo Test Questions

E.4.1 Analysis of Placebo Test Responses

In our analyses, we use both non-standardized and standardized versions of the placebo response variable. Unless noted, we analyze the results for the Basic, Covariate Control, and the ENE designs separately. Define $Y_{i,j}^N$ as subject $i$’s non-standardized response to placebo test question $j$. There are a total of $Q$ respondents; each respondent $i$ is assigned to vignette type $V_i$. The mean non-standardized placebo response within each vignette type $v$ is $Y_{j,v}^N = \frac{1}{Q} \sum_{i=1}^{Q} [Y_{i,j}^N 1(V_i = v)]$.

Next, we define the standardized version $Y_{i,j}$ as:

$$Y_{i,j} = \frac{Y_{i,j}^N - \bar{Y}_{j,v}^N}{\sqrt{\frac{1}{Q} \sum_{i=1}^{Q} (Y_{i,j}^N - \bar{Y}_{j,v}^N)^2}}$$

(3)

For each placebo test and vignette type, we produce coefficient plots that show the estimated standardized and non-standardized difference-of-means between the democracy condition and the non-democracy condition. We use heteroscedasticity-robust standard errors in our regressions. In the coefficient plots, we also report the 95 percent (thick line) and 99 percent (thin line) confidence intervals.

We predict the difference-in-means will be largest (furthest from 0 in the positive direction) for the Basic Vignettes, medium for the Control Vignettes, and smallest (closest to 0) for the ENE Vignettes.32

For each vignette type, we estimate $E(\tau_{i,j}) = E[Y_{i,j}(Z_i = 1) - Y_{i,j}(Z_i = 0)]$ using $\hat{\beta}_{1,j}$ from the regression $E(Y_{i,j}|Z_i) = \beta_{0,j} + \beta_{1,j}Z_i$.

E.4.2 Placebo Test Results

Responses to the placebo test questions are presented in this subsection. We visualize the distribution of responses using bar charts and density plots. We show the results from our regression analysis using coefficient plots.

32To simplify the presentation, all placebo variables are coded so that more positive values correspond to the values more common in democracies.
Figure 11: Placebo A: Most Likely Regions Distribution of Responses

This figure shows the distribution of responses across different geographic regions (Middle East/North Africa, Sub-Saharan Africa, East Asia, Central Asia, Western Europe, North America) for two treatment assignment regimes: Basic Covariate Control and Covariate Control. The y-axis represents the geographic regions, and the x-axis shows the proportion of respondents who selected each region as a most likely region. The bars are color-coded to indicate the treatment assignment regime type: non-democracy (yellow) and democracy (blue).

The graphs indicate that the proportion of respondents selecting each region varies significantly across regions and treatment assignment regimes.
Figure 12: Placebo B: GDP per Capita Distribution of Responses
Figure 13: Placebo C: Likelihood of Being Majority Christian Distribution of Responses
Figure 14: Placebo D: Likelihood of Not Having Large Oil Reserves Distribution of Responses
Figure 15: Placebo E: Likelihood of Being Majority White Distribution of Responses

Treatment Assignment Regime Type
- Non-democracy
- Democracy
Figure 16: Placebo F: Likelihood of Military Alliance with the U.S. since World War II
Distribution of Responses
Figure 17: Placebo G: Level of Trade with the U.S. Distribution of Responses

The figure shows the distribution of responses for Total Trade Flow (log millions USD) across different treatment assignment regimes. The x-axis represents the total trade flow in log millions USD, ranging from 3 to 9. The y-axis represents the density of responses.

Three panels are shown:
- **Basic Covariate Control**: This panel shows the distribution for non-democracy and democracy regimes. The densities for both regimes are visualized using different colors, with non-democracy in orange and democracy in gray.
- **Covariate Control**: This panel provides a closer look at the basic covariate control scenario, maintaining the same color scheme for non-democracy and democracy.
- **ENE**: This panel focuses on the ENE scenario, again with the same color scheme for non-democracy and democracy.

The figure highlights the differences in trade flow distribution between non-democracy and democracy regimes, providing insights into the impact of regime type on total trade flow.
Figure 18: Placebo H: Likelihood of Joint Military Exercise with the U.S. Distribution of Responses
Figure 19: Placebo I: Level of Investment in U.S. Businesses Distribution of Responses
Figure 20: Placebo J: Military Spending Distribution of Responses

![Graph showing military spending distribution for different likelihoods of not having large oil reserves, categorized by treatment assignment regime type (Non-democracy and Democracy).]

**Treatment Assignment Regime Type**

- Non-democracy
- Democracy

**Density**

- Very Unlikely
- Unlikely
- Odds About Even
- Likely
- Very Likely

**Likelihood of Not Having Large Oil Reserves**
Figure 21: Placebo K: Most Likely Countries Distribution of Responses
For Placebo Outcomes B, G, and J, we took the natural log of the non-standardized USD outcome before standardizing within vignette type.
Figure 23: Placebo Test Questions Results (Standardized) including Military Spending

For Placebo Outcomes B, G, and J, we converted the non-standardized USD outcomes to ordinal values (0 to 6 for Placebo Outcome B; 0 to 4 for Placebo Outcomes G and J) before standardization.
For Placebo Tests B, G, and J, the outcomes are in their original USD values.
E.4.3 Joint Test of Placebo Outcomes Using NPC

We will jointly test if there exists imbalances among the placebo test responses A through I using the non-parametric combination test (NPC).\textsuperscript{33} Note that the null hypothesis we assume for NPC is different from those we assumed for the previous hypothesis tests. For NPC, we assume the global sharp null, that is the regime type of the country in the scenario has no effect for every subject’s response to every placebo test question.

For each vignette type, we use the following algorithm to calculate a global p-value:

A. Calculate a vector of observed test statistics $T^{\text{obs}} = (T^{\text{obs}}_A, T^{\text{obs}}_B, ..., T^{\text{obs}}_m, ..., T^{\text{obs}}_I)$ corresponding to the nine partial placebo tests. We use the difference-of-means between the treatment and control group as our test statistic:

$$T^{\text{obs}}_m = \mathbb{E}(Y_m | Z = 1) - \mathbb{E}(Y_m | Z = 0)$$

for each placebo test $m$.

B. Repeat the following $Q$ times:

a) Randomly permute the group labels $Z$ of units that are exchangeable under the sharp null.

b) In each permutation $q \in \{1, ..., Q\}$, calculate the vector $T^*_q = (T^*_A, T^*_B, ..., T^*_m, ..., T^*_I)$ of values of the nine test statistics.

C. Presuming that the partial test statistics are expected to be large in the alternative, let $\bar{U}_j(t) = Q^{-1} \sum_{q=1}^{Q} 1(T^*_m \leq t)$ be the estimated significance level for any test statistic $t \in \mathbb{R}^1$ corresponding to partial test $m$. Calculate the vector of estimated significance levels for the observed data: $\hat{p} = (\hat{p}_A, \hat{p}_B, ..., \hat{p}_m, ..., \hat{p}_I)$, where $\hat{p}_m = \bar{U}_m(T^{\text{obs}}_m)$. Then, for each permutation $q$, calculate the vector of pseudo p-values $\hat{U}^*_q = (\hat{U}^*_A, \hat{U}^*_B, ..., \hat{U}^*_m, ..., \hat{U}^*_I)$, where $\hat{L}^*_m = \hat{L}(T^*_m)$.

D. Using combining function $\psi$, combine the vector of nine estimated significance levels into a global test statistic $T''^{\text{obs}} = \psi(\hat{p})$, which captures the observed divergence from the null across all nine partial tests. Then calculate the analogous statistic $T''^*_b = \psi(\hat{L}^*_b)$ for each permutation $q$. We will report results using the following three combining functions:

a) Fisher’s: $\psi_a = -\sum_m \log(p_m)$

b) Liptak’s: $\psi_b = -\sum_m \Phi^{-1}(p_m)$, where $\Phi^{-1}$ is the inverse of the normal CDF

c) Tippett’s: $\psi_c = -\min_m (p_m)$

E. Estimate the combined p-value of the global test as $\hat{p}_\psi'' = Q^{-1} \sum_{q=1}^{Q} 1(T''^*_q \geq T''^{\text{obs}})$.

In Table 8, we report the NPC results by vignette type. We use all three combining functions and use $Q = 10000$ permutation for each analysis. As the results show, the global p-values are orders of magnitude smaller for the Basic and Covariate Control design than for the ENE design. This suggests that the imbalance in placebo outcomes, as a whole, are far worse in those first two designs than in the ENE design.

Table 8: p-values from Joint Test of Placebo Outcomes Using NPC Combining Function Used

<table>
<thead>
<tr>
<th></th>
<th>Basic</th>
<th>Covariate Control</th>
<th>ENE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher’s</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>0.0055</td>
</tr>
<tr>
<td>Liptak’s</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>0.0191</td>
</tr>
<tr>
<td>Tippett’s</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>0.0101</td>
</tr>
</tbody>
</table>

E.5 Treatment Measures

E.5.1 Coding and Analysis of Treatment Measure Results

Type Question 1: Probability of Being in Each Regime Type Our Treatment Measure 1 measures subjects’ beliefs about how democratic the target country is. We call this latent variable $D_i$, which we proxy using our imputed measure $R_1 \ i$ based on subject $i$’s response to Treatment Measure 1.

Define $K_{i,j}$ as subject $i$’s response to regime type category $j \in \{1, 2, ..., 5\}$. Using these responses, we impute $R_1 \ i$, which ranges from -10 to 10 — much like the Polity score. The procedure for imputing $R_1 \ i$ is:

A. First, we impute $K_{i,j}$, the probability subject $i$ assigned to regime type category $j$. Recall that in the survey, each subject $i$ selected a probability interval $[K_{a,i,j}, K_{b,i,j}]$ for each regime type category $j$. We reduce the dimensionality of each subject’s responses by defining $K_{i,j}$ as the mean of the probability interval $[K_{a,i,j}, K_{b,i,j}]$.

$$K_{i,j} = (K_{a,i,j} + K_{b,i,j})/2$$

B. Let $R_{1,i,j}$ be the normalized probability subject $i$ assigns to regime type category $j$. For $j \in \{1, 2, ..., 5\}$, we normalize $K_{i,j}$ so that $\sum_j K_{i,j} = 1$, meaning that the probabilities each subject assigned to the regime type categories will sum to one.

$$R_{1,i,j} = \frac{K_{i,j}}{\sum_j K_{i,j}}$$

C. Finally we impute $R_{1,i}$. For $j \in \{1, 2, ..., 5\}$, we multiply the mean polity score of the $j$th regime type category $O_j$ \(^{34}\) by $R_{1,i,j}$ then we sum these five products. In short, we calculate the expected value of the “Polity score” for each subject $i$’s response.

$$R_{1,i} = \sum_j (O_j \cdot R_{1,i,j})$$

Define the treatment effect of the democracy condition ($Z$) on responses to this treatment measure question as $\tau_{i,R_1} = R_1 (Z_i = 1) - R_1 (Z_i = 0)$. For each vignette type, we estimate $\mathbb{E}(\tau_{i,R_1})$ using $\hat{\beta}_{1,R_1}$ from the regression: $\mathbb{E}(R_1 \mid Z_i) = \beta_{0,R_1} + \beta_{1,R_1} Z_i$.

\(^{34}\)The five regime types we present in our survey are fully democratic, democratic, somewhat democratic/somewhat non-democratic, non-democratic, and fully non-democratic. These correspond to the following Polity 4 regime types: full democracy (10), democracy (6 to 9), open anocracy (1 to 5), closed anocracy (-5 to 0), and autocracy (-10 to -6). We choose not to use the Polity 4 terms because they are too specialized for our respondents to understand.
**Regime Type Question 2: Characteristics of Democracies**  
We define \( R_2i \) as the number of democratic characteristics respondent \( i \) selected, which serves as a proxy for how democratic respondent \( i \) thought the target country is.

Define the treatment effect of the democracy condition \((Z)\) on responses to this treatment measure question as \( \tau_{i,R_2} = R_2i(Z_i = 1) - R_2i(Z_i = 0) \). For each vignette type, we estimate \( E(\tau_{i,R_2}) \) using \( \hat{\beta}_{1,R_2} \) from the regression: \( E(R_2i|Z_i) = \beta_{0,R_2} + \beta_{1,R_2}Z_i \).

**E.5.2 Treatment Measures Results**

Figure 25: Treatment Measure: Probability of Each Regime Type

![Graph showing the probability of each regime type](image)

We compare the mean probability subjects assigned to each regime type between those who received the Democracy vignette and those who received the Non-democracy vignette. For each subject, we normalize the probability she assigned to each regime type so that her probabilities sum up to 100 percent.
In the coefficient plot above, we estimate the difference in the likelihood of respondents assign to each regime type by vignette type.
We impute “Polity scores” based on responses using the method described in Subsection E.5.1.

In this coefficient plot above, we present the difference in the mean imputed “Polity score” between treatment and control for each vignette type.
In this coefficient plot above, we present the difference in the proportion of respondents who selected each characteristic of democracies by each vignette type.

In the coefficient plot above, we present the difference in the mean number of characteristics selected between treatment and control by each vignette type.
E.6 Substantive Outcome: Support for War

E.6.1 ITT and IV Analyses

ITT Estimates  First, we estimate the effect of the democracy condition on support for military action against the target country. We call this the intention-to-treat (ITT) estimate. We define $S_i$ as subject $i$’s response to the support for using force question (the typical outcome variable used in experimental studies of the democratic peace), such that $S_i = 4$ for “strongly oppose” and $S_i = 0$ for “strongly support.”\(^{35}\) Define the treatment effect of the democracy condition $Z$ on response $S$ as $\zeta_i = S_i(Z_i = 1) - S_i(Z_i = 0)$.

For each vignette type, we estimate $E(\zeta_i)$ using the coefficient estimate $\hat{\beta}_1$ from the regression:

$$E(S_i|Z_i) = \beta_0 + \beta_1 Z_i.$$  

As a robustness check, we repeat the same analysis but using a dichotomous outcome variable $S^*$ such that $S^*_i = 1$ if $S_i > 3$ and $S^*_i = 0$ otherwise.

IV Estimates of Democracy’s Effect on Support for Military Action  For our IV estimate of democracy’s effect on support for military action $S$, we use the treatment assignment $Z$ as an instrument for subjects’ perceptions of the target country’s level of democracy. We use $R1_i$, the imputed “Polity score” from Treatment Measure 1, as a proxy for subjects’ perceptions of the target country’s democracy level. For ease of exposition in our coefficient plots, we scale the effect to a perceived increase of 10 Polity points. Note that $R1_i$ might be measuring $D_i$, the latent variable, with error; therefore, we define the causal effects we seek to estimate with respect to $R1_i$ and not $D_i$.

Define $\rho = E[S_i(R1_i = 1) - S_i(R1_i = 0)|R1_i(Z_i = 1) > R1_i(Z_i = 0)]$ as the local average treatment effect (LATE). We estimate the LATE $\rho$ using the Wald Estimator since $Z$ is binary:

$$\rho_{WALD} = \frac{E(S_i|Z_i = 1) - E(S_i|Z_i = 0)}{E(R1_i|Z_i = 1) - E(R1_i|Z_i = 0)}.$$  

For each vignette type, we estimate $\rho$ and calculate a confidence interval. Note that the IV estimates are likely biased for vignette types that violate the exclusion restriction. Significant imbalance on dispositive placebo variables for the Basic and Covariate Control design provides evidence against the exclusion restrict.

As a robustness check, we repeat the analysis above using the dichotomous measure of support for using force $S^*$ and a dichotomous treatment measure $R1^*$. Let $R1^*_i = 1$ if respondent $i$ indicates the aggressor country has a higher probability of being democratic or fully democratic than being non-democratic or fully non-democratic and let $R1^*_i = 0$ if she indicates otherwise.

---

\(^{35}\)Subjects who answered "don’t know" to the question were assigned $S = 2$. 
E.6.2 ITT and IV Results

Figure 29: ITT and IV Estimates: Ordinal Measure of Support for Using Force

The dependent variable is support for using force measured using a 5 point ordinal scale. Those who
strongly favor using force is coded 4 and those who strongly oppose using force is coded 0. Those who
responded with “don’t know” is coded 2.

For the Imputed Polity Score Treatment Measure, we combine the probabilities each respondent assign to
the five regime types into a single score from -10 to 10, akin to the Polity score. The score is calculated by
summing the product of the probability respondents assign to each regime type and the mean real-world
Polity score for that regime type. See E.5.1 for more details.

For the Dichotomous Treatment Measure, we code that respondents perceive the country is a democracy
when they indicate the country has a higher probability of being democratic or fully democratic than being
non-democratic or fully non-democratic.
The dependent variable is dichotomous measure of support for using force. Those who “favor strongly” and “favor somewhat” using force are coded 1 and 0 otherwise.

For the Imputed Polity Score Treatment Measure, we combine the probabilities each respondent assign to the five regime types into a single score from -10 to 10, akin to the Polity score. The score is calculated by summing the product of the probability respondents assign to each regime type and the mean real-world Polity score for that regime type. See E.5.1 for more details.

For the Dichotomous Treatment Measure, we code that respondents perceive the country is a democracy when they indicate the country has a higher probability of being democratic or fully democratic than being non-democratic or fully non-democratic.
We perform the same type of analysis seen in Figures 29 and 30 except we examine the two different versions of the ENE Design. Recall that in one version of the ENE Design, the country started out a fragile democracy and in the other version, the country started out as a fragile dictatorship.
E.7 Abstract Encouragement Design

Respondents had $\frac{1}{2}$ probability of being randomly assigned to read instructions that encouraged them to consider the vignette scenario in the abstract. They were told “For scientific validity the situation is general, and is not about a specific country in the news today.” We determine those assigned to the Abstract Encouragement Design do not exhibit less imbalance in their placebo test responses. Figures 32 and 33 show that respondents in both groups exhibit similar levels of imbalance and imbalance in the same direction in almost all the placebo outcomes.
Figure 32: Effect of the Abstract Encouragement Design (Standardized)
Figure 33: Effect of the Abstract Encouragement Design (Non-standardized)
E.8 Question Block Order Does Not Affect Substantive Outcome Measure

E.8.1 Question Block Order in Together-Placebos Design

We examine whether the order of the question blocks changes how respondents answer the substantive outcome question (i.e., support for using force). In the Together-Placebos, we randomized the order of four question blocks relative to the substantive outcome block. In Figure 34, for each of the four blocks, we estimate the difference-in-mean in support for using force between respondents who saw the block before the substantive outcome block and those who saw it after. We find that the order of the question blocks mostly do not affect respondents’ support for using force.
E.8.2 Placebo Test Questions Do Not Affect Substantive Outcome Measure

In the Separated-Placebos Design, we randomize whether zero, one, two, or three placebo test question(s) appear(s) before the support for using force question. The three placebo test questions we use are:

• Regions of the world
• GDP per capita
• Likelihood of being majority Christian
When placebo test question(s) appear(s) before the support for using force question, which questions are asked and the order of those questions (when there are two or more such questions) are randomized. Let \( X_1 \) be a dummy for whether the GDP per capita question appeared before the outcome question; let \( X_2 \) be a dummy for whether the religion question appeared before the outcome question; let \( X_2 \) be a dummy for whether the regions question appeared before the outcome question; and let \( Y \) be responses to the outcome measure.

First, we test if there is a significant difference-in-means in \( Y \) depending on \( X_j \) for \( j \in \{1, 2, 3\} \). The results are reported in Figure 35. Whether each of the placebo test question appeared before the support for using force question did not affect respondents’ support for using force.

Figure 35: Separated-Placebos Design Results

\[
\begin{align*}
\text{DV: Support for Using Force (Ordinal Scale: 0 to 4)} \\
\text{Placebo Test Question} & \quad \text{Regions} \quad \text{GDP per Capita} \quad \text{Religion} \\
\end{align*}
\]

\[
\begin{align*}
\text{Difference in Support of Using Force} \\
-0.2 & -0.1 & 0.0 & 0.1 & 0.2 \\
\end{align*}
\]

\[
\begin{align*}
\text{DV: Support for Using Force (Dichotomous Measure)} \\
\text{Placebo Test Question} & \quad \text{Regions} \quad \text{GDP per Capita} \quad \text{Religion} \\
\end{align*}
\]

\[
\begin{align*}
\text{Difference in Proportion Who Support of Using Force} \\
-0.10 & -0.05 & 0.00 & 0.05 \\
\end{align*}
\]

We also estimate the effect of the placebo test questions appearing before the support for using force question using a fully interacted regression:

\[
E(Y_i|X_{i,1}, X_{i,2}, X_{i,3}) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \beta_4 X_{i,1} X_{i,2} + \beta_5 X_{i,1} X_{i,3} + \beta_6 X_{i,2} X_{i,3} + \beta_7 X_{i,1} X_{i,2} X_{i,3}
\]  \hspace{1cm} (4)

95
As Table 9 shows, none of the coefficients are statistically significant. Respondents’ support for using force is unchanged by all combinations of placebo test questions appearing before it.

Finally, we consider whether the placebo test questions appearing before the support for using force question affects our ITT estimates (estimates of the effect of treatment assignment Z on Y). For \( j \in \{1, 2, 3\} \), we estimate \( \beta_{3,j} \) from the following regression:

\[
E(Y_i|X_{ij}, Z_i) = \beta_{0,j} + \beta_{1,j}X_{ij} + \beta_{2,j}Z_i + \beta_{3,j}X_{ij}Z_i
\]

In Tables 10 and 11, we show that \( \hat{\beta}_{3,j} \) for all \( j \) are not statistically significant at \( p = 0.05 \). This means that each of the placebo test questions appearing before the support for war question does not seem to affect our ITT estimates.

Table 9: Results from the Fully Interacted Regressions

<table>
<thead>
<tr>
<th></th>
<th>DV 1: Support for Using Force (Ordinal Scale)</th>
<th>DV 2: Support for Using Force (Dichotomous Measure)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>GDP per Capita</td>
<td>0.058</td>
<td>-0.031</td>
</tr>
<tr>
<td></td>
<td>(0.129)</td>
<td>(0.049)</td>
</tr>
<tr>
<td>Religion</td>
<td>0.042</td>
<td>-0.039</td>
</tr>
<tr>
<td></td>
<td>(0.131)</td>
<td>(0.050)</td>
</tr>
<tr>
<td>Regions</td>
<td>0.005</td>
<td>-0.056</td>
</tr>
<tr>
<td></td>
<td>(0.130)</td>
<td>(0.047)</td>
</tr>
<tr>
<td>GDP per Capita \times Religion</td>
<td>-0.183</td>
<td>-0.042</td>
</tr>
<tr>
<td></td>
<td>(0.204)</td>
<td>(0.075)</td>
</tr>
<tr>
<td>GDP per Capita \times Regions</td>
<td>0.063</td>
<td>0.091</td>
</tr>
<tr>
<td></td>
<td>(0.204)</td>
<td>(0.075)</td>
</tr>
<tr>
<td>Religion \times Regions</td>
<td>-0.083</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>(0.204)</td>
<td>(0.075)</td>
</tr>
<tr>
<td>GDP per Capita \times Religion \times Regions</td>
<td>0.175</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>(0.289)</td>
<td>(0.106)</td>
</tr>
<tr>
<td>Constant</td>
<td>1.727***</td>
<td>0.370***</td>
</tr>
<tr>
<td></td>
<td>(0.067)</td>
<td>(0.025)</td>
</tr>
</tbody>
</table>

*Note:* *p<0.05; **p<0.01; ***p<0.001
Table 10: Effect of Placebo Tests on ITT Estimates

<table>
<thead>
<tr>
<th>DV: Support for Using Force (Ordinal Scale)</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democracy</td>
<td>−0.595***</td>
<td>−0.455</td>
<td>−0.436***</td>
</tr>
<tr>
<td></td>
<td>(0.090)</td>
<td>(0.363)</td>
<td>(0.102)</td>
</tr>
<tr>
<td>Regions</td>
<td>−0.053</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.093)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Democracy × Regions</td>
<td>0.208</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.130)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GDP per Capita</td>
<td>−0.031</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.028)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Democracy × GDP per Capita</td>
<td>−0.003</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.041)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Religion</td>
<td>−0.005***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Democracy × Religion</td>
<td>0.0005</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>2.029***</td>
<td>2.290***</td>
<td>2.257***</td>
</tr>
<tr>
<td></td>
<td>(0.066)</td>
<td>(0.242)</td>
<td>(0.065)</td>
</tr>
</tbody>
</table>

*Note:* *p<0.05; **p<0.01; ***p<0.001
Table 11: Effect of Placebo Tests on ITT Estimates

<table>
<thead>
<tr>
<th>DV 2: Support for Using Force (Dichotomous Measure)</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democracy</td>
<td>−0.173(*)</td>
<td>−0.044</td>
<td>−0.127(*)</td>
</tr>
<tr>
<td></td>
<td>(0.034)</td>
<td>(0.133)</td>
<td>(0.037)</td>
</tr>
<tr>
<td>Regions</td>
<td>−0.031</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.036)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Democracy × Regions</td>
<td>0.062</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.048)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GDP per Capita</td>
<td>−0.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Democracy × Regions</td>
<td>−0.010</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Religion</td>
<td></td>
<td>−0.003(*)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.001)</td>
<td></td>
</tr>
<tr>
<td>Democracy × Religion</td>
<td>0.0004</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.426(*)</td>
<td>0.432(*)</td>
<td>0.490(*)</td>
</tr>
<tr>
<td></td>
<td>(0.025)</td>
<td>(0.094)</td>
<td>(0.024)</td>
</tr>
</tbody>
</table>

Note: \*p<0.05; \*\*p<0.01; \*\*\*p<0.001

E.9 Substantive Outcome Question Does Not Affect Severity of Imbalance in Placebo Outcomes

In the Together-Placebos design, we randomize whether the substantive outcome question (i.e., support for using force) comes before or after the placebo test question block. In this subsection, we analyze whether the support for using force question impacts the severity of imbalance in the placebo outcomes using a difference-in-difference approach.

Let $F_i$ be an indicator variable for whether subject $i$ answered the support for using force question before the placebo test questions. Adopting notation from previous subsections, let $Y_{i,j}$ be subject $i$’s standardized response to placebo test question $j$ and $Z_i$ be her treatment assignment. We estimate
\[ \mathbb{E} \{ Y_{i,j}(Z_i = 1, F_i = 1) - Y_{i,j}(Z_i = 0, F_i = 1) \} - \mathbb{E} \{ Y_{i,j}(Z_i = 1, F_i = 0) - Y_{i,j}(Z_i = 0, F_i = 0) \} \]

using \( \hat{\beta}_{3,j} \) from the regression:

\[
\mathbb{E}(Y_{i,j}|Z_i, F_i) = \beta_{0,j} + \beta_{1,j}Z_i + \beta_{2,j}F_i + \beta_{3,j}Z_iF_i \tag{5}
\]

We report our estimates for \( \beta_{3,j} \) in Figure 36 for each vignette type. The results demonstrate that answering the substantive outcome question does not affect the severity of imbalance respondents exhibit in their placebo outcomes.
### E.10 Attention Checks

We use three measures to check how much respondents are paying attention to our survey. First, we analyze respondents’ answers to the attention check question. Thirty-eight respondents, or 1.31 percent of respondents, failed the attention check.
Second, we examine whether respondents took too little or too much time to complete the survey. Those who spent too little time likely rushed through the questions; those who took too much time might have been pre-occupied with other activities. The average amount of time respondents’ took to complete the survey is 12.91 minutes and the median is 10.70 minutes. We code respondents who were in the bottom and top five percentile of time spent (less than 4.65 minutes or more than 25.83 minutes) as inattentive.

Finally, we look at how frequently respondents chose the first or last answers in the multiple-choice placebo test questions; respondents who are rushing through the survey are likely to simply click on the first or last answer choice. Thirty-three percent of respondents did not select the first/last answer choices for any of the 10 questions; only six respondents exclusively selected the first/last answer choices.

For dimension reduction, we using principal component analysis (PCA) to combine the three measures into a single principal component that measures attentiveness. We use the PCA score to test whether respondents paid attention produced greater imbalance in their placebo tests. Note that a higher PCA score means that the respondent is more attentive.

In Table 12, the interaction effect between treatment assignment Z and the PCA score is statistically significant at $\alpha = 0.05$ for seven out of 10 placebo outcomes. Furthermore, the positive signs suggests that the more attentive respondents are, the more likely they think the country described as a democracy in the scenario has real-world characteristics of democracies.
Table 12: Attentiveness and Responses to Placebo Test Questions

<table>
<thead>
<tr>
<th>DV: Responses to Placebo Test Questions A through E</th>
<th>(A)</th>
<th>(B)</th>
<th>(C)</th>
<th>(D)</th>
<th>(E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democracy</td>
<td>0.455***</td>
<td>0.293***</td>
<td>0.403***</td>
<td>0.334***</td>
<td>0.140***</td>
</tr>
<tr>
<td></td>
<td>(0.037)</td>
<td>(0.037)</td>
<td>(0.035)</td>
<td>(0.035)</td>
<td>(0.037)</td>
</tr>
<tr>
<td>PCA Score</td>
<td>−0.115***</td>
<td>−0.163***</td>
<td>−0.410***</td>
<td>−0.477***</td>
<td>−0.040</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.036)</td>
<td>(0.030)</td>
<td>(0.030)</td>
<td>(0.032)</td>
</tr>
<tr>
<td>Democracy × PCA Score</td>
<td>−0.042</td>
<td>0.102*</td>
<td>0.119*</td>
<td>0.171***</td>
<td>−0.037</td>
</tr>
<tr>
<td></td>
<td>(0.046)</td>
<td>(0.048)</td>
<td>(0.048)</td>
<td>(0.047)</td>
<td>(0.046)</td>
</tr>
<tr>
<td>Constant</td>
<td>−0.218***</td>
<td>−0.132***</td>
<td>−0.164***</td>
<td>−0.121***</td>
<td>−0.068**</td>
</tr>
<tr>
<td></td>
<td>(0.022)</td>
<td>(0.027)</td>
<td>(0.024)</td>
<td>(0.024)</td>
<td>(0.025)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DV: Responses to Placebo Test Questions F through J</th>
<th>(F)</th>
<th>(G)</th>
<th>(H)</th>
<th>(I)</th>
<th>(J)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democracy</td>
<td>0.284***</td>
<td>0.117**</td>
<td>0.184***</td>
<td>0.172***</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>(0.036)</td>
<td>(0.036)</td>
<td>(0.037)</td>
<td>(0.036)</td>
<td>(0.037)</td>
</tr>
<tr>
<td>PCA Score</td>
<td>−0.358****</td>
<td>−0.337***</td>
<td>−0.374***</td>
<td>−0.369***</td>
<td>−0.025</td>
</tr>
<tr>
<td></td>
<td>(0.031)</td>
<td>(0.032)</td>
<td>(0.031)</td>
<td>(0.032)</td>
<td>(0.035)</td>
</tr>
<tr>
<td>Democracy × PCA Score</td>
<td>0.113*</td>
<td>0.162***</td>
<td>0.175***</td>
<td>0.158**</td>
<td>−0.023</td>
</tr>
<tr>
<td></td>
<td>(0.044)</td>
<td>(0.046)</td>
<td>(0.045)</td>
<td>(0.049)</td>
<td>(0.049)</td>
</tr>
<tr>
<td>Constant</td>
<td>−0.107****</td>
<td>−0.027</td>
<td>−0.053*</td>
<td>−0.050*</td>
<td>−0.002</td>
</tr>
<tr>
<td></td>
<td>(0.025)</td>
<td>(0.025)</td>
<td>(0.026)</td>
<td>(0.026)</td>
<td>(0.026)</td>
</tr>
</tbody>
</table>

Note: *p<0.05; **p<0.01; ***p<0.001
F Replication and Expansion of DeSante (2013)

F.1 Survey Overview

The design of our survey experiment closely follows that of DeSante’s (2013) survey experiment. Respondents are presented with two one-page welfare applications (based on the North Carolina Work First’s welfare application), side-by-side. Each application contains the name, household information, and welfare history of the applicant. The name on the right application is always “Laurie.” The name on the left application is “Latoya” with 1/2 probability (the treatment condition) and “Emily” with 1/2 probability (the control condition). For the Basic Design, we do not provide any additional information about each applicant. For the Covariate Control Design, we provide additional information about the applicant’s worker quality, which is rated as either “Poor” or “Excellent.” Respondents are assigned to the Basic Design with 1/2 probability and the Covariate Control Design with 1/2 probability.

Each applicant is described as needing $900, and the participants are asked to divide $1,500 between the two applicants. Respondents also have the option to give any dollar amount to offset the state’s budget deficit. After respondents allocate money to the two applicants, they are asked six placebo test questions.

F.2 Text of the Survey

In this subsection, we present the text of the survey in full.

F.2.1 Age

What is your age?

[Textbox]

F.2.2 Introduction

Researchers have been hired to consult with North Carolina Work First, that state’s welfare agency. On the next page, you will find two applicants for state assistance. These forms have been redacted to hide information that may identify individual applicants.

Each applicant has a state-assessed level of need of $900 per month. Your task is to allocate $1,500 between the two applicants. You can allocate any amount between $0 and $900 to each applicant. Any remaining funds will be used to offset the state’s budget deficit.

F.2.3 Welfare Application

Note: In Table 13, we list the experimental conditions in the survey experiment. We vary the name of the two applicants, whether the worker quality was listed, and if worker quality was listed, the worker quality assessment.
Table 13: Experimental Conditions in Replication and Expansion of DeSante (2013)

<table>
<thead>
<tr>
<th>Left Applicant Name</th>
<th>Right Applicant Name</th>
<th>Left Applicant Worker Quality</th>
<th>Right Applicant Worker Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emily</td>
<td>Laurie</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Emily</td>
<td>Laurie</td>
<td>Poor</td>
<td>Excellent</td>
</tr>
<tr>
<td>Emily</td>
<td>Laurie</td>
<td>Excellent</td>
<td>Poor</td>
</tr>
<tr>
<td>Latoya</td>
<td>Laurie</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Latoya</td>
<td>Laurie</td>
<td>Poor</td>
<td>Excellent</td>
</tr>
<tr>
<td>Latoya</td>
<td>Laurie</td>
<td>Excellent</td>
<td>Poor</td>
</tr>
</tbody>
</table>

Below are examples of applications respondents saw.

Figure 37: Welfare Applications from Replication and Expansion of DeSante (2013): No Worker Quality Assessment

<table>
<thead>
<tr>
<th>WORK FIRST ASSISTANCE APPLICATION</th>
<th>WORK FIRST ASSISTANCE APPLICATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applicant Name: Emily</td>
<td>Applicant Name: Laurie</td>
</tr>
<tr>
<td>Address:</td>
<td>Address:</td>
</tr>
<tr>
<td>Date of Application:</td>
<td>Date of Application:</td>
</tr>
<tr>
<td>Telephone:</td>
<td>Telephone:</td>
</tr>
<tr>
<td>County:</td>
<td>County:</td>
</tr>
<tr>
<td>Case No.</td>
<td>Case No.</td>
</tr>
<tr>
<td>District No.</td>
<td>District No.</td>
</tr>
<tr>
<td>HOUSEHOLD: List all household members requesting Emergency Assistance:</td>
<td>HOUSEHOLD: List all household members requesting Emergency Assistance:</td>
</tr>
<tr>
<td>Non-applicant household members are not required to provide a social security number, immigrant citizenship status</td>
<td>Non-applicant household members are not required to provide a social security number, immigrant citizenship status</td>
</tr>
<tr>
<td>Name</td>
<td>Date of Birth</td>
</tr>
<tr>
<td>------</td>
<td>---------------</td>
</tr>
<tr>
<td></td>
<td>09/01/2004</td>
</tr>
<tr>
<td></td>
<td>02/10/2005</td>
</tr>
<tr>
<td></td>
<td>02/12/2004</td>
</tr>
<tr>
<td></td>
<td>02/12/2005</td>
</tr>
<tr>
<td></td>
<td>02/12/2006</td>
</tr>
<tr>
<td></td>
<td>02/12/2007</td>
</tr>
<tr>
<td>Does the household include a child who meets the work-first age requirement?</td>
<td>X Yes</td>
</tr>
<tr>
<td>Is the child living with an adult who meets the work-first kinship requirement?</td>
<td>X Yes</td>
</tr>
<tr>
<td>Has anyone listed on the EA Application ever received EA?</td>
<td>X Yes</td>
</tr>
<tr>
<td>Does anyone live in the home that is not listed on the EA Application?</td>
<td>X Yes</td>
</tr>
<tr>
<td>If yes, is the individual(s) a co-resident?</td>
<td>X Yes</td>
</tr>
<tr>
<td>Total assessed monthly need:</td>
<td>$900.00</td>
</tr>
</tbody>
</table>

Applicant 1

Applicant Statement: I understand that it is against the law for me to make false statements and that I am subject to prosecution if I do. I declare under penalty of perjury that the information I have provided is true and complete statement of facts according to my best knowledge and belief. I certify, under penalty of perjury, that all persons for whom I am applying are U.S. citizens or qualified immigrants. I give the agency permission to verify any information necessary to determine my eligibility for Emergency Assistance.

Witness’s Signature: 
Applicant’s Representative’s Signature: 
Date: 

DSS-8169 (rev. 04/03) 
Family Support and Child Welfare Services Section

Applicant 2

Applicant Statement: I understand that it is against the law for me to make false statements and that I am subject to prosecution if I do. I declare under penalty of perjury that the information I have provided is true and complete statement of facts according to my best knowledge and belief. I certify, under penalty of perjury, that all persons for whom I am applying are U.S. citizens or qualified immigrants. I give the agency permission to verify any information necessary to determine my eligibility for Emergency Assistance.

Witness’s Signature: 
Applicant’s Representative’s Signature: 
Date: 

DSS-8169 (rev. 04/03) 
Family Support and Child Welfare Services Section
Figure 38: Welfare Applications from Replication and Expansion of DeSante (2013): With Worker Quality Assessment

F.2.4 Allocation of Money to Applicants

Your task is to allocate $1,500 between the two applicants. You can allocate any amount between $0 and $900 to each applicant. Any remaining funds will be used to offset the state’s budget deficit. Please enter three numbers below.

Amount allocated to Applicant 1 [$ ]
Amount allocated to Applicant 2 [$ ]
Amount allocated to reduce budget deficit [$ ]
Total [$ ]

Note: The total amount adjusts as respondents enter the amount they want to allocate to Applicant 1, Applicant 2, or the reduction of budget deficit. Respondents can only allocate $1,500.
F.2.5 Pop-up Window

Note: Respondents see this following screen before they answer all other questions. This allows them to open up a pop-up window with the applications.

Next, we are going to ask you some questions about each of the applicants. CLICK HERE to view the applications in a new window so you can refer to them.

We strongly recommend you click on the link above.

F.2.6 Placebo Test Questions

Note: Respondents are asked a series of questions about Applicant 1 followed by the same set of questions about Applicant 2. Within each applicant block, the order of the questions are randomized.

For each question, respondents are given the following answer choices:

- Very Unlikely
- Unlikely
- Odds About Even
- Likely
- Very Likely

The placebo test questions are:

- How likely do you think it is that [Name of the Applicant] has a high school diploma or GED?
- How likely do you think it is that [Name of the Applicant] has worked full-time, part-time or temporary during the previous 12 months?
- How likely do you think it is that [Name of the Applicant] has pending criminal charge(s) or criminal conviction(s)?
- How likely do you think it is that [Name of the Applicant] grew up in a low-income family?
- How likely do you think it is that [Name of the Applicant] has good parenting skills?
- How likely do you think it is that [Name of the Applicant] will have another child in the next two years?
F.2.7 Demographics Questions

Party ID  Generally speaking, do you usually think of yourself as a Republican, Democrat, Independent, or what?

- Strong Democrat
- Weak Democrat
- Independent, leaning Democrat
- Independent
- Independent, leaning Republican
- Weak Republican
- Strong Republican
- Other [Textbox]

Income  What is your total household income?

- Less than $10,000
- $10,000-$19,999
- $20,000-$29,999
- $30,000-$39,999
- $40,000-$49,999
- $50,000-$59,999
- $60,000-$69,999
- $70,000-$79,999
- $80,000-$89,999
- $90,000-$99,999
- $100,000-$149,000
- More than $150,000
- I prefer not to say
Education  What is the highest level of education you have completed?

- Less than high school
- High school
- Associate’s/Junior College
- Bachelor’s
- Graduate’s (Master’s, MBA, PhD, MD)
- I don’t know

Political Ideology  On the scale below, 1 means extremely liberal and 7 means extremely conservative.

Where would you place yourself on the 7-point scale?

[7-point scale]

F.2.8 Racial Resentment Measure

Note: For the racial resentment measures, respondents are given a series of statements and asked whether they agree with them. For each statement, they select their response from the following choices:

- Agree Strongly
- Somewhat Agree
- Neither Agree Nor Disagree
- Somewhat Disagree
- Disagree Strongly

We randomize the order in which the following statements appear.

Statements:

- Irish, Italians, Jewish and many other minorities overcame prejudice and worked their way up. Blacks should do the same without any special favors.
- Generations of slavery and discrimination have created conditions that make it difficult for blacks to work their way out of the lower class.
- It’s really a matter of some people not trying hard enough; if blacks would only try harder, they could be just as well off as whites.
- Over the past few years, blacks have gotten less than they deserve.
F.3 Results of Placebo Test Questions

F.3.1 Coding and Analyzing the Placebo Outcomes

Each respondent provided responses to the six placebo test questions. For each non-standardized response $R_{i,j}^N$ to placebo test question $j$, we code “Very Unlikely” as 1, “Unlikely” as 2, “Odds About Even” as 3, “Likely” as 4, and “Very Likely” as 5. We construct the standardized response $R_{i,j}$ using the method in the Democratic Peace survey experiment (see E.4.1 for details). The placebo outcome of interest is the difference between respondents’ assessment of the left applicant and the right applicant, i.e., the standardized paired difference $Y_{i,j} = R_{i,j}^{Left} - R_{i,j}^{Right}$. As a robustness check, we also analyze the non-standardized paired difference.

Let $Z_i$ be an indicator variable for whether respondent $i$ saw Latoya as the left applicant. For each vignette type, we estimate $E(\tau_{i,j}) = E[Y_{i,j}(Z_i = 1) - Y_{i,j}(Z_i = 0)]$ using $\hat{\beta}_{1,j}$ from the regression $E(Y_{i,j}|Z_i) = \beta_{0,j} + \beta_{1,j}Z_i$. We calculate heteroskedasticity-robust standard errors for our coefficients and present the 95 and 99 percent confidence intervals in our coefficient plots.

F.3.2 Placebo Test Results

In Figure 39, we plot the distribution of the paired differences for each placebo variable by vignette type. In Figure 40, we use coefficient plots to show our estimates of non-standardized paired difference for each placebo variable by vignette type. Overall, we find that the Covariate Control design exhibits imbalance in fewer placebo outcomes than the Basic design.
Figure 39: Replication and Expansion of DeSante (2013): Distribution of Responses to Placebo Test Questions by Treatment Assignment
G  Latura’s (2015) Survey Experiment

G.1  Overview of Experiment

Our experiment is embedded within a survey experiment performed by Audrey Latura, a PhD student at Harvard University. The substantive goal of Latura’s study is to examine whether people are more likely to accept a time-consuming promotion if their firm provides subsidized high-quality extended hours childcare. (Latura also examines the moderating effect of an information manipulation, but this is not relevant to our study.)

Our study involves examining whether respondent beliefs about other aspects of the firm in the scenario are affected by the manipulation about the availability of subsidized childcare. In the “Basic Design”, after reading about other aspects of their situation and the firm, some respondents are informed that “The company you work at subsidizes the cost of high-quality, extended-hours childcare for employees.” In the Embedded Natural Experiment (ENE) Design, all respondents are informed that their firm operates an “on-site, high-quality, extended-hours day-care center open from 6:00 AM to 10:00 PM on weekdays. The center is free for employees, but slots are allocated via random lottery.” The control group is informed that they did not win a day-care slot, the treatment group that they did.

Respondents are asked three (placebo) questions about the company: (1) Does the company offer other employee benefits; (2) does the company help employees to balance work-family issues; (3) does the company expect employees to answer work-related email on the weekends.
G.2 Text of the Survey

G.2.1 Demographics Questions

Education  What is the highest level of education you have completed?

- Less than high school
- High school / GED
- Some college
- 2-year college degree
- 4-year college degree
- Masters degree
- Doctoral degree
- Professional degree (JD, MD, MBA)

Age  What is your current age?

- Less than 18
- 18 to 21
- 22 to 25
- 26 to 30
- 31 to 35
- 36 to 40
- 41 to 45
- 46 to 50
- 51 to 60
- 61 to 70
- Over 70

Gender  What is your gender?

- Female
- Male
Passport  Do you currently have a valid passport?

- Yes
- No

G.2.2  Directions

In the next section, you will be presented with a brief text. Please read the text carefully, and when you have finished reading click on "Next." You will then be presented with an opinion question about the text.

G.2.3  Experimental Vignettes

Note: Respondents had 1/2 probability of being randomly assigned to the Basic Design or the ENE Design. Within each vignette design, respondents had 1/2 probability of being assigned to the treatment condition (subsidized childcare) or control condition (no subsidized childcare). Female respondents saw an extra paragraph at the end of the ENE vignettes.

Basic Design Text  You work at a company where you have recently won an award for talented junior employees. Now, you have been promoted to a mid-level management position. Past employees in this position have often moved into more senior management jobs with the company, although working in senior management entails longer hours. You are married with a two-year old child. [The company you work at does not subsidize the cost of childcare arrangements for employees./The company you work at subsidizes the cost of high-quality, extended-hours childcare for employees.]

ENE Design Text  Imagine yourself in the following scenario.

You work at a company where you have recently won an award for talented junior employees. Now, you have been promoted to a mid-level management position. Past employees in this position have often moved into more senior management jobs with the company, although working in senior management entails longer hours, it comes with a higher salary and more leadership opportunities.

You are also married with a two-year old child. Currently, your child is in day-care for about 40 hours per week. If you moved into senior management, your child would need to be in day-care for at least 50 hours per week.

For the last several years, your firm has been designated by Forbes magazine as one of the “100 best companies to work for” and has now opened an on-site, high-quality, extended-hours day-care center open from 6:00 AM to 10:00 PM on weekdays. The center is free for employees, but slots are allocated via random lottery. [Today you find out that you have not won a day-care slot for your child in the center./Today you find out that you have won a day-care slot for your child in the center.]

Only a subset of female respondents see the next paragraph:

Later, you read a news story reporting that in a nationally-representative survey, more than 50% of college-educated women under age 45 said that the ideal situation for women
with young children is working part-time outside the home, while 30% said not working at all outside the home. Only 10% said that the ideal situation for women with young children is working full-time.

G.2.4 Substantive Outcome Question
If you were in the situation described above, what is the likelihood you would try to advance into a senior management position? Using the slide rule below, position the slide approximately where in the scale you feel your likelihood falls.

[0 to 100 scale; 0 = Highly Unlikely; 100 = Highly Likely]

G.2.5 Placebo Test Questions
A. How likely do you think it is that this company offers employees benefits other than childcare that would be important to you? Using the slide rule below, position the slide approximately where in the scale you feel your likelihood falls.

[0 to 100 Scale; 0 = Highly Unlikely; 100 = Highly Likely]

B. How likely do you think it is that this company helps employees balance work-family issues? Using the slide rule below, position the slide approximately where in the scale you feel your likelihood falls.

[0 to 100 Scale; 0 = Highly Unlikely; 100 = Highly Likely]

C. How likely do you think it is that this company expects employees to answer work-related email on the weekends? Using the slide rule below, position the slide approximately where in the scale you feel your likelihood falls.

[0 to 100 Scale; 0 = Highly Unlikely; 100 = Highly Likely]

G.2.6 Mechanisms
In this space, please tell us briefly why you answered the way you did about your likelihood to try to advance into a senior management position in this company.

[Textbox]

G.2.7 More Demographics Questions
Number of Children How many children (including stepchildren) do you have?

[Drop down menu: 0 to more than 3]

Political Ideology How would you describe yourself politically?

- Very liberal
- Liberal
- Moderate, leaning liberal
- Neither liberal nor conservative
Moderate, leaning conservative
Conservative
Very conservative

**Employment status**

- Employed full-time
- Employed part-time
- Employed part-time and part-time student
- Full-time student
- Independent contractor with varying hours
- Stay-at-home parent
- Unemployed, but looking for work
- Unemployed, not looking for work
- Retired
- Other

**Marital Status**

What is your marital status?

- Married
- Divorced
- Separated
- Single, never married
- Widowed
- Civil union or domestic partnership

---

36 The actual text of the survey was incorrectly written as “Martial Status”.
G.3 Results of Placebo Test Questions

G.3.1 Coding and Analyzing the Placebo Outcomes

Each respondent provided responses to the three placebo test questions. Each non-standardized response $Y_{i,j}^N$ to placebo test question $j$ is along a 100-point likelihood scale. We construct the standardized response $Y_{i,j}$ using the method in the Democratic Peace survey experiment (see E.4.1 for details). As a robustness check, we also analyze the non-standardized placebo outcomes $Y_{i,j}^N$.

Let $Z_i$ be an indicator variable for whether respondent $i$ is told they have subsidized childcare provided by the hypothetical company. For each vignette type, we estimate $E(\tau_{i,j}) = E[Y_{i,j}(Z_i = 1) - Y_{i,j}(Z_i = 0)]$ using $\hat{\beta}_{1,j}$ from the regression $E(Y_{i,j}|Z_i) = \beta_{0,j} + \beta_{1,j}Z_i$. We calculate heteroskedasticity-robust standard errors for our coefficients and present the 95 and 99 percent confidence intervals in our coefficient plots.

G.3.2 Placebo Test Results

In Figure 41, we plot the distribution of the placebo outcomes by treatment assignment and vignette type. The distributions of placebo outcomes are more similar for the treatment and control groups in the ENE design than in the Basic design. In Figure 42, we use coefficient plots to show our estimates of the non-standardized difference-of-means in the placebo outcomes by vignette type. Overall, we find that the imbalance in placebo outcomes is smaller in the ENE design than in the Basic design. Nevertheless, for two of the placebo outcomes in the ENE design, we still detect statistically significant imbalance in placebo outcomes at $\alpha = 0.05$.

Figure 41: Latura (2015): Distribution of Responses to Placebo Test Questions by Treatment Assignment and Vignette Type
Figure 42: Latura (2015): Placebo Test Questions Results (Non-standardized)