Confounding in Survey Experiments

Allan Dafoe[1], Baobao Zhang[1], and Devin Caughey[2]

[1]Department of Political Science, Yale University

[2]Department of Political Science, Massachusetts Institute of Technology

This draft: April 18, 2014*

Early Draft. Comments Welcome.

Figures are not yet compatible with grayscale printing.

Abstract

Inferences from scenario-based survey experiments are often confounded. Scholars randomly manipulate features of a vignette. This is intended to manipulate a specific belief of the respondent, while **not** manipulating other beliefs. When other beliefs are also, unintentionally, manipulated, inferences about the causal factor of interest will be confounded. Survey experimental manipulation of vignettes should thus be understood using the framework of instrumental variables: the manipulation works when it only changes the intended beliefs about the scenario; it fails when it also manipulates other unspecified beliefs (the exclusion restriction is violated). This paper examines this methodological situation in the context of the study of the democratic peace. We recommend *placebo tests* as a diagnostic for confounding in survey experiments. We evaluate two strategies for reducing confounding: controlling for other factors in the vignette and embedding a natural experiment in the scenario. We find that confounding exists and is most severe when neither strategy is employed. Like in observational analysis, controlling reduces confounding on controlled and correlated factors, but not on other factors. Embedded natural experiments, when correctly designed, eliminate confounding on all factors, though the instrument may be weak. Scenario-based survey experiments should be analyzed using the IV framework: scholars should measure manipulation of the causal factor of interest; IV estimation should be used for estimating causal effects; estimated effects are **local** since they depend on all aspects of the vignette including the size and nature of the manipulation of the causal factor of interest.

*For helpful comments, we would like to thank Peter Aronow, David Broockman, Alan Gerber, Donald Green, Sophia Hatz, Josh Kalla, Jason Lyall, Elizabeth Menninga, Cyrus Samii, Sean Zeigler, and participants of the University of North Carolina Research Series and Yale Institution for Social and Policy Studies Experiments Workshop. For support, we acknowledge the MacMillan Institute at Yale University, and the National Science Foundation Graduate Research Fellowship Program.

1 Introduction

Survey experiments involving hypothetical scenarios ("scenario-based survey experiments") play an increasingly important role in political science research. Because they involve random experimental manipulation, it is widely believed that survey experiments allow researchers to overcome the problems of spurious correlation — confounding — that plague observational methods. Contrary to this conventional wisdom, we argue and demonstrate that inferences from scenario-based survey experiments are likely to suffer from confounding, often of a kind similar to that which is present in observational research. The reason is that manipulation of one component of a scenario will generally change subjects' beliefs about other aspects of the scenario. We argue that manipulation of a vignette in a survey should be conceptualized as a potential instrumental variable (Z) for the actual causal factor of interest (D): the beliefs of the respondent about a specific feature of the scenario. Often, as we will show, the manipulation of the vignette fails as an instrument because it violates the exclusion restriction: manipulation of Z changes beliefs about other unspecified features of the scenario (E) that are likely to influence the outcome (Y). We develop our argument by examining this problem in two important research topics: (1) the study of the democratic peace (Mintz and Geva, 1993; Tomz and Weeks, 2013); (2) the study of racial discrimination in preferences about welfare allocation (Desante, 2013).

To address confounding in survey experiments, we introduce and evaluate a set of tools. First, we recommend using placebo tests to diagnose possible confounding in scenario-based survey experiments. Placebo tests consist of questions about components of the scenario that are not consequences of D and therefore should be unaffected by Z (the manipulation of the vignette) if Z is a valid instrument. Second, we summarize and evaluate two possible strategies for addressing confounding: (1) controlling for potential confounds by specifying them in the scenario; (2) embedding a plausible natural experiment in the scenario so that respondents' beliefs about the treatment (D) are uncorrelated with their beliefs about other aspects of the scenario. We discuss subtle issues that arise with these strategies, such as what happens when controls make a vignette unrealistic, and how different embedded natural experiments change the underlying counterfactual.

Section 2 reviews the use of survey experiments in prominent political science publications. Section 3 develops our theoretical foundation. Section 4 then introduces our primary research design for evaluating our theory and tools. We build upon a recent, important, and excellent study of the Democratic Peace that used a scenario-based survey experiment (Tomz and Weeks, 2013); this study examined whether, in the context of an aggressive country developing nuclear weapons, Americans are less supportive of using force against this country if it is a democracy. We evaluate three types of vignettes: a "basic design" in which the possibility of confounding is ignored; a "control design" in which certain potentially confounding features of the scenario are controlled (held constant); an "embedded natural experiment design" in which we embed a plausibly as-if random manipulation of the causal factor. We ask nine placebo questions about characteristics of the target country in the scenario (or about countries "most likely to fit the scenario"), including about the country's region of the world, GDP/capita, military alliance with the US, trade and investment with the US, religious composition, and languages spoken.

Section 5 summarizes our findings. As theorized, we found extensive confounding in the *basic design* in which only the regime type was manipulated in the vignette. Further, this confounding is in the same direction and roughly of the same magnitude as the kinds of confounding that would arise from an analogous observational study. For example, respondents believe scenarios involving

a non-democracy, as opposed to a democracy, are less likely to take place in Europe and more likely to take place in the Middle East.

In our *control design* we control additional features of the scenario, specifically the target country's trade with the US and whether it has a military alliance with the US. We find that controlling for features of the scenario reduced confounding on those features that were controlled for, as well as those that are correlated with the controls in the real world such as foreign direct investment (FDI) in the US. Controlling, however, did not reduce confounding on those features that do not have a strong positive correlation with the controls. Just as with observational studies, controlling for a set of characteristics will tend to reduce confounding on those characteristics, but not necessarily on other characteristics.

Finally, we evaluate our *embedded natural experiment design* in which we embed a plausibly as-if random manipulation of the regime type of the country. In our vignettes, we describe a fragile democracy being held together by a popular president. We then describe an assassination attempt by the military, and manipulate whether the attempt succeeds or not. The natural experiment design passes all but one of the placebo tests, showing that, again as with observational studies, a successful natural experiment provides a means to overcome all forms of unobserved confounding. However, our natural experiment has a relatively weak effect on the regime type, which will amplify any confounding that remains. We also demonstrate that respondent's beliefs need not move as instructed by the vignette; for example in the "democracy" condition of the basic design, reflecting the context that the country is aggressively developing nuclear weapons, more respondents perceive the country to be a "Full Autocracy" than to be a "Full Democracy".

Section 7 concludes by proposing a set of standards to guide designers of survey experiments and readers of studies that use this research methodology. We also discuss a proposal to evaluate our theory and tools in the context of DeSante's 2013 survey experiment, which tests whether Americans have racial preferences when allocating welfare. In his study, DeSante used racialized first names to test if subjects discriminated against African-American applicants. We suggest that using racialized first names (e.g., Keisha versus Emily) may induce subjects to think about the applicant's other characteristics such as background socio-economic status and level of education. We also propose and test alternative strategies to manipulate survey respondents' perception of an applicant's race that reduce confounding.

2 Survey Experiments in Political Science

Survey experiments have become increasingly popular as a research method within political science. In Figure 1, we show the growing use of survey experiments, particularly over the last decade. Although survey experiments have existed since the 1950s, it only became a popular methodology among social scientists in the late 1990s with the rise of computer-assisted telephone interviewing. Proponents of survey experiments often claim that this strategy overcome the problems of traditional, non-experimental surveys. Instead of running regressions with dozens of controls to minimize confounding, researchers purportedly can unravel cause and effect by randomly assigning treatments and observing outcomes (Brady, 2000; Gilens, 2002; Mutz, 2011). While we agree that survey experiments are powerful, and underused, tools for social science, our paper points out that their contribution to causal inference depends in subtle ways on their design.

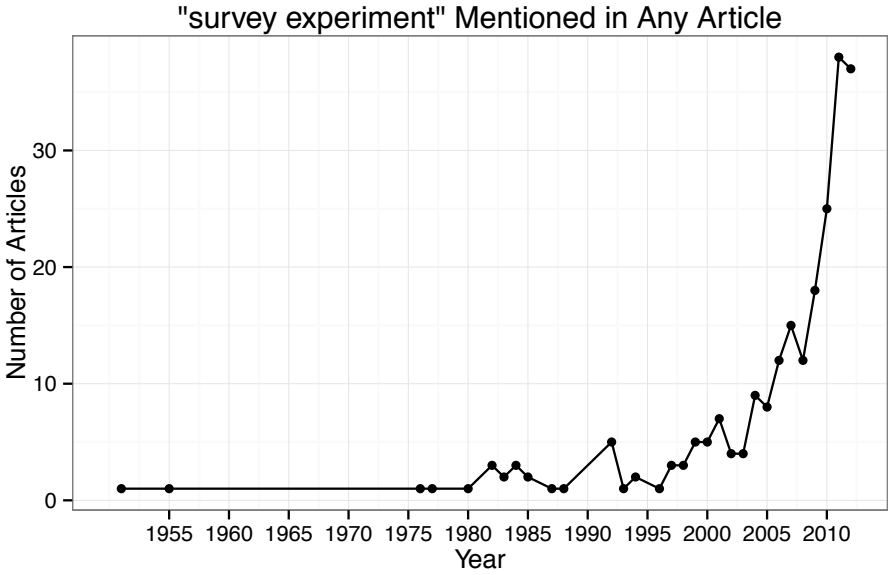


Figure 1: Mentions of "Survey Experiments" in Political Science Journals
The article counts come from searches within political science journals in JSTOR.

Gilens (2002) reviews the different types of survey experiments. The simplest and most common type involves manipulating a hypothetical scenario. We call these *scenario-based survey experiments*. One prominent example is the "welfare mother experiment" on the 1991 Race and Politics Survey conducted by the University of California, Berkeley.[1] Survey subjects were randomly sorted into two experimental groups. Each subject was given a vignette about a hypothetical welfare mother; keeping other features of the mother constant, the researchers told the control group that she is white and the treatment group that she is black. Then subjects were asked to predict the behavior of the welfare mother and other questions about race and welfare. Other examples include surveys that manipulate the race of criminals (Hurwitz and Peffley, 1997), the gender of political candidates (Sanbonmatsu, 2002), the branch of government that made a policy decision (Gibson, Caldeira and Spence, 2005), the race of immigrants (Schildkraut, 2009), and the party that endorses a particular foreign policy (Trager and Vavreck, 2011).

A second type of survey experiments involves changing the question contex; we call these "framing" or "priming" experiments. An often-cited example is Sniderman and Carmine's (1997) "mere mention" study, reported in *Reaching Beyond Race*. Subjects were randomly assigned to two experimental groups. The ones in the control group were asked to rate blacks as a group on a series of traits and stereotypes. The ones in the treatment group were first asked a question about affirmative action and then proceeded to the traits and stereotype questions. Sniderman and Carmine interpret the significant difference between the two groups' responses to mean that merely mentioning affirmative action can "prime" negative perceptions of blacks. Other examples include priming subjects with correct news information (Gilens, 2001), political arguments about immi-

[1]The 1991 Race and Politics Survey was directed by Paul M. Sniderman, Philip E. Tetlock, and Thomas Piazza with support from the National Science Foundation (SES-8508937).

gration (Sniderman, Hagendoorn and Prior, 2004), and messages from political elites (Druckman and Nelson, 2003).

A third type of survey experiment, the list experiment, is used to detect subjects' true sentiments on sensitive topics (Kuklinski 1997). This is not so much an experiment for causal inference as it as measurement technique, since what is randomly assigned are the answer options. For instance, subjects might feel uncomfortable disclosing choices that might appear racist in a survey. In the list experiment, the respondents hear a list of items that they are told "might make people angry or upset." The respondents were told to indicate how many of the items — not which ones — make them angry or upset. In the control version, the subject is given four items. In the treatment version, the subject is given five items; the additional item is the experimental treatment of interest. By comparing the mean number of upsetting items in the baseline condition (4 items) with the treatment condition (5 items), researchers can calculate the percentage of the sample population who express hostility towards the item of interest.

In this paper, we focus on scenario-based survey experiments.[2] Nevertheless, confounding could also exist in the other two types of experiments. For instance, in priming/framing experiments, the stimulus researchers present to subjects might trigger more than one emotion or subconscious thought. In list experiments, the exclusion violation might be violated if adding the fifth item (the treatment item) changes how subjects view the other four items.

2.1 Prior Work on Confounding and Local Effects in Survey Experiments

Some scholarship on survey experiments is aware of some of the risks to inference that we describe in this paper, in particular the possibility that the causal factor of interest may be confounded because of other unintended effects of the manipulation.[4] The work that most explicitly articulates these problems that we are aware of is (Tomz and Weeks, 2013, 5). Tomz and Weeks refer to the problem of confounding as "information leakage", noting that manipulation of the regime-type of the target country may lead respondents to draw inferences about other characteristics of the target country ("leak information" about these other characteristics). In recognition of this threat to inference, many survey experiments employ what we call Controls Designs: they include additional details in the scenario, sometimes experimentally manipulated, to control respondent's beliefs about these potentially confounding characteristics (Bechtel and Scheve, 2013; Desante, 2013; Grieco et al., 2011; Johns and Davies, 2012).

Recent work on "conjoint analysis" (Hainmueller, Hopkins and Yamamoto, 2014) shares themes with this paper in being concerned about improving causal inference in scenario-based survey experiments. However, Hainmueller, Hopkins, and Yamamoto (2014) are concerned with a fundamentally different problem. They confront the problem that some survey experiments manipulate multiple aspects of a vignette, such as a design that varies the ethnicity of a person by altering the immigrant's "face, name, and country of origin" (2). By manipulating multiple aspects of a vignette in a collinear manner, it is not possible to identify the specific effects of each of these

[2]To evaluate the state of the literature, we reviewed all articles that used survey experiments published in the *American Political Science Review* and the *American Journal of Political Science* during the past two decades.[3] In our coding, 53% of these (20/38) involved scenarios, and 34% involved framing (13/38). We include a list of the articles we reviewed in the Appendix.

[4]Our early thinking on this topic was discussed by Cyrus Samii in a March 2011 blog, referring to a conversation about Allan Dafoe's work.

words. Conjoint analysis solves this problem by independently manipulating each relevant feature of a vignette. The problems that we discuss—of confounding and local effects—remains even if one manipulates a *single word* of a vignette, or multiple single words in a factorial design as done in conjoint analysis. We are concerned with how manipulation of a vignette, be the manipulation of a single or multiple words, will change a set of beliefs in addition to the beliefs that the scholar wishes to manipulate.

In various works scholars have also noted what we refer to as the problem of *local effects* in survey experiments. For example, Tomz and Weeks (2013, ft 7) note that an aspect of Johns and Davies (2012) vignette (the government favored air strikes) may have "leaked additional information", changing the causal estimand ("reduced the estimated effect of democracy"). Other works are attentive to the possibility that the causal estimand, and thus the results, will depend on the level of abstraction (Mutz, 2011, 66; Herrmann, Tetlock and Visser, 1999, 566). For instance, previous studies have shown that subjects express greater empathy and support for individuals described in vignettes than for policies that affect populations or sub-populations at large (Cao, 2014; Iyengar, 2012). We contribute to this scholarship by explicitly articulating this problem as involving a change in the causal estimand due to conditioning on other aspects of the scenario or framing; while there is no one "correct" estimand, we argue that there are more and less relevant estimands for various theoretical questions and scholars should be aware of how their design will determine which local effects they estimate.

3 Theory

Researchers use scenario-based survey experiments because they wish to manipulate subjects' beliefs about a specific aspect of a scenario. When the manipulation of the vignette—the actual text describing the scenario—unintentionally manipulates the respondents' beliefs about other features of the scenario, the causal inference may be confounded. The causal factor of interest (the respondent's beliefs about a specific feature of the scenario) may be correlated with other unmeasured causes (respondent's beliefs about other features of the scenario) of the outcome.[5] This section shows why scenario-based survey experiments should be conceptualized as relying on instrumental variables (IV) for causal inference. The causal inference is sound when (1) the IV is itself unconfounded, (2) the IV affects the outcome (Y) only through the causal factor of interest (the exclusion restriction is satisfied), and (3) the IV is believed to have monotonic effects on the causal factor of interest (D). The causal inference will be confounded when the exclusion restriction is violated so that respondent's beliefs about other unspecified but relevant characteristics are also manipulated by the manipulation of the vignette. We then formally introduce models of how sub-

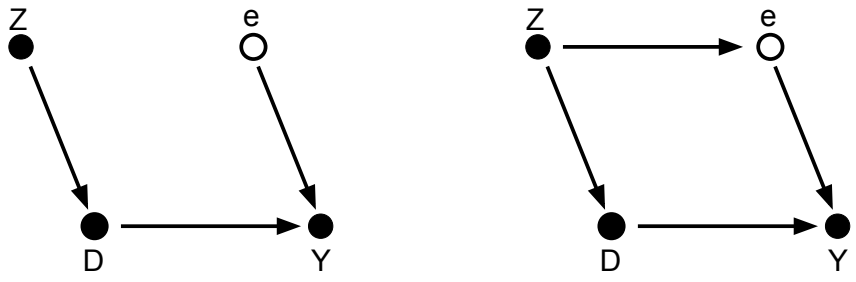
[5]There are a number of different methodological vocabularies one can use to characterize this problem. In the study of field experiments and medical trials (Gerber and Green, 2012; Hernan and VanderWeele, 2011) methodologists worry about how compound (or bundled) treatments can complicate causal inference; in the case of survey experiments the manipulation of a vignette or framing may be a compound treatment, changing the respondent's beliefs and feelings along multiple dimensions. In the study of instrumental variables methodologists worry about the *exclusion restriction*, which states that the instrument did not affect the outcome except through the causal factor of interest. While we occasionally use the language of compound treatments and IV estimation, we prefer to frame this problem as one of confounding: there is a common cause (manipulation of the survey) of the causal factor of interest and other causes of the outcome. We prefer this language because it draws attention to the near perfect mapping to the methodological problem of confounding in the analogous observational studies.

jects form and revise their beliefs, which allows us to articulate *ex ante* the kinds of confounding we could expect from different designs. Finally, we connect these theoretical concepts with the study of the democratic peace.

3.1 Design problems and design-based solutions

Scholars manipulate features of vignettes (denote as Z) so as to draw inferences about the effects of changes in respondents' beliefs about a feature of the real world (denote as D). For example, in studies of the democratic peace scholars might want to investigate the effects of the regime type of a target country on the U.S. public's support for using force against that country. To do so, a researcher would manipulate whether a target country in a scenario is *described as a democracy or non-democracy* (Z). He/she would then ask the respondent about their support for using force in such a scenario (Y). The researcher would like to infer that the observed effects (the change in Y) correspond to the *effects of regime type* (D) on public support for using force. However, this inference only holds if Z is a valid instrument for D. Z is a valid instrument if it has a monotonic moderately sized effect on D and does not affect other determinants (e) of the outcome Y (Morgan and Winship, 2007). Figure 2 represents this issue using non-parametric causal graphs.

Figure 2: Vignettes as Instruments



● represents an observed random variable and ○ represents a latent random variable. A ⟶ B means A causes B. The absence of an arrow implies the absence of causation.

Manipulations of vignettes should be viewed as instruments (Z) for beliefs about the causal factor to which they refer (D). To be a valid instrument for D, Z must only affect Y through D (left figure). Inferences about D using Z will be confounded if Z changes other determinants (e) of the outcome (right figure).

We can also conceptualize this issue using the notation of linear models. Write out the determinants of Y as (where all variables are $n \times 1$ vectors):

$$Y = \alpha_1 + \delta D + e$$

and the determinants of D as:

$$D = \alpha_2 + \gamma Z + u$$

Then the IV estimator for δ is:

$$\hat{\delta}_{IV} = \frac{Cov(Y,Z)}{Cov(D,Z)} = \frac{\delta Cov(D,Z) + Cov(e,Z)}{Cov(D,Z)} \tag{1}$$

$\hat{\delta}_{IV}$ provides a good estimate of δ, the effect of D, to the extent that

1. $Cov(e,Z)$ is small or zero. This can be broken up into two conditions. (a) The exclusion restriction: Z only affects Y through D. (b) No fundamental confounding: except for the causal effect of Z on Y, Z is otherwise independent of Y. In correctly implemented randomized survey experiment we expect (b) to be true, since characteristics of the respondent should be independent of the experimental conditions received. If both (a) and (b) are true we would say that there is no confounding with the causal factor of interest.

2. $Cov(D,Z)$ is not small; the instrument is not weak. This condition can be evaluated empirically.

3. γ, the effect of Z on D, is constant. If it is heterogeneous, but monotonic, then our IV estimate will be a weighted average based on these unit-specific effects. If it is not monotonic then IV estimation requires other assumptions.

We argue that it is often the case in scenario-based survey experiments that condition (1a) is false: $Cov(e,Z) \neq 0$, so that inferences about D will be confounded. Our research will evaluate the existence and size of $Cov(e,Z)$, as well as strategies for reducing $Cov(e,Z)$.

Analyzing scenario-based survey experiments using an IV framework reveals other important considerations analysts should keep in mind. First, scholars usually estimate their quantity of interest using some function of the covariance of Y and Z, such as through regression or difference in means: $E[Y|Z=1] - E[Y|Z=0]$. This estimation strategy corresponds to an intention-to-treat effect (ITT) which, given the IV conditions, will give us the sign of the effect of D. However, this estimation strategy does not give us the magnitude of the effect of D, which is often of interest. To estimate the magnitude of the effect of D analysts must take into account the covariance of Z and D using an IV estimator, such as through two-stage least squares or the Wald estimator. This implies that survey experiments must measure the respondents' beliefs about the causal factor of interest, in our case how democratic they perceive the target country to be. We refer to this as a *manipulation measure*.

Second, except for the unlikely case of constant effects, analysis of instrumental variables yields estimates of *local* causal effects in which different units and parts of the response surface receive varying weight; units and areas of the response surface that experience more variation in D induced by Z receive more weight in IV estimation (Imbens and Angrist, 1994).[6] Thus, conceptualizing scenario-based survey experiments using an IV framework properly directs our attention to understanding the kinds of variation in D induced by our Z, and the kinds of individuals who most respond to Z. Specifically, in our study of the democratic peace we find that our intervention (Z) only moves perception of regime type (D) by, on average, a few Polity points, often around the middle of the scale. Thus, the ITT estimates reported by most studies of the democratic peace are

[6]For that matter, all causal estimates are weighted averages of unit-specific effects, except when we implausibly assume otherwise (Aronow and Samii, 2013).

probably substantial underestimates of the true effect on public opinion of whether a target country is a democracy or autocracy.

3.2 Model of Respondents' Beliefs

What should we believe about the kinds of confounding present in any particular scenario-based survey experiment? We introduce here three models about $Cov(e,Z)$ which can be articulated prior to performing a study and yield precise testable implications. These models are (1) the No Confounding Null Model in which $Cov(e,Z) = 0$; (2) the Roughly Realistic Bayesian Model in which respondents' beliefs are approximately what a Bayesian with realistic beliefs about the world would draw; (3) the Ignorant Bayesian Model in which belief updating is based upon Bayesian inference, but beliefs about the world are not realistic. Each of these models assumes (1b) that there is no fundamental confounding, which is true if the experiment was properly administered.

1. **No Confounding Null Model**: The manipulation of the vignette *only* manipulates the intended beliefs of the respondent. This is the model implicitly assumed by most analyses of scenario-based survey experiments.

2. **Roughly Realistic Bayesian Model**: Respondents have a model of the distribution of characteristics of a scenario based on the actual distribution of the characteristics of similar scenarios in the real world. For instance, when told that a country suffers from malaria, subjects think this country has a high probability of being in the tropics, since in the real world proportion of countries suffering from malaria is much greater for countries in the tropics than those not. Further, respondents should revise their beliefs approximately according to the laws of conditional probability (Bayesian updating). Respondents will use characteristics specified in a vignette to condition their beliefs about unspecified characteristics, as well as the meaning of the other words employed in the vignette. So long as we have information about the actual covariance of characteristics in the real world, this model yields precise implications about how respondents will revise their beliefs. [7]

3. **Ignorant Bayesian Model**: This model is like the Roughly Realistic Bayesian Model, except we allow the respondents to hold prior beliefs that are not consistent with the real world. Respondents may draw their beliefs from inaccurate stereotypes from the media and other sources. For instance, Americans inaccurately believe that the US government gives much more foreign aid, per US citizen, than it in fact does. If a scenario were to specify the level of US foreign aid, a US respondent could draw inferences about other aspects of the scenario that are unrealistic. In order to test this model, the character of these "ignorant" beliefs need to specified. Scholars could first ask respondents for their beliefs about relevant features of the world in order to establish the baseline beliefs. In our work so far we do not invoke or test the Ignorant Bayesian Model, but we mention it as a possible alternative.

[7] 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 rational beliefs and do in fact revise according to conditional probability. For example, humans often over- and under- weight very low probabilities, believe the probability of a scenario increases as restrictive details are added, and do not give enough consideration to alternative hypotheses (Bowers and Davis, 2012). 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.

We conjecture that the Roughly Realistic Bayesian Model better accounts for respondents' beliefs than the No Confounding Null. Specifically, this means that describing a country in a scenario as a "democracy", vs "non-democracy" will lead respondents to believe that this country is more likely to have other characteristics correlated with democracies in the real world, such as being in Europe, having liberal values and norms, being wealthier, being more economically interdependent, and sharing strategic interests with the US.

3.3 Design-based solutions

To detect the problem of confounding in survey experiments, we propose the use of placebo tests. Placebo tests are tests of known (usually zero) effects (Rosenbaum, 2002, Ch 6), used to evaluate a design and estimator (Dafoe and Tunón, 2014; Sekhon, 2009). Specifically, our placebo tests are survey questions that measure whether the experimentally manipulated features of the vignette (Z) affected subjects' beliefs in unintended ways. For instance, in the democratic peace survey experiment, we use placebo tests to evaluate whether describing the aggressor country as a dictatorship makes subjects more likely to think the aggressor country is poor or located in the Middle East.

Some scholars are implicitly aware of confounding in survey experiments and adopt what we call a *control design* in which potential confounds are explicitly mentioned in the vignette and thereby held constant. We argue that a control design will have similar effects as conditioning strategies in observational studies: they will reduce or eliminate confounding on the variables specified, and usually reduce confounding on other characteristics that are correlated with the controls. However, as with observational studies, a control design will not address confounding on characteristics not correlated with the controls, and can even induce confounding.

We examined Tomz and Weeks' *control design* in which *military alliance* and *trade with the US* are held constant. As expected, imbalance on these variables becomes closer to zero and insignificant. In addition, imbalance on the similar variable of *Foreign Direct Investment in the US* also becomes close to zero. However, all other potential confounds that we examined (*region, GDP per capita, religion, English speaking*) remained significantly imbalanced across treatment levels and by the same magnitude (except *English speaking*). Controlling reduces confounding on the variables controlled for, and on correlated variables, but not on other variables.

We innovate by suggesting a new strategy for overcoming confounding in scenario-based survey experiments: basing the hypothetical scenario on a natural experiment, an as-if random manipulation of the causal factor of interest.[8] Just as natural experiments in the real world allow observational studies to identify plausibly as-if random variation in the causal factor of interest, so we conjecture that scenarios based on plausible natural experiments will induce respondents to believe that the variation in the causal factor of interest is as-if random in the scenario: independent of all other features of the scenario that are not a consequence of the causal factor of interest. Our natural experiment involves a narrative about a country that remains a democracy largely because of the influence of the president; an assassination attempt then either just succeeds or just fails, leading to the as-if random persistence or overthrow of the democracy. As we conjecture, the *natural experiment design* exhibits the least confounding. Further, it has similar balance on *military alliance* and *trade with the US* as the *control design*, even though it does not explicitly control for these factors. One concern with the embedded natural experiment design is that it may produce a

[8] In a future draft we will formally deduce this conjecture.

weak instrument. In our study the change in the democracy level of the aggressor country induced by the natural experiment was relatively weak, and barely significant. Since IV bias increases with the weakness of the instrument, this means we should be careful that our embedded natural experiments induce sufficient variation in subjects' beliefs about the causal factor of interest.

We discuss a few prominent survey experiments that might have experience the confounding problem. Hainmueller and Hiscox (2010) evaluate whether Americans oppose immigration because they fear economic competition. Half of the respondents were asked of their attitudes towards "low-skilled immigrants" and the other half were asked of their attitudes towards "highly skilled immigrants." Aside from changing subjects' beliefs about the immigrants' skill-level, the vignettes might have affected subjects' beliefs about the immigrants' race or ethnicity. The reason is that in the U.S., a large majority of low-skilled immigrants come from Latin America; likewise, a majority of highly skilled immigrants come from Asia.

Survey experiment vignettes using racial names to study racial discrimination also might be confounded. For example, DeSante (2013) manipulates the name of workers applying for government assistance, and evaluates the effect of evaluations of the worker's work ethics and the worker's implied race on support for assistance. However, in addition to changing beliefs about race, changing the worker's name from Laurie to Latoya could change the respondent's beliefs about other characteristics of the worker, such as her level of education, her criminal background, her work history, and the socioeconomic class of her parents.

Some survey experiments, by necessity, change two (or more) features of the vignette simultaneously. This bundles separate treatments, making it impossible to estimate their independent effects. For instance, Cohen (2003) manipulates the political party endorsing a policy position, as well as the justification for the policy position. This simultaneous change is necessary to have a realistic counterfactual. Nevertheless, because the endorsing party and the policy justification are collinear it is not possible to separate their effects.

4 Research Design

4.1 Studies of the Democratic Peace

Increasing numbers of IR scholars have turned to scenario-based survey experiments to test important theories, such as audience cost (Tomz, 2007), the legitimizing role of international organizations (Grieco et al., 2011), and rationalist explanations for opposing immigration (Hainmueller and Hiscox, 2010). For our project, we use a survey experiment to evaluate the democratic peace. One of the most robust and widely noted empirical associations in international relations is the democratic peace: democracies are much less likely to threaten and use force against each other. Studies of the democratic peace, however, are almost entirely based on the observed behavior of countries. As such, causal inference about the democratic peace faces the deep challenge of disentangling the many possible causal mechanisms and potential confounds. The democratic peace could be caused by a variety of mechanisms: greater respect for the policies of democratically elected governments, greater sensitivity to the costs of war, information revelation through public debate and a free press,electoral incentives for leaders to uphold their commitments, and greater caution in selecting disputes and greater competence in fighting them. Similarly, scholars have suggested that the democratic peace could be caused by a variety of confounding factors: shared foreign policy

preferences especially during the Cold War, economic interdependence, stable borders, or reverse causation (democracy flourishes in peaceful regions).

There exist several important survey experiments on the democratic peace. Mintz and Geva's (1993) and Rousseau's (2005) studies show that Americans express greater support for going to war against dictatorships than democracies. More recent studies have controlled for other aspects of the opponent country. Johns and Davies (2012) test whether subjects would respond differently to democracies versus autocracies, as well as to the majority religion of the opponent country (Christian versus Muslim). In several large-*N* survey experiments, Tomz and Weeks found that subjects are more likely to support military strikes against autocracies than democracies, even after controlling for the aggressor country's military capabilities and alliances (Tomz and Weeks, 2013) and that this effect seems to be mediated by respect for human rights (Tomz and Weeks, 2012) rather than procedural aspects of democracy. Further, Tomz and Weeks (2013) found that in vignettes involving democracies (versus non-democracies), respondents had similar expectations of the costs of conflict and the probability of failure, but decreased perceptions of threat and increased perceptions of the immorality of a US attack. This provides highly informative evidence about the possible mechanisms of the democratic peace.

These survey experiments have contributed valuable insight about the democratic peace. Survey experiments on public opinion provide a crucial new empirical domain for evaluating theories of the democratic peace and they provide a level of control that enables careful designs that tease apart the many possible mechanisms of and alternatives to the democratic peace. We expect that similarly careful and creative survey experimental designs will provide further progress in the study of this important question.

Our contribution to this literature is to evaluate the conditions under which and extent to which these scenario-based survey experiments provide evidence of the causal effect of the causal factor of interest—the regime-type of the country in the scenario—as opposed to the possible effects of other characteristics of the country in the scenario or features of the scenario. As we will show, we as yet cannot rule out that the possibility that the US public's aversion to using force against a country described as democratic is actually because of beliefs about other features of the country, such as its liberal norms, the religion and culture of its citizens, its history of conflict with the West and willingness to be a "responsible" global citizen, the orientation of its economy, or other factors correlated—in the real world and in the minds of respondents—with its regime-type. Among other research strategies, future survey experiments that are sensitive to these challenges have much promise of furthering our understanding of this important phenomenon.

4.2 Our Survey Experiments

We conduct two survey experiments that build and expand upon Tomz and Weeks's (2013) survey experiment about the democratic peace. The original survey experiment tests whether American and British respondents show greater support for attacking a non-democratic aggressor country than a democratic aggressor country. We evaluate confounding using placebo tests. Furthermore, we measure the manipulation of the causal factor of interest: to what extent do the different designs change subjects' perception of the aggressor country's level of democracy. We also measured the main outcome (i.e., support for military action against the aggressor) and mediating variables of the Tomz and Weeks survey experiment.

Table 1: Differences Between the Two Surveys

Survey 1: January Wave	Survey 2: March Wave
3 vignette types: Basic, Controls, Embedded Natural Experiment 1	5 vignette types: Basic, Controls, Embedded Natural Experiment 1, 2, & 3
8 placebo test questions	10 placebo test questions
unidimensional scale manipulation measure	multidimensional manipulation measure
no outcome measures	outcome measures

We conducted these survey experiments using Amazon.com's Mechanical Turk. The two surveys test the same hypotheses and have similar questions, although the second one contains more questions. We refer to the first survey as the *January Wave* and the second survey as the *March Wave*. Differences between the two surveys are presented in Table 1.

4.3 Vignette Types

In both the January Wave and March Wave, we used three types of vignettes to measure the amount of confounding associated with each type. The three types are Basic, Controls, and Embedded Natural Experiment. In the January Wave, we only tested one version of the Embedded Natural Experiment. In the March Wave, we tested three versions.

Respondents taking the survey are randomly assigned to one vignette from the three types. All vignettes describe an aggressor country determined to manufacture nuclear weapons despite international condemnation; the respondents are also told that if the aggressor country acquires nuclear weapons, it can use them to attack any country in the world. In the experimental part of the vignettes, we vary characteristics of the aggressor country in the following ways.

For the first type, the *Basic Vignettes*, we manipulate only the regime type of the aggressor country. The aggressor country is either a democracy or not a democracy. The subjects assigned to read about the democratic aggressor are in the treatment group; those assigned the non-democratic aggressor are in the control group. This delineation between treatment and control follows for the other vignette types.

The second type, the *Controls Vignettes*, vary the regime type of the aggressor country, whether the country had signed a military alliance with the U.S., and whether the country has low or high levels of trade with the U.S. The Controls Vignettes have a $2 \times 2 \times 2$ experimental design. This is the vignette type used in Tomz and Weeks's 2013 survey experiment; we based our design on the survey wording in (Tomz and Weeks, 2013).

In addition, we introduce a new vignette type, the *Embedded Natural Experiment Vignettes*. This design allows us to make the treatment (e.g., the regime type of the country) seem to the respondent to arise in the scenario in an as-if random manner. For this type of vignette, subjects read about the recent history of the aggressor country, called Country A, which is depicted as a fragile democracy, headed by a popular president. Subjects are also informed that a well-researched U.S. governmental report predicts there is a high chance the country would become a military dictatorship if the president were to die. In the January Wave, we only tested one type of the Embedded Natural Experiment Vignettes (Type 1). In the March Wave, we tested three types (Type 1, Type 2,

Table 2: Differences in Information Provided by the Three Type of Embedded Natural Experiment Vignettes.
As denotes information about assassination attempt; *Re* denotes information about regime-type.

Type 1: manipulate *As*, *Re*	Type 2: manipulate *Re*	Type 3: manipulate *As*
details about the aggressor country	details about the aggressor country	details about the aggressor country
outcome of the assassination attempt	NO mention of the assassination attempt	outcome of the assassination attempt
regime change	regime change	NO mention of regime change

Type 3). Figure 2 outlines the main differences between the three types of the Embedded Natural Experiment Vignettes.

In Embedded Natural Experiment Type 1, for the non-democracy scenario, a disgruntled military officer shot the president of Country A, hitting him in the head, leading to the president's death and a military takeover of the government. For the democracy scenario, a disgruntled military officer shot the president of Country A, hitting him in the shoulder; the president survived the attack and the democracy was preserved.[9]

We also wanted to evaluate whether we can construct an embedded natural experiment by presenting subjects with the details of the aggressor country without the assassination attempt. For Embedded Natural Experiment Type 2, we presented the aggressor country's background information. Subjects in the treatment condition were told the country remained a democracy while subjects in the control condition were told the country became a military dictatorship. We withheld any specific information about why the country did or did not experience regime change.

Although Embedded Natural Experiment Type 1 may seem reasonable, it is not the correct way to exploit an embedded natural experiment. The reason is that in it we manipulated both the as-if random event and the regime type of the country; while manipulation in the former should be independent of all other features of the scenario (that are not consequences of the assassination attempt), manipulation in the latter will suffer from similar confounding as in the Basic and Control designs. This double manipulation would only not be a problem if the specification about regime type was uninformative, so that respondents believe that the effect of the assassination attempt was deterministic: the probability of the country being "democracy" given the failed assassination attempt is 1, and 0 for a successful assassination.

In Embedded Natural Experiment Type 3 we describe the outcome of the assassination attempt,

[9]Jones and Olken's 2009 study using successful and unsuccessful assassinations as a natural experiment inspired our vignettes. The exact wording of our vignettes are as follows: "Five years ago, 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 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 shoulder/in the head] and [survived/did not survive] the attack. [Country A's democratically elected government survived the political turmoil, and is still a democracy today./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.]"

but we provide no information about the country's subsequent regime type. In this way the only feature of the scenario that we vary is a feature that is plausibly believed to be as-if random in the scenario. The challenge for this design is to induce sufficient variation in the causal factor of interest.

4.4 Placebo Tests

After reading a vignette, each subject answers questions that serve as our placebo tests. A placebo test is defined as a test of an association that should be zero if the design is working as it should be. These questions ask subjects about their perceptions of characteristics of the scenario that should not be a consequence of the causal factor of interest, such as geography and religion.[10] If there is a significant difference between the responses from the treatment group and the responses from the control group about these placebo characteristics then we have evidence of confounding.

In the January Wave, we conducted 8 placebo tests (Placebo Tests A-H). In the March Wave, we included two new placebo tests (Placebo Tests I and J), in addition to those in the January Wave.

The first two placebo tests ask about the geographic location of countries likely to experience the scenario described. In *Placebo Test A*, we ask subjects to list real countries likely to fit the scenario they have read. In *Placebo Test B*, subjects are given 11 regions of the world. They are asked to pick the two most plausible regions and the two least plausible regions to experience the scenario. *Placebo Test C* is an open-ended question: subjects are told to write down at least five characteristics of countries likely to fit the scenario they have read.

Placebo Tests D through H are multiple choice questions. *Placebo Test D* asks respondents to select the GDP per capita range of countries likely to experience the scenario. Since the Controls Vignettes provided details about the aggressor's military alliance (or lack of) with the U.S. and trade levels with the U.S., we ask about these two factors in *Placebo Tests D* and *E*, respectively. Finally, we are interested in how respondents perceive the culture of the aggressor country. Thus, *Placebo Test G* asks subjects to think about countries that fit the scenario and how likely it is that these countries' populations are mostly Christian. For *Placebo Test H*, respondents select the percentage of people in these countries who are English speaking.[11]

Finally, we are interested to determine whether the Controls Vignettes reduced confounding in other characteristics correlated with military alliance and trade. *Placebo Test I* asks subjects how likely they think the aggressor country had fought alongside the U.S. in the Iraq War. *Placebo Test J* asks subjects what level of direct investment the aggressor country has in U.S. businesses.

For each of the placebo tests, we have the following hypotheses:

- H_0: Responses from the treatment and control group come from the same underlying distri-

[10] We adopted two ways of phrasing this question. Under the more intuitive phrasing, we ask about characteristics of the country in the scenario; however, this question has an absurd quality to it: a respondent will be asked by the creators of the scenario to describe aspects of the country that have not been specified in the scenario. Under our other phrasing, we ask respondents about characteristics of countries "likely to experience the scenario described" because implies a well defined question about the real world for which there is, in principle, a correct answer. We do not detect a difference in responses according to these two ways of phrasing the placebo questions.

[11] In the January Wave, respondents were asked how likely it is that an average person in the aggressor country can speak English. Because the question is difficult to interpret, we do not analyze the results from that version of the question in this paper.

bution.[12]

- $H_{a,1}$: Responses from the treatment and control group do not come from the same underlying distribution.
- $H_{a,2}$: The differences in the treatment and control groups' answers are similar to differences between the characteristics of democracies and non-democracies in the real world that are likely to experience events such as the scenario.

4.5 Manipulation Measure and Plausibility Measure

In addition to the Placebo Test questions, we included a *manipulation measure* and a *plausibility measure* in our survey. We performed a manipulation measure to examine whether the different vignette types induced different changes in the causal factor of interest: how democratic was the aggressor. For the January Wave manipulation measure, subjects were shown a 21-point Polity scale with 11 reference countries above the scale. Using a slider, each subject gives the aggressor country what amounts to a Polity score. For the March Wave manipulation measure, subjects were given a series of categories ("Full Autocracy", "Semi-Autocracy" ...) with example countries next to each; we then imputed a Polity value for each of these categories (see Section 5.5).

We also wanted to measure how plausible subjects find the scenario. We asked them to rate how likely they think it is that such a scenario would occur in the next 10 years. The multiple choice answers we provide include qualitative descriptions with percentages attached to them.[13] Researchers interpreting observational studies worry about extrapolating outside the convex hull of the covariates because such inference is more model dependent King and Zeng (2006). Likewise, we worry that assigning subjects to read about implausible scenarios might cause them to become confused, angry, or problematically conscious of the survey's artificiality.

4.6 Outcome Measures

In the March Wave, we asked about the same outcomes as used in (Tomz and Weeks, 2013). For our main outcome measure, we asked whether respondents would favor or oppose using the U.S. military to attack the aggressor's nuclear development sites. In addition, we asked the multiple choice mediation questions Tomz and Weeks asked to measure subjects' perceptions of threat, cost, success, and morality. To gauge perceptions of threat, we asked which of the following events had more than a 50 percent chance of happening if the U.S. *did not* attack:

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

[12] In our hypothesis tests using parametric models, we use the null of no difference between the mean responses of the treatment and control groups. In our hypothesis tests conducted using randomization inference, we use the sharp null of no difference between all subjects' potential outcomes under treatment and control.

[13] We used words of estimative probability used by the Central Intelligence Agency. See Kent 1964.

- The country would would launch a nuclear attack against the U.S. or a U.S. ally

To assess the perceptions of costs and success, we asked which, if any, of the following events would have more than a 50 percent chance of happening if the U.S. *did* attack:

- The country would respond by attacking the U.S. or a U.S. ally
- The U.S. military would suffer many casualties
- The U.S. economy would suffer
- U.S. relations with other countries would suffer
- The U.S. would prevent the country from making nuclear weapons in the short run.
- The U.S. would prevent the country from making nuclear weapons in the long run.

For the two questions above, respondents could select as many events as they thought probable or none of them. Finally, to evaluate perceptions of morality, we asked subjects whether it would be "morally wrong for the U.S. military to attack the country's nuclear development sites."

4.7 Measurement Uncertainty Principle and Question Order

One of the concerns we had when conducting the March Wave was how to order the questions. Psychologists have long known that one can use previous questions to prime subjects to consciously or subconsciously think of topics that influence their answers to later questions Harley (2013). A prominent example of such a priming experiment is Peffley, Hurwitz and Sniderman's (1997) experiment on sterotypes and whites' attitudes towards blacks. The researchers prime their subjects to think about stereotypes of African-Americans using a series of questions before the outcome measures. In contrast, we do not want the order of our questions to produce a priming effect.

The priming problem poses a conundrum similar to the Heisenberg Uncertainly Principle: our desire to measure confounding might affect how subjects answer the outcome measure questions. Therefore, in the March Wave, we randomize the order in which three blocks of questions appear. The three blocks include 1) the placebo tests and the plausibility measure, 2) the manipulation measure, and 3) the outcome measures. The randomization would allow us to test whether viewing the placebo test questions affects subjects' response to the outcome measures and vise versa. Furthermore, we randomized the order of the outcome measure questions because we wanted to test whether there is a difference between asking subjects the mediation questions before the main outcome question or after the main outcome question.

5 Evaluating the Survey Experiments

We conducted our first survey (the January Wave) through Amazon.com's Mechanical Turk in January 2014. A total of 554 respondents completed the survey. Again, "treatment" refers to vignettes in which the aggressor is a democracy, "control" refers to the corresponding vignette in which the aggressor is not a democracy. We conducted our second survey (the March Wave)

through Amazon.com's Mechanical Turk in March 2014. A total of 671 respondets completed the survey. In both waves we tested for whether respondent characteristics were imbalanced across treatment conditions; we found no evidence that they were, implying successful randomization (see Appendix).

Because our data have many moving parts, we discuss our approach to analysis. First, we combine data from the January Wave and the March Wave whenever the questions from each wave follow the same format. These include Placebo Tests A, D, E, F, and G. When analyzing data from the combined surveys, we included fixed effect for survey wave in regression analysis and block for survey wave in permutation tests. Secondly, in the March Wave, we tested whether question order had a priming effect before deciding to include fixed effects for regressions or blocks for permutation tests. We discovered that question order did not have an effect on answers to the placebo tests, the manipulation measure, or the outcome measure questions. We also noted that it does not make a difference to ask subjects to think about the aggressor country or to think about countries that are likely to experience the scenario. Therefore, we elected not to control for question order or question phrasing in our analysis in the March Wave.

5.1 Summary of Results

We present summary of our findings in Table 3, which contains the results from our hypothesis tests using randomization inference. We tested whether subjects in treatment and control groups gave different answers to each of the placebo tests across the five vignette types.

Our results show that the Basic Vignettes exhibit the most confounding, as we predicted. Subjects not only thought the democratic aggressor and non-democratic aggressor were different in terms of regime type but also in terms of geographic location and other characteristics that might impact their decision to attack the aggressor. Respondents thought the democratic aggressor is more likely to be located in East Asia, particularly North Korea, than the non-democratic aggressor. Furthermore, subjects thought the democratic aggressor is wealthier, more likely to be an American ally, has higher levels of trade with the U.S., more likely to be mostly Christian, has a higher percentage of English speakers, more likely to have fought alongside the U.S. in the Iraq War, and have invested more in U.S. businesses when compared with the non-democratic aggressor. In short, the Basic Vignettes failed to pass eight of the nine placebo tests we analyzed.[14]

The Controls Vignettes performed better than the Basic Vignettes in that it passed five of the nine placebo tests. Not surprisingly, subjects given the Controls Vignettes thought the democratic aggressor is equally likely to have signed a military alliance with the U.S. and to have a similar level of trade with the U.S. when compared with the non-democratic country. The Control Vignettes controlled for these variables by providing additional details about military alliance and trade with the U.S. Furthermore, we did not detect overall confounding along geographic dimensions. Nevertheless, the Controls Vignettes failed the other four tests; the direction of confounding is similar — although sometimes bigger in magnitude — to the confounding in the Basic Vignettes. Finally, the Embedded Natural Experiment Vignettes were the most successful design for reducing confounding. Out of the three sets of nine placebo tests, we were only able to reject the null hypothesis of no confounding once. While confounding was diminished in the Embedded Natu-

[14]We had a total of ten placebo tests. For this paper, we did not analyze answers to the open-ended question that asked respondents to write five characteristics of countries likely to fit the scenario.

Table 3: Summary Results from Hypothesis Tests

Placebo Test	Basic	Controls	Emb. Natural Exp. 1	Emb. Natural Exp. 2	Emb. Natural Exp. 3
A: Plausible Countries	○	○	○	○	○
B: Most Plausible Regions	●	○	○	○	○
D: GDP per Capita	●	●	○	○	○
E: Likelihood of Military Alliance with the U.S.	●	○	○	○	○
F: Level of Trade with the U.S.	●	○	○	○	○
G: Likelihood of Being Mostly Christian	●	●	○	○	○
H: Percentage English Mpeaking	●	●	○	○	○
I: Likelihood of Being Iraq War Ally	●	●	○	○	●
J: Level of Investment in the U.S. Businesses	●	○	○	○	○

○ means we fail to reject H_0 of no confounding. ● means we reject H_0. We conduct one-sided hypothesis tests at $\alpha = 0.05$ using non-parametric combination test (using the Fisher combining function) for Tests A and B and uni-variate randomization inference for Tests D-J. Recall in Emb. Natural Exp. 1, we manipulate both the assassination attempt outcome (Z) and the regime type (D). In Emb. Natural Exp. 2, we manipulate only the regime type (D). In Emb. Nautrla Exp. 3, we manipulate only the assassination attempt outcome (Z).

ral Experiment Vignettes, the change in perceptions of democracy was also much smaller. This implies that these designs may suffer from bias-amplification due to having weak instruments.

5.2 Placebo Tests A and B: Geographic Location of the Aggressor

The first two of our placebo tests are about respondents' perceived locations of the aggressor country. Geographic confounding was most severe in the Basic Vignettes, less severe in the Controls Vignettes, and least but somewhat present in the Embedded Natural Experiment Vignettes. Most notably, subjects given the Basic Vignettes and the Controls Vignettes associated the non-democratic aggressor with Iran and North Korea.

Placebo Test A is an open-ended question that asks subjects to list countries they think are likely to experience the vignette. For our analysis, we select the top 10 countries mentioned by the subjects. For each country, we compare what proportion of treated subjects mentioned the country and what proportion of control subjects mentioned the country. These differences-in-proportions, along with 95 percent confidence intervals, are shown in Figure 2.[15] The Basic Vignettes exhibit much confounding: there was a significant difference between treatment and control for six out of 10 countries. The Controls Vignettes performed better in that there exists imbalance for three of the 10 countries. The Embedded Natural Experiment Vignettes Type 3 performed the best: we could not detect significant differences between treatment and control in any country. We detected confounding for one and two countries out of 10 in Types 2 and 3, respectively.

We also investigate whether subjects think the democratic aggressor is more likely to be in certain regions of the world compared with the non-democratic aggressor (Placebo Test B). Once

[15]The 95 percent confidence intervals are estimated using OLS with robust standard errors. The countries in Figure 2 are listed (top to bottom) in descending Polity IV score.

again, the Basic Vignette performed poorly and we rejected the null hypothesis of no confounding using nonparametric combination tests (NPC). Subjects given the Basic Vignettes thought the democratic aggressor is more likely to be in East Asia and less likely to be in North America. Subjects given the Controls Vignettes thought the democratic aggressor is more likely to be in the Middle East and less likely to be in Western Europe. Within the Embedded Natural Experiment Vignettes, Type 1 and 2 did not exhibit confounding for any region, but subjects assigned to Type 3 indicted that the democracy is less likely to be in South Asia and East Asia.

Using nonparametric combination (NPC) tests, we attempt to detect aggregate confounding in (A) countries mentioned or (B) regions deemed most plausible. We report one-sided p-values using the Fisher combining function in Table 4.[16] Our results, reported in Table 4, suggests there seems to be no aggregate confounding in Placebo A for all vignette types.[17]

In Figure 3, we present the real distribution of democracies and autocracies using a world map displaying each country's Polity score represented by colors. The greener the color, the more democratic the country is; the redder the color, the more autocratic the country is. The Middle East and East Asia, unsurprisingly, contain higher percentages of non-democracies than Western Europe and North America. Furthermore, these non-democracies in the Middle East (e.g., Iran, Egypt) and East Asia (e.g., North Korea, China) are salient to most Americans. The confounding we observe in the Basic Vignettes and the Controls Vignettes appears similar to the difference in geographic location of real democracies and non-democracies. This similarity is consistent with our Roughly Realistic Bayesian Model of belief updating.

Table 4: NPC p-values for Placebo Tests A and B

A: Plausible countries	global p-values
Basic	0.185
Controls	0.863
Natural Experiment 1	0.604
Natural Experiment 2	0.682
Natural Experiment 3	0.105
B: Most plausible regions	**global p-values**
Basic	0.001
Controls	0.289
Natural Experiment 1	0.518
Natural Experiment 2	0.301
Natural Experiment 3	0.235

The global p-values are produced using the Fisher combining function.

[16]For hypothesis tests where there are multiple outcomes (Placebo Tests A and B), we use non-parametric combination tests, as discussed in Caughey, Dafoe and Seawright 2013.

[17]Despite the imbalance on several of the countries for the Basic Vignettes, we fail to reject the sharp null of no confounding. We cannot attribute this finding to our blocking on survey waves. Without blocking, the p-values for the Basic Vignettes are 0.222, 1.000,

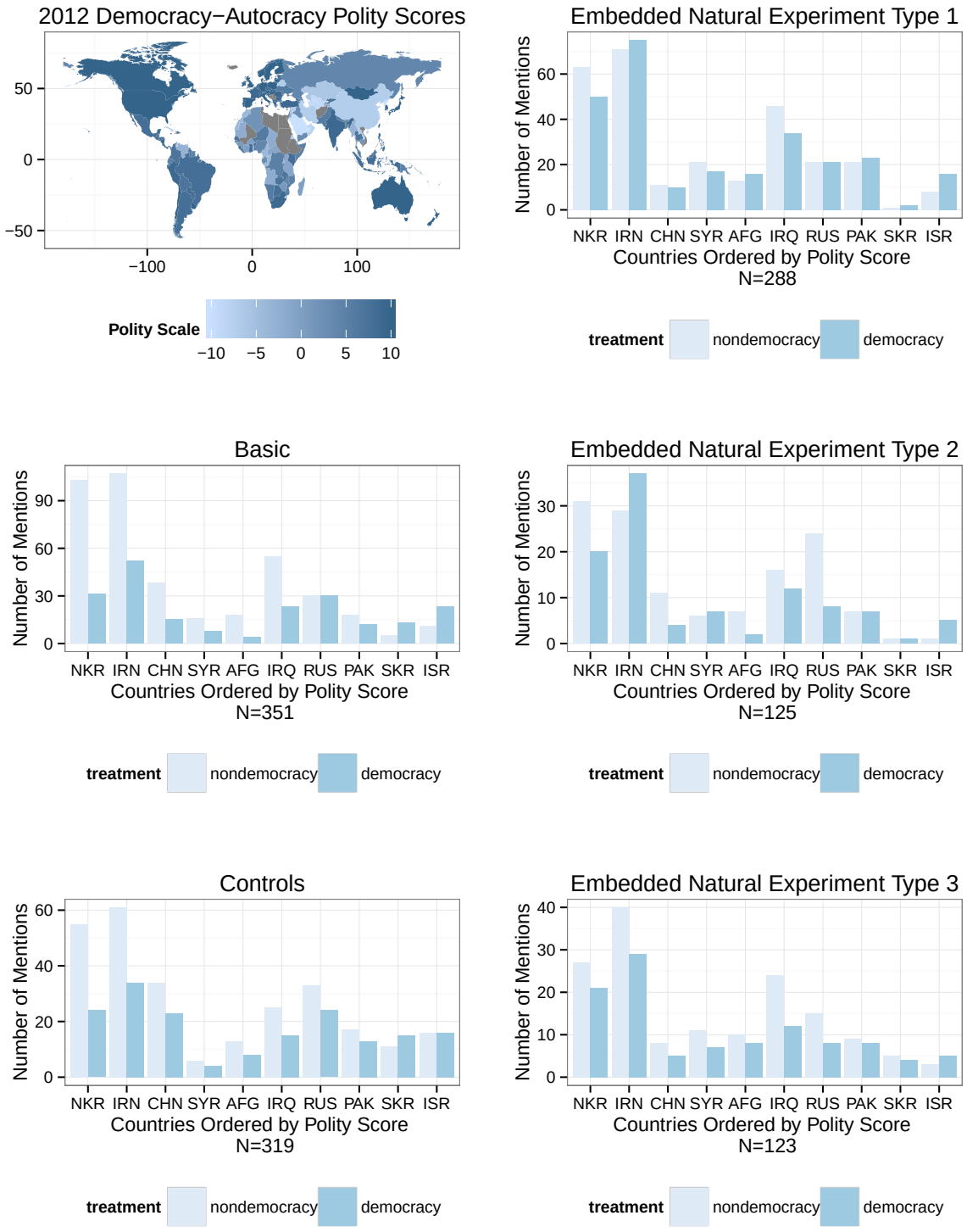


Figure 3: Placebo Test A: Countries Mentioned in Open-Ended Response
The 10 countries above are the most frequently mentioned by all subjects in response to "Please list some countries that you think are most likely to fit the scenario." Country abbreviations: NKR=North Korea, IRN=Iran, CHN=China, SYR=Syria, AFG=Afghanistan, IRQ=Iraq, RUS=Russia, SKR=South Korea, ISR=Israel

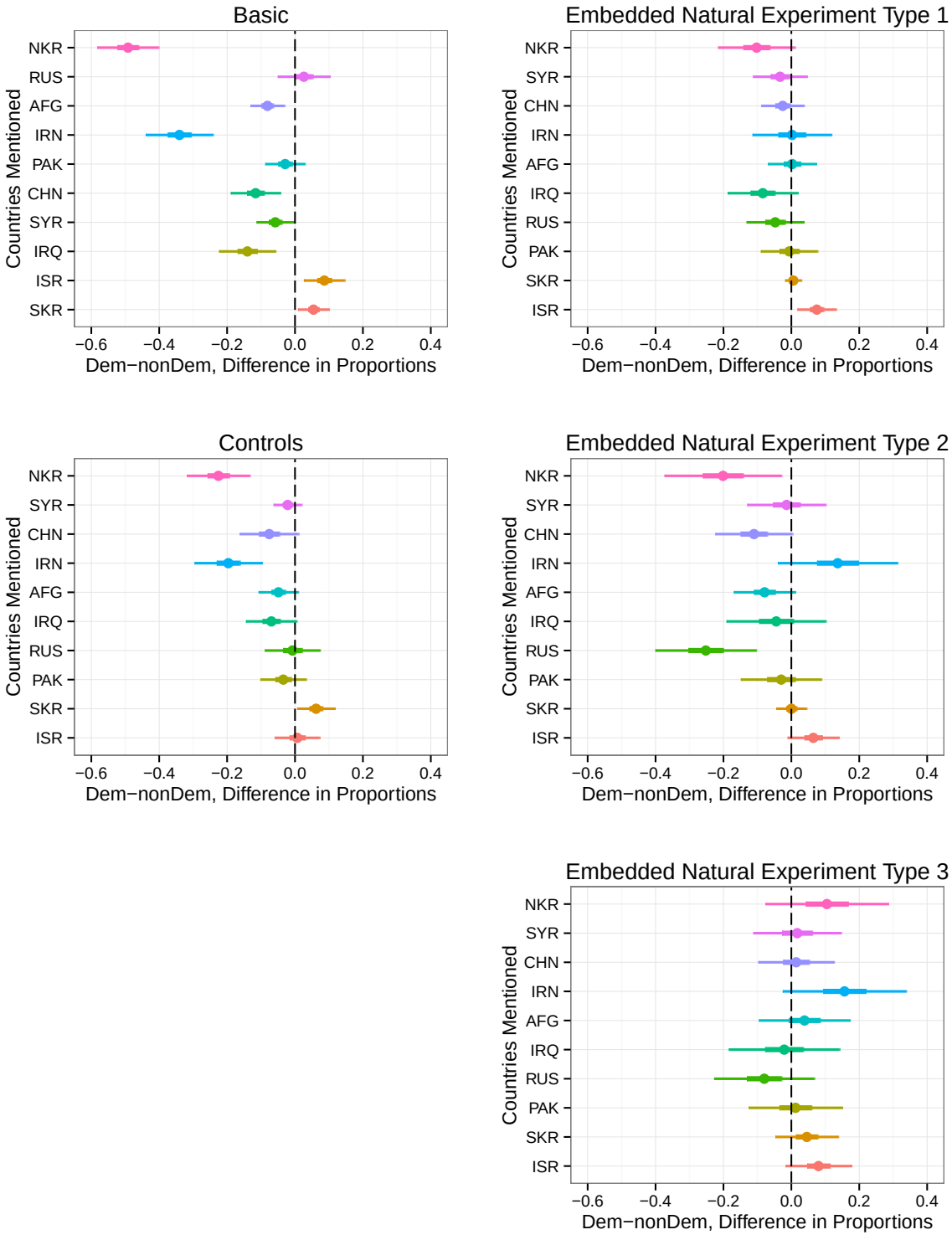


Figure 4: Placebo Test A: Difference in Proportions Who Mentioned Each Country
Dependent variable: proportions of subjects who mentioned each country.

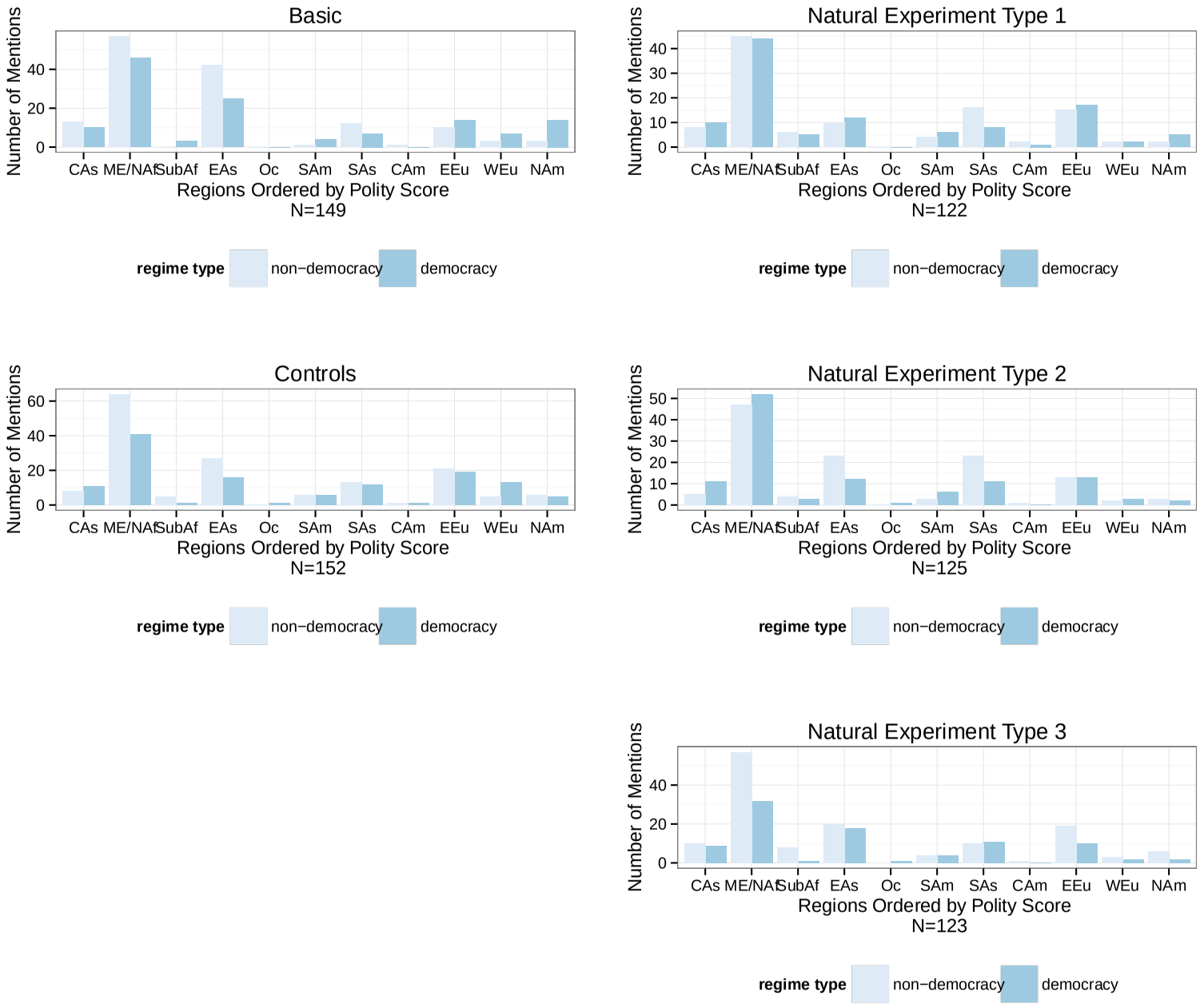


Figure 5: Placebo Test B: Ranking of Likely Regions

Dependent variable: rank of each region, in response to "Please select the two regions most likely to experience the scenario." Region abbreviations: CAs=Central Asia; ME/NAf = Middle East and North Africa, SubAf=Subsaharan Africa, EAs = East Asia, Oc = Oceania, SAm=South America, SAs=South Asia, CAm=Central America, EEu=Eastern Europe, WEu=Western Europe, NAm=North America

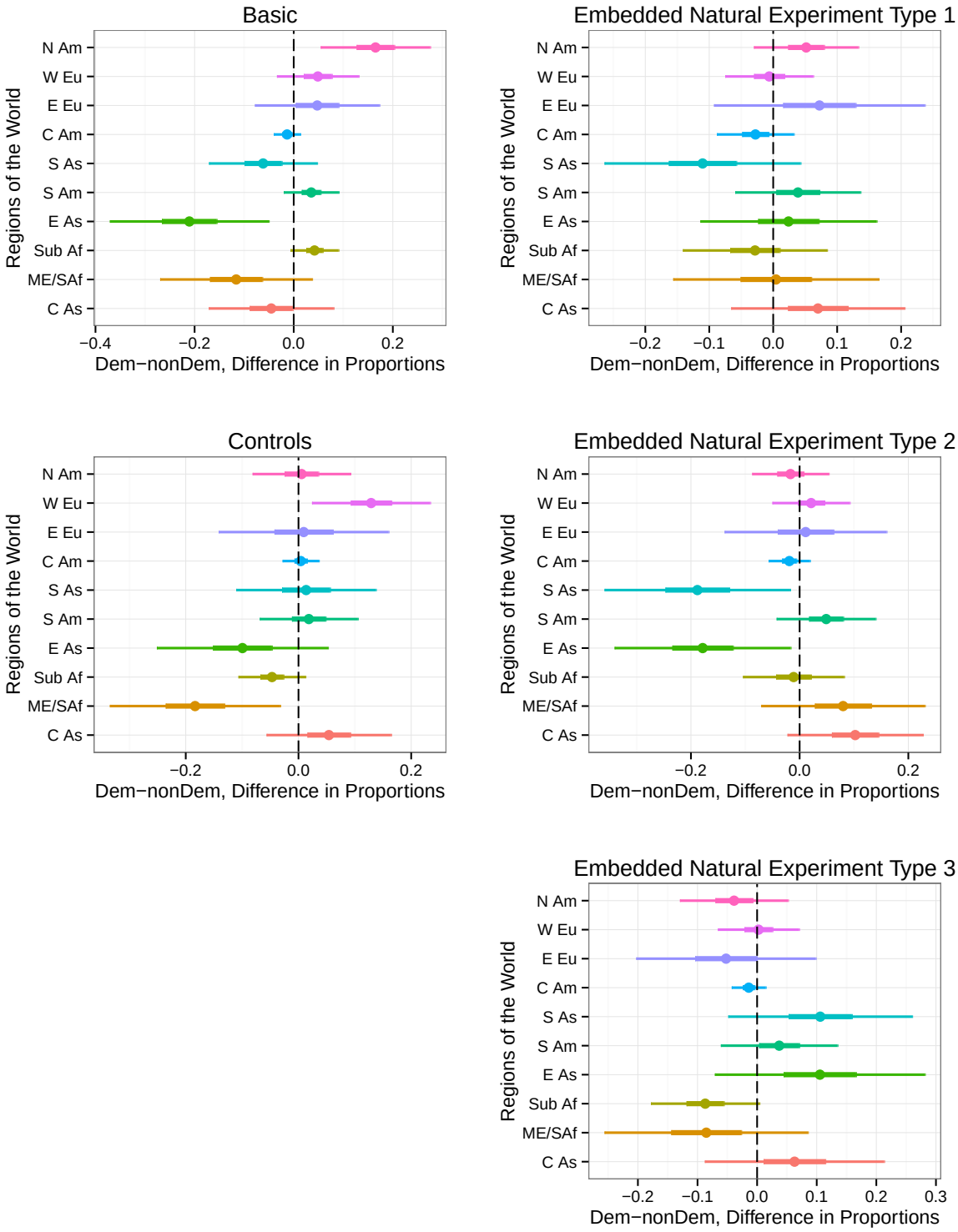


Figure 6: Placebo B: Difference in Mean Likelihood Ranking
Dependent variable: difference in mean likelihood ranking for each of the 11 regions.

5.3 Placebo Tests D, G, H: Characteristics of the Aggressor Country Not Controlled For

For the univariate placebo tests, we began with those that ask about characteristics of the aggressor country not controlled for in the Controls Vignettes. These placebo tests asks subjects to estimate the aggressor country's GDP per capita (Placebo Test D), likelihood of its being mostly Christian (Placebo Test G), and the percentage of its population that speaks English (Placebo Test H). We find that the Basic Vignettes and the Controls Vignettes, but not the Embedded Natural Experiment Vignettes, exhibit confounding in all three placebo tests. This suggests that the Controls Vignettes are ineffective in controlling for characteristics not highly correlated with the controls.

Only the Embedded Natural Experiment Vignettes passed Placebo Test D, which examines whether subjects think the democratic aggressor is wealthier than the non-democratic aggressor. In contrast, subjects given the Basic Vignettes perceived the democratic aggressor to be more than $7,000 wealthier in GDP per capita than the non-democratic aggressor. Subjects given the Controls Vignettes perceived a difference of around $5,800. These differences in treated subjects and control subjects' perceptions are large and significant. In reality, the average difference between democracies and non-democracies' GDP per capita is $8987, not too far from the differences in wealth that our respondents inferred. This similarity, again, supports the Roughly Realistic Bayesian Model of how subjects update their beliefs.

It would be difficult to compare subjects' answers to Placebo Tests G and H with real world distributions because the metric of our placebo questions does not map naturally on to real world data. Nevertheless, we observe that the direction of confounding in the Basic and Controls Vignettes can be predicted by real world data. For instance, Democracies have a higher average percentage of the population that is Christian and a higher average percentage of that is English speaking. In Placebo Test G, subjects given the Basic Vignettes and the Controls Vignettes thought the democratic aggressor is 28 percent and 32 percent — respectively — more likely to be mostly Christian when compared with the non-democratic aggressor. In Placebo Test H, subjects given the Basic Vignettes and the Controls Vignettes thought the percentages of English speakers in the democratic aggressor are 11 points and 7 points higher — respectively — than in the non-democratic aggressor. This evidence further suggests that subjects have some understanding of real world distributions and are using them to fill in the unspecified aspects of the scenarios they read.

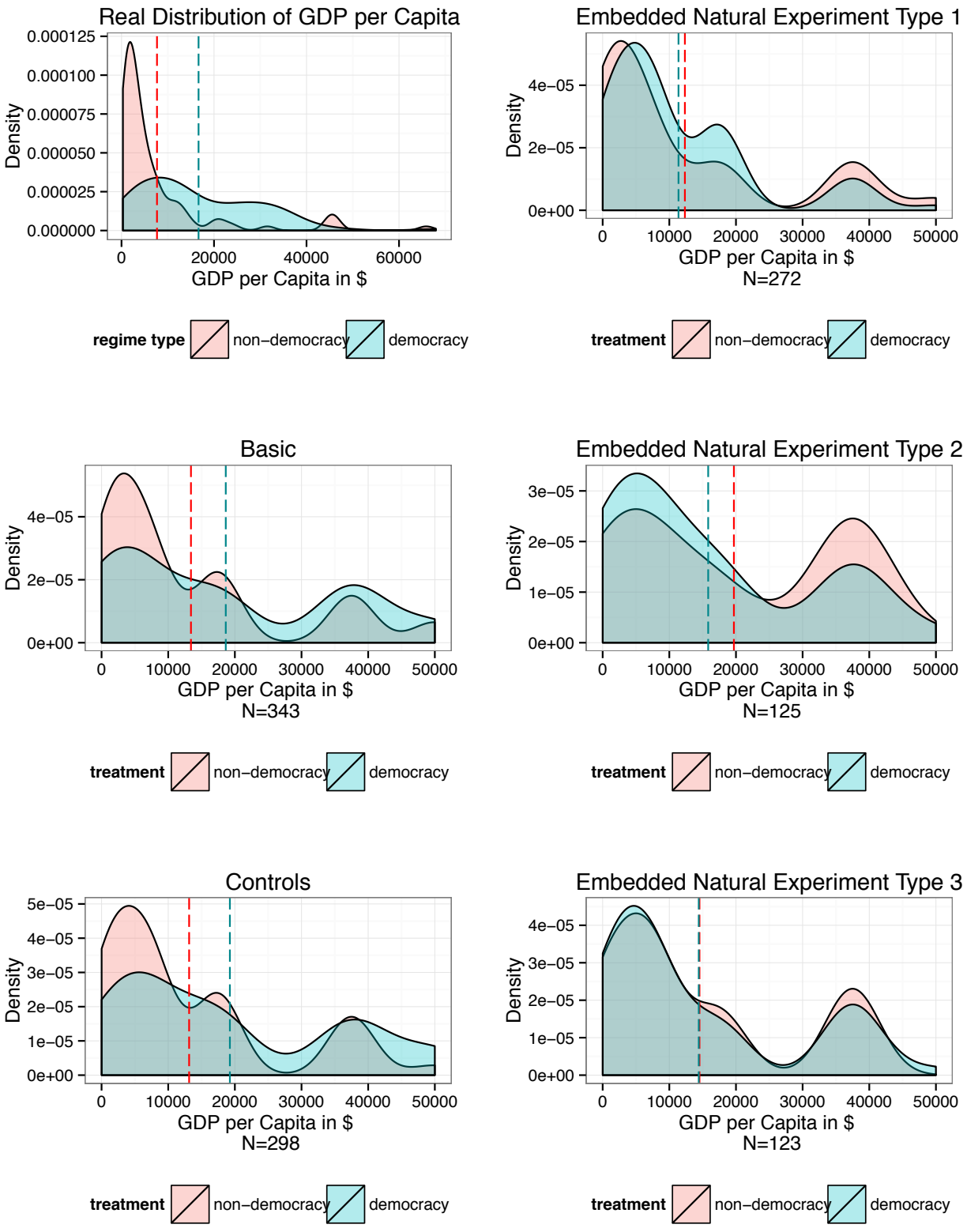


Figure 7: Placebo D: GDP per Capita
Includes data from the January Wave and the March Wave for the Basic, Controls, and Embedded Natural Experiment Type 1 Vignettes.

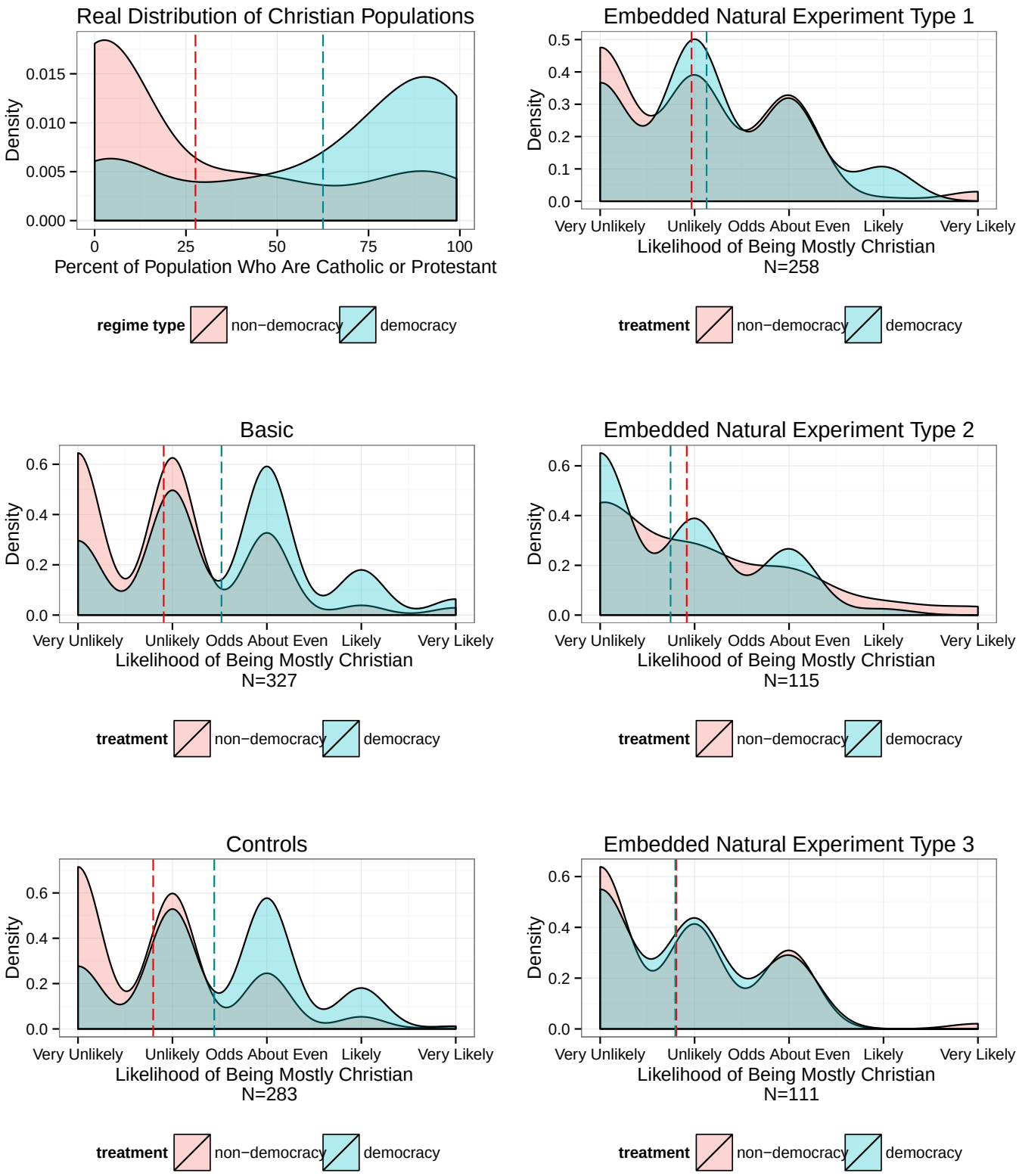


Figure 8: Placebo G: Likelihood of Being Mostly Christian
Includes data from the January Wave and the March Wave for the Basic, Controls, and Embedded Natural Experiment Type 1 Vignettes.

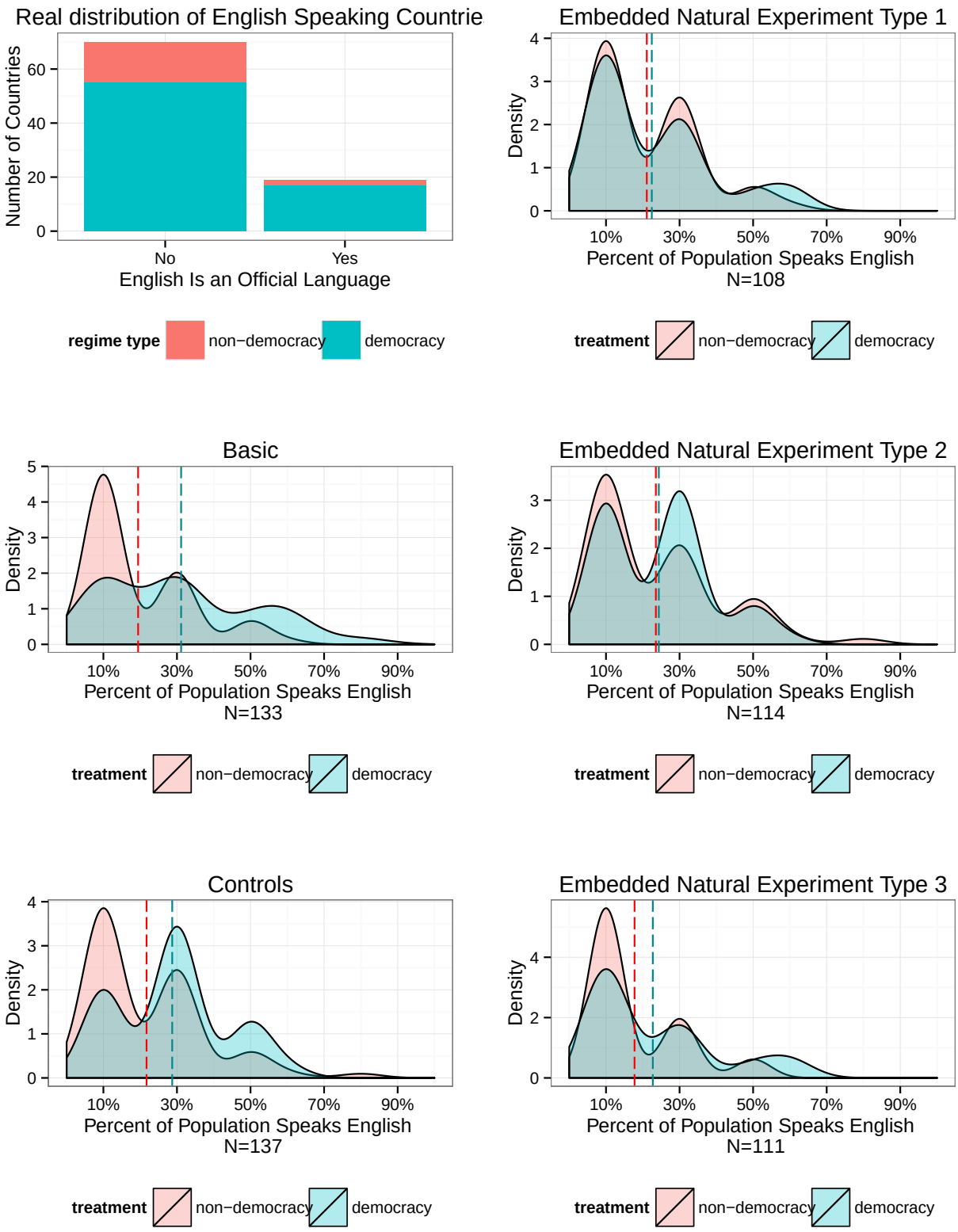


Figure 9: Placebo H: Likelihood of Being English Speaking

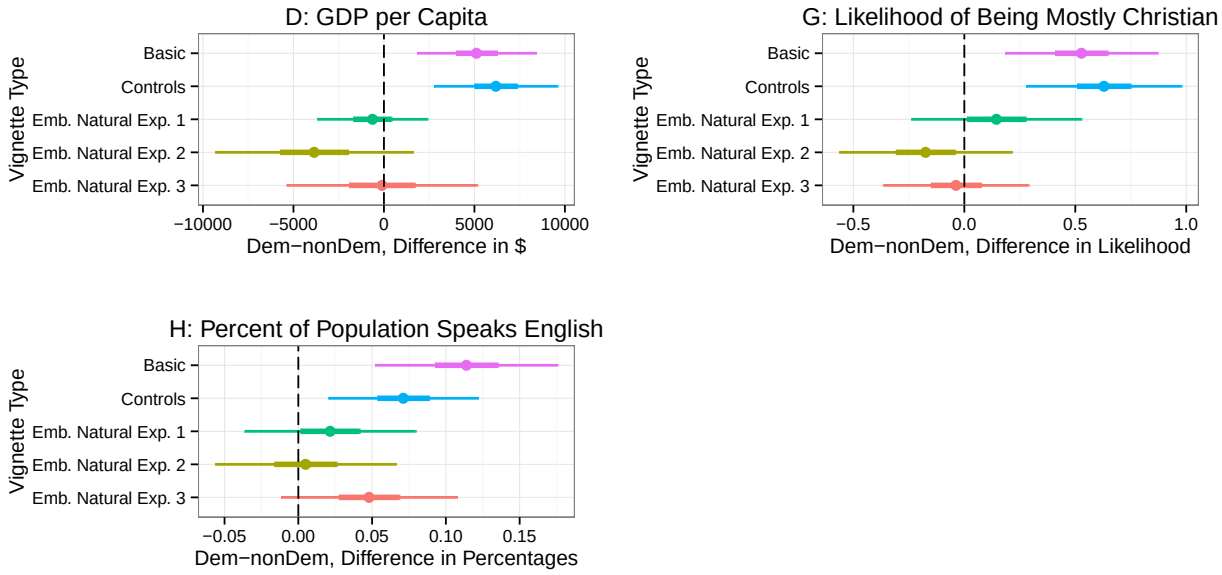


Figure 10: Coefficient Plot: Placebo Tests D, G, H

Table 5: *p*-values from Randomization Inference: Placebo Tests D, G, H

D: GDP per Capita	One-sided p-value
Basic	0.001
Controls	<0.001
Embedded Natural Experiment 1	0.662
Embedded Natural Experiment 2	0.661
Embedded Natural Experiment 3	0.270
G: Likelihood of Being Mostly Christian	**One-sided p-value**
Basic	<0.001
Controls	<0.001
Embedded Natural Experiment 1	0.103
Embedded Natural Experiment 2	0.848
Embedded Natural Experiment 3	0.572
H: Percentage English speaking	**One-sided p-value**
Basic	<0.001
Controls	0.003
Embedded Natural Experiment 1	0.342
Embedded Natural Experiment 2	0.409
Embedded Natural Experiment 3	0.053

We conducted permutation inference using difference-in-means as our test statistic. In Placebo Tests D and G, we block on survey waves for the Basic, Controls, and Embedded Natural Experiment Type 1.

5.4 Placebo Tests E, F, I, J: Characteristics of the Aggressor Country Controlled For

Next, we consider the univariate placebo tests that ask about characteristics of the aggressor country that were explicitly controlled for in the Controls Vignettes—military alliance with the US (Placebo Test E) trade with the US (Placebo Test F)—as well as characteristics highly correlated with those controls. We were also interested to learn whether controlling for military alliance and trade implicitly controlled for other characteristics correlated with these two variables. The two characteristics we chose are the likelihood of the aggressor country being a U.S. ally in the Iraq War (Placebo Test I) and the level of direct investment it has in U.S. businesses (Placebo Test J). In the world world, the correlation between having signed a military alliance with the U.S. and being a member of Multi-National Force–Iraq is 0.42. The correlation between the volume of trade with the U.S. and foreign direct investment in the U.S. is 0.41.

We discovered that the Controls Vignettes were very good at reducing confounding along the military alliance (p-value 0.07) and trade dimensions (p-value 0.16). Furthermore, subjects given the Controls Vignettes thought the democratic aggressor invested about the same in U.S. businesses as the non-democratic aggressors. However, despite the high real world correlation, the Controls Vignettes did not reduce imbalance on Likelihood of Being Iraq War Ally (Placebo Test I). One conjecture for why this control failed to reduce imbalance on correlated variable is that the Bush administration justified the Iraq War by claiming that Iraq, a non-democracy, was secretly building nuclear weapons. Therefore, respondents may have thought it unlikely that a non-democracy engaged in nuclear proliferation, much like Iraq, would ally with the U.S. in that conflict.

The results from the Controls Vignettes are consistent with our expectation that including additional details in the vignettes are unlikely to reduce confounding on variables not closely correlated with the controls. Although the Embedded Natural Experiment Vignettes, particularly Type 1, did not explicitly control for military alliance or trade, they exhibited almost no confounding in the four placebo tests. Like the use of natural experiments in observational studies, if respondents perceive a characteristic of a scenario to be as-if random, then variation of this characteristic should be independent of all other factors that are not consequences of this characteristic, leading to balance (in expectation) on “pre-treatment" covariates without the use of controls. However, despite the seemingly greater balance on all possible confounds, it remains unclear which type of Vignette would actually have less bias from confounding since the IV bias also depends on the strength of the instrument. We now turn to investigating the strength of the manipulations of the causal factor of interest in these different vignette types.

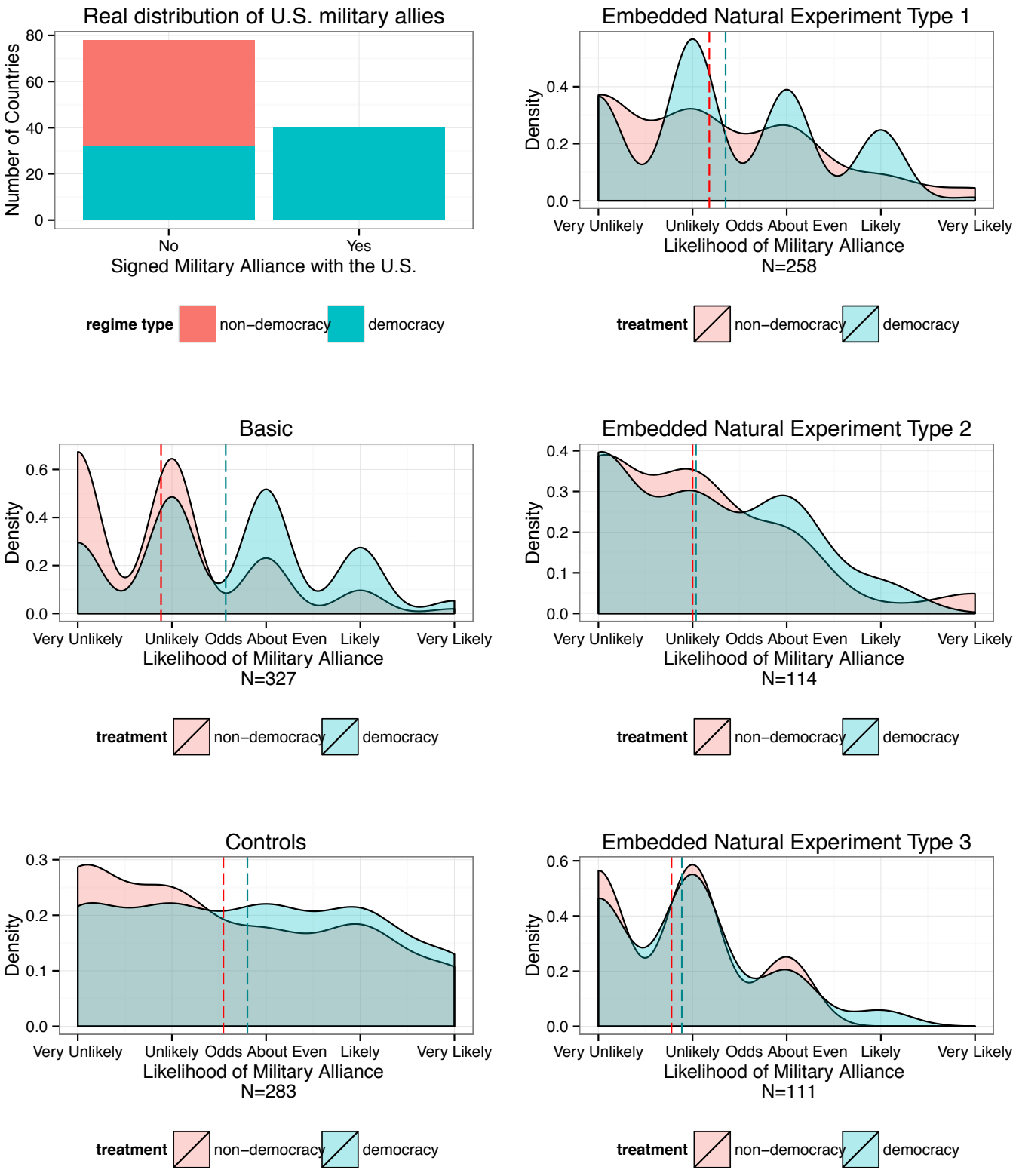


Figure 11: Placebo E: Likelihood of Military Alliance with the U.S

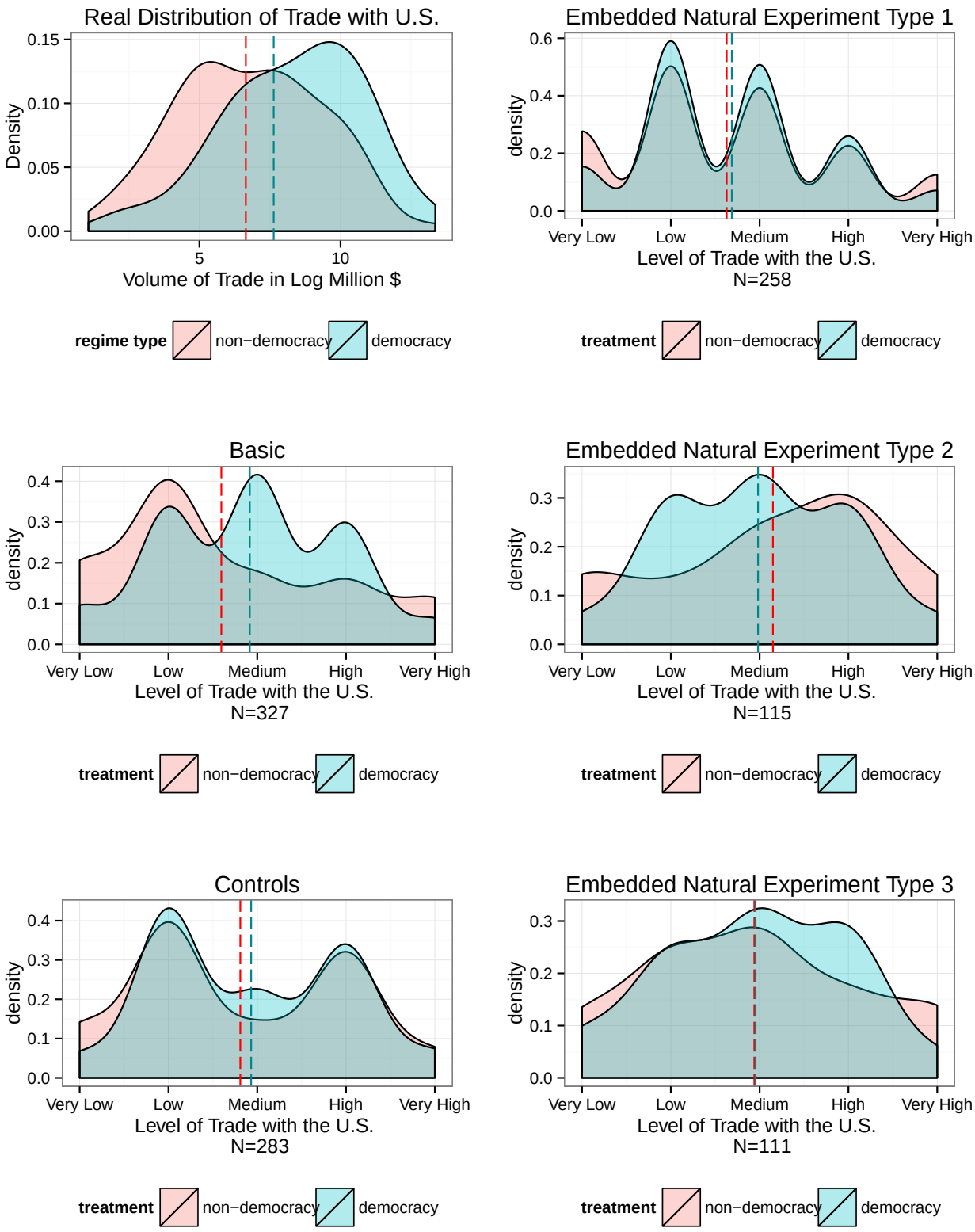


Figure 12: Placebo F: Trade with the U.S.
Includes data from the January Wave and the March Wave for the Basic, Controls, and Embedded Natural Experiment Type 1 Vignettes.

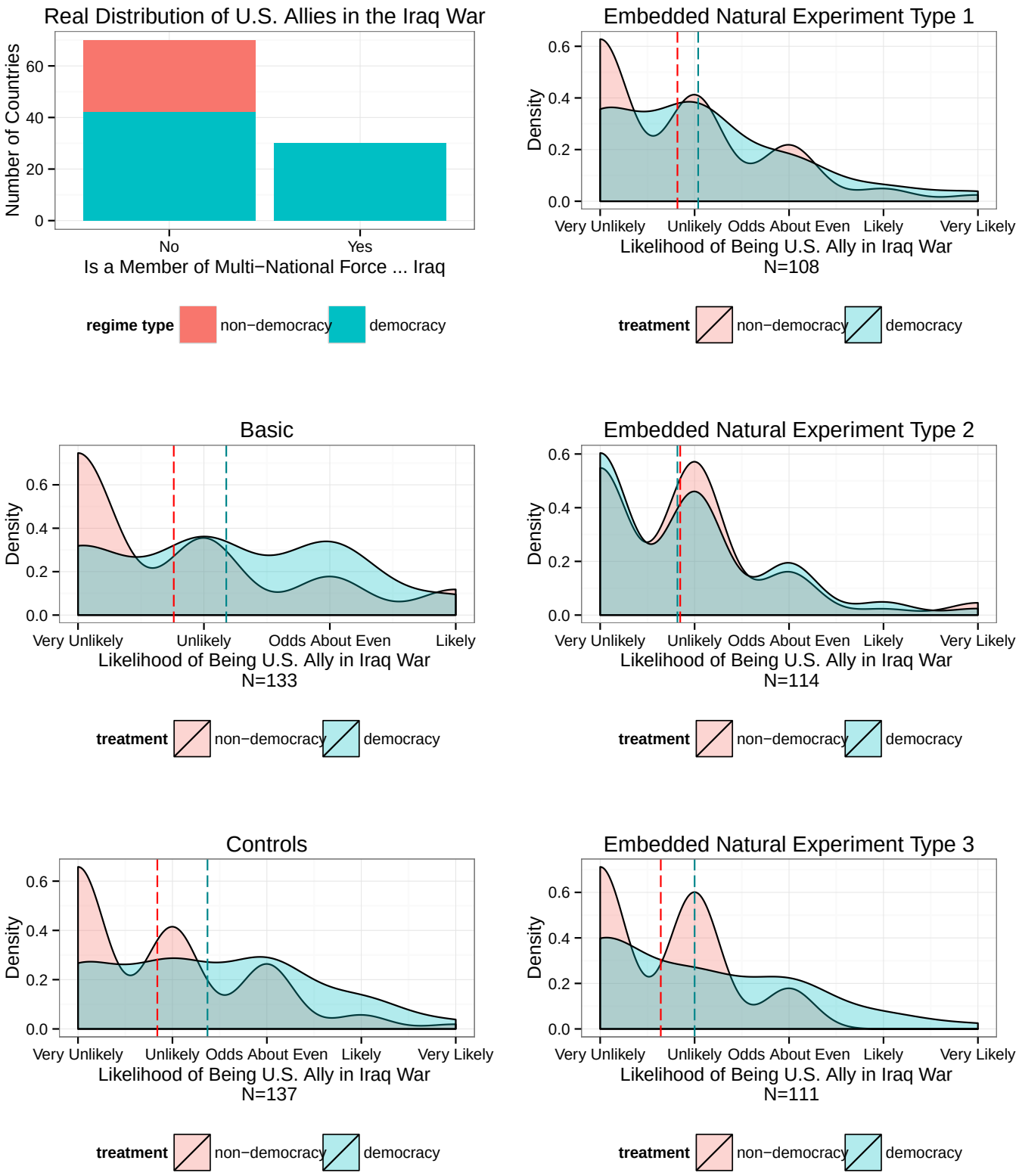


Figure 13: Placebo E: Likelihood of Being U.S. Ally in the Iraq War
Includes data from the January Wave and the March Wave for the Basic, Controls, and Embedded Natural Experiment Type 1 Vignettes.

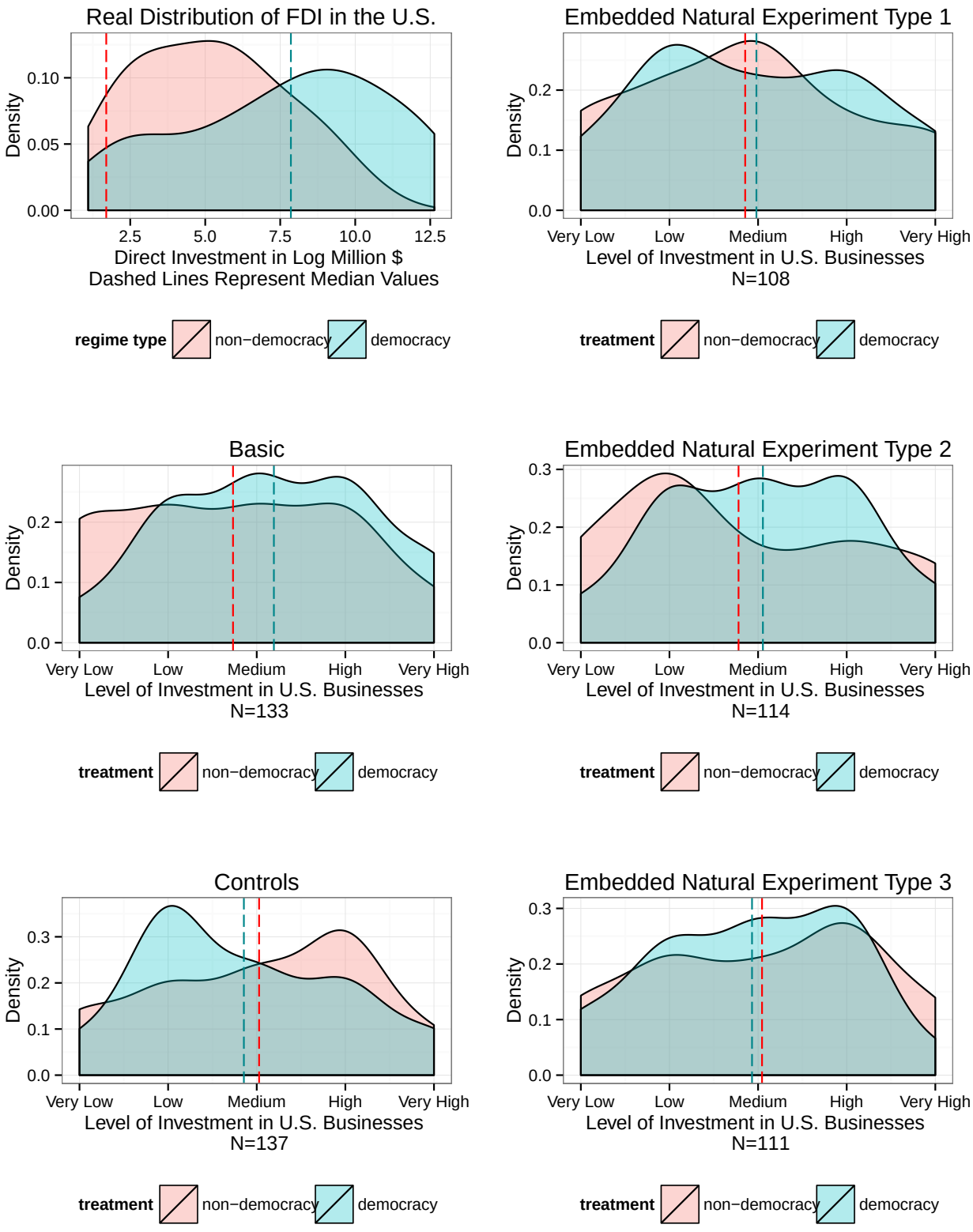


Figure 14: Placebo J: Investment in U.S. Businesses

Table 6: *p*-values from Randomization Inference: Placebo Tests E, F, J, I

E: Likelihood of Military Alliance	One-sided p-value
Basic	<0.001
Controls	0.066
Embedded Natural Experiment 1	0.100
Embedded Natural Experiment 2	0.463
Embedded Natural Experiment 3	0.265
F: Trade with U.S.	**One-sided p-value**
Basic	0.006
Controls	0.165
Embedded Natural Experiment 1	0.205
Embedded Natural Experiment 2	0.801
Embedded Natural Experiment 3	0.512
I: Likelihood of Being Iraq War Ally	**One-sided p-value**
Basic	0.007
Controls	0.003
Embedded Natural Experiment 1	0.148
Embedded Natural Experiment 2	0.610
Embedded Natural Experiment 3	0.022
J: Investment in U.S. Businesses	**One-sided p-value**
Basic	0.021
Controls	0.823
Embedded Natural Experiment 1	0.324
Embedded Natural Experiment 2	0.140
Embedded Natural Experiment 3	0.710

We conducted permutation inference using difference-in-means as our test statistic. In Placebo Tests E and F, we block on survey waves for the Basic, Controls, and Embedded Natural Experiment Type 1.

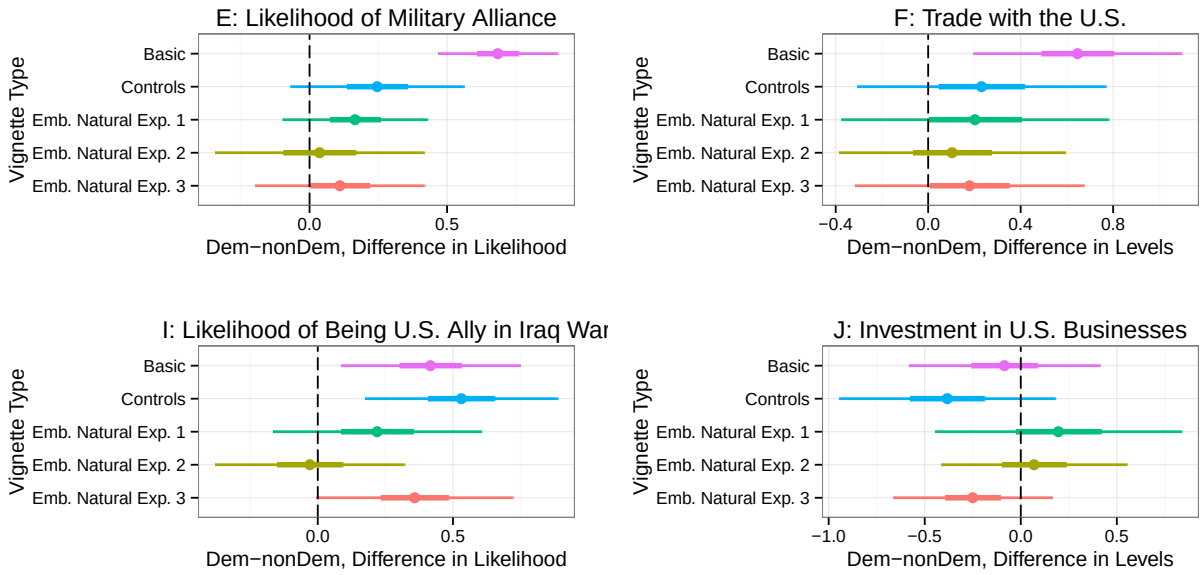


Figure 15: Coefficient Plot: Placebo Tests E, F, J, I

5.5 Manipulation Measure

We performed a manipulation measure to examine whether the different vignette types induced different changes in the causal factor of interest: how democratic was the aggressor.[18] In some surveys we asked respondents to either think about the aggressor country; in others we asked respondents to think about "countries that are most likely to experience the scenario you read." We then ask them how likely the country (countries) they have in mind has (have) the following regime types: full democracy, democracy, semi-democracy, semi-autocracy, and full autocracy.[19] The format of the manipulation measure question is presented in Figure 16. Note that we provide example countries for each regime type so respondents have some reference.

Figure 16: Manipulation Measure Question

	Very unlikely (0-20% chance)	Unlikely (21-40% chance)	Chances About Even (41-60% chance)	Likely (61-80% chance)	Very likely (81-100% chance)
Full democracy (ex: Canada, Japan)	○	○	○	○	○
Democracy (ex: India, Brazil)	○	○	○	○	○
Semi-democracy (ex: Russia, Iraq)	○	○	○	○	○
Semi-autocracy (ex: Egypt, Uganda)	○	○	○	○	○
Full autocracy (ex: China, Saudi Arabia)	○	○	○	○	○

Respondents beliefs about the regime-type of the country changed as we intended in all vignettes: under the democracy vs the non-democracy conditions, respondents assigned a higher likelihood to the country being a full democracy, democracy, or semi-democracy, and a lower likelihood of being a semi-autocracy or full autocracy (see Figures 17 and 18). However, these baseline levels were very different from what we anticipated, radically changing the meaning of the results emerging from studies such as this one.

Respondents perceived the country in the scenario to be much less democratic than we anticipated and than a literal interpretation of the regime type portion of the vignette would suggest. We expected, as we believe would most scholars reading work such as this would, that respondents would interpret "a country that is DEMOCRACY and shows every sign that it will remain a democracy" to be a country with a Polity score at least as large as 6, such as Brazil, India, or Hungary, if not Japan or Canada. However, respondents assigned the highest likelihood for the Democracy condition for most vignette types to Semi-Democracy (Polity score 1 to 5), for which our examples in the survey were Russia or Iraq. Under the Basic Design, respondents' actually assigned a higher probability to the "democracy" being a "Full Autocracy" than a "Full Democracy". Similarly, the respondents' beliefs about the country in the non-democracy condition were much more autocratic

[18] The manipulation measure data we analyze in this section are all from the March Wave.

[19] We based these regime types on Polity IV's categorization: full democracy (Polity score=10), democracy (6 to 9), open anocracy (1 to 5), closed anocracy (-5 to 0), and autocracy (-10 to -6). We replaced the terms "open anocracy" with "semi-democracy" and "closed anocracy" with "semi-autocracy" because we suspect most respondents are unfamiliar with the definition of anocracy.

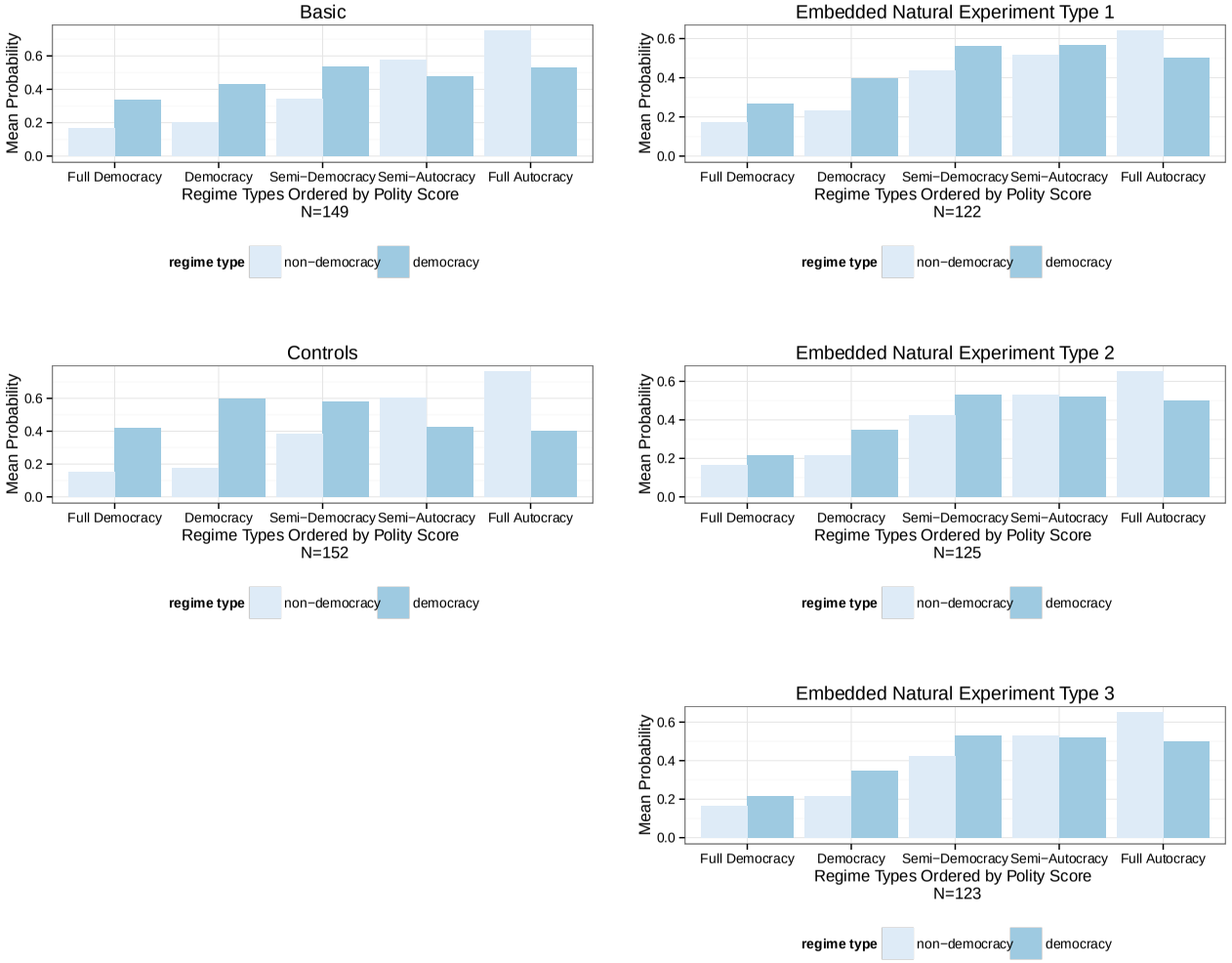
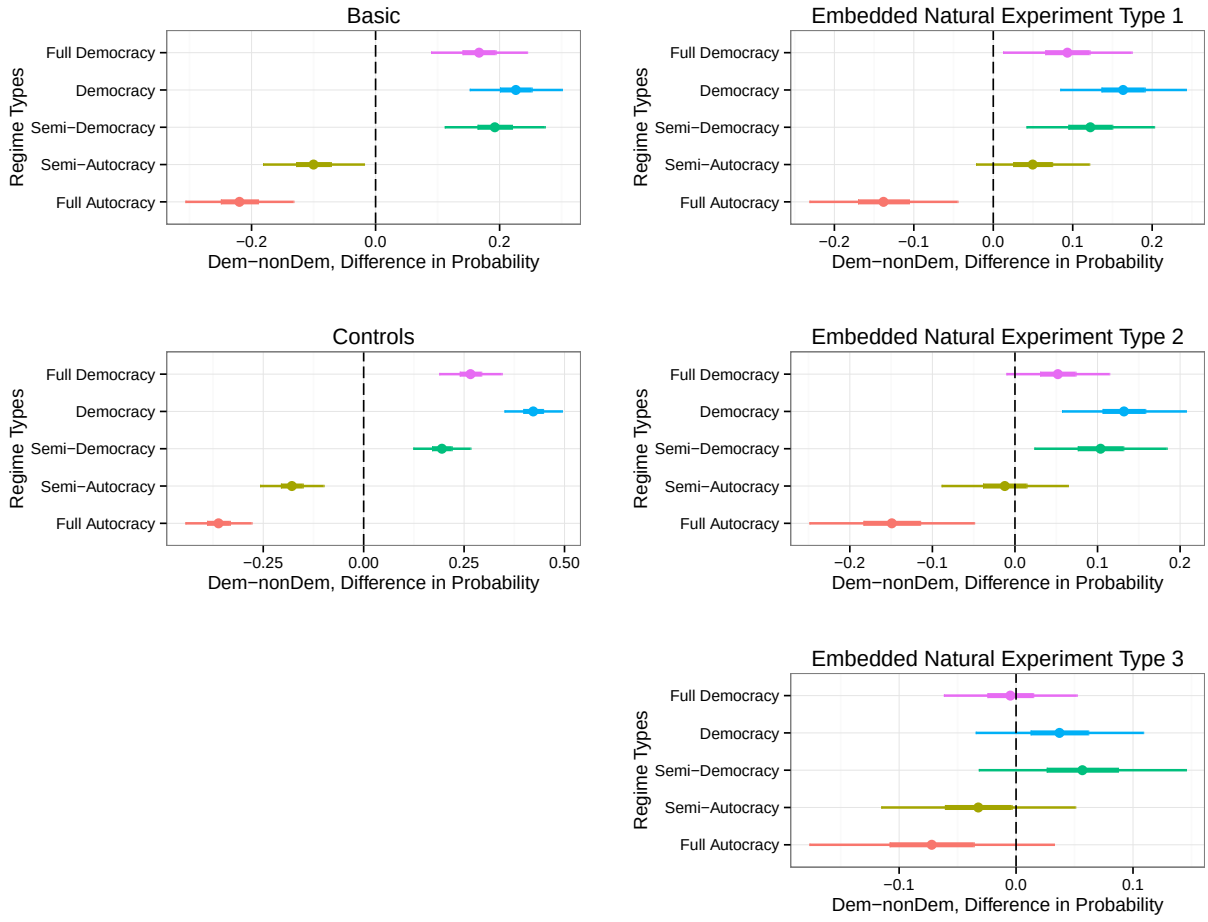


Figure 17: Manipulation Measure Histograms

In the survey questions, subjects indicated the likelihood by selecting a range of probabilities (e.g., Very Unlikely: 0-20% chance). For ease of analysis, we converted each range to point values by taking the mean of each range. Note that values need not sum to 1 because we did not impose a forced sum on the responses.

Figure 18: Coefficient Plot: Manipulation Measure



than we expected. The scenario read that "the country is NOT a DEMOCRACY and shows no sign of becoming a democracy." Respondents assigned the highest probability to Full Autocracy (Polity score -10 to -6), for which our examples were China and Saudi Arabia. Respondents thus interpreted the terms "democracy" and "non-democracy" in the vignettes very differently from the literal meaning.

The reason for this seems clear: respondents did not read the sentence about regime-type independent of the other features of the scenario. The fact that this country was developing nuclear weapons, "had refused all requests to stop its nuclear weapons program" and is otherwise portrayed as a threat led respondents to condition their interpretation of the sentence pertaining to regime-type. While Russia and Iraq are not the countries one thinks of when reading about "DEMOCRACY", they seem to be the kinds of countries one thinks about when reading about democracies building nuclear weapons in a threatening manner. For future waves of this study we will vary the countries we use as examples to make sure that this is not driving respondents answers to the manipulation measure. In general this result speaks to the broader message of this paper that it is rarely possible to simply manipulate a specific feature of a scenario-based survey experiment without also manipulating the respondents' interpretation and understanding of other features of the scenario.

Respondents also perceived the difference in the level of democracy to be much smaller than a literal interpretation of the regime-type portion of the vignettes would suggest. If we interpret the "democracy" phrasing to refer to countries centered in the middle of our "democracy" category (Polity $= 8$), and the "non-democracy" phrasing to refer to countries centered on our "semi-autocracy" and "semi-democracy" categories (Polity $= 0$), then the change in level of democracy should be about 8 points on the Polity scale. We could also adopt a more autocratic interpretation of "non-democracy", centering the interpretation at about a Polity$= -3$. The effect of our vignettes on perceived level of democracy, then, should be about 8 to 11 points on the Polity scale. We found that the level of democracy increased by 3 and 6.5 points for the Basic and Controls Vignettes, substantially smaller than this literal interpretation. Furthermore, the Embedded Natural Experiments Vignettes had even weaker effects, generating changes in perceived level of democracy of 1.0, 1.7, and 0.8(expressed on the Polity Scale) for the ENE1, ENE2 and ENE3 vignettes, respectively. These differences are significant at the 0.1 percent level for ENE1 and ENE2, but not (p=0.373) for ENE3.[20] This highlights a potential problem with embedded natural experiments: while they seem best suited to reduce confounding, they also may reduce changes in the causal factor of interest. IV bias becomes larger as the instrument becomes weaker. To assess bias, scholars should do a formal assessment of IV bias that takes into account the strength of the instrument.[21]

The magnitude of these effects matter for how we should interpret the results of scenario-based survey experiments. As we show below, the embedded natural experiments actually induce similar effects in some outcome measures as the Basic and Controls Vignettes. However, these changes are generated from a much smaller change in the level of democracy of the country. If it truly is the perceived level of democracy of the country that is causing these changes then these differences need to be taken into account when estimating the effect of perceived level of democracy on public opinion. Put simply, given similar sized ITT estimates, as we do in fact observe for Basic, Controls, ENE1, and ENE3, the implied average effect of democracy assuming linear effects on the Polity

[20] NPC p-values using the Fisher combining function.

[21] We will do this in our next draft.

scale is three to seven times larger for the ENEs than for the Basic and Control Vignettes.

Finally, even if we assume that this method is able to recover the true causal effects of target country regime-type on US public support for using force, the ITT estimates do not correspond to the intended average causal effect. The reason is that the original question—the quantity of interest—is about the counterfactual effect of a target country being a democracy vs being a non-democracy. However, such a question implies a difference, as argued above, of approximately an 8 point change in the Polity score, while the actual variation induced was between 1 and 7 points. Thus, for estimating the magnitudes of effects one needs to go beyond ITT estimates to IV estimates, rescaling by the strength of the instrument. Scenario-based survey experiments of the democratic peace reporting only ITT estimates are thus likely to underestimate the intended quantity of interest because the other details in the scenario will tend to shift the respondent's interpretation of "democracy" and "non-democracy" towards each other.

5.6 Plausibility Measure

When we first planned our plausibility measures, we were concerned that subjects would consider some of the scenarios to be more plausible than others. In particular, we worried that subjects might think scenarios involving a non-democratic aggressor in the Basic and Controls Vignettes to be more plausible than those involving a democratic aggressor. This could occur because there are far more real world non-democracies than democracies engaged in nuclear proliferation that threatens the U.S.

We hypothesize a few possible consequences of scenario implausibility. First, respondents may experience a distancing effect, or *verfremdungseffekt*, in which respondents become more conscious of their role as a respondent or conscious of the basis for their answers (Kiralyfalvi, 1990). Second, respondents may lose respect for the study, possibly leading them to answer less sincerely or to leave the survey. Third, respondents may not be able to answer the questions, increasing "I don't know" answers, noise, and attrition from the survey. Finally, the above effects may depend on characteristics of the respondent, such as their education, their realism, and their world-view. This could lead to a systematic bias if respondents who are more likely to perceive implausibility, such as educated respondents or respondents with a realist epistemology, are more likely not to complete the survey, to answer "I don't know," or to have more noisy responses. These problems can confound inference in survey experiments if the treatment and control vignettes vary greatly in plausibility.

Fortunately for our survey experiment, subjects in treatment and control — for all five vignette types — thought the scenarios they read had similar plausibility. (See Figure 20.) Across the vignettes, subjects on average thought the scenarios have a 45 to 65 percent chance of happening in the next five years. (See Figure 19.) Despite the dearth of real world democracies threatening to build nuclear weapons, subjects think there is around a 50-50 chance that the scenarios involving a democracy could occur in the near future. This again may reflect how respondents interpret "democracy" to refer more to what we called semi-democracies such as Russia. We also note that the Embedded Natural Experiment Vignettes are not viewed as less plausible than the other two vignette types, although subjects rated Type 3 as the least plausible of the vignette types. Because we did not tell subjects what happened to the government of the country following the assassination attempt in Type 3, subjects might feel the scenario is less plausible due to the lack of information.

Figure 19: Coefficient Plot: Plausibility Measure

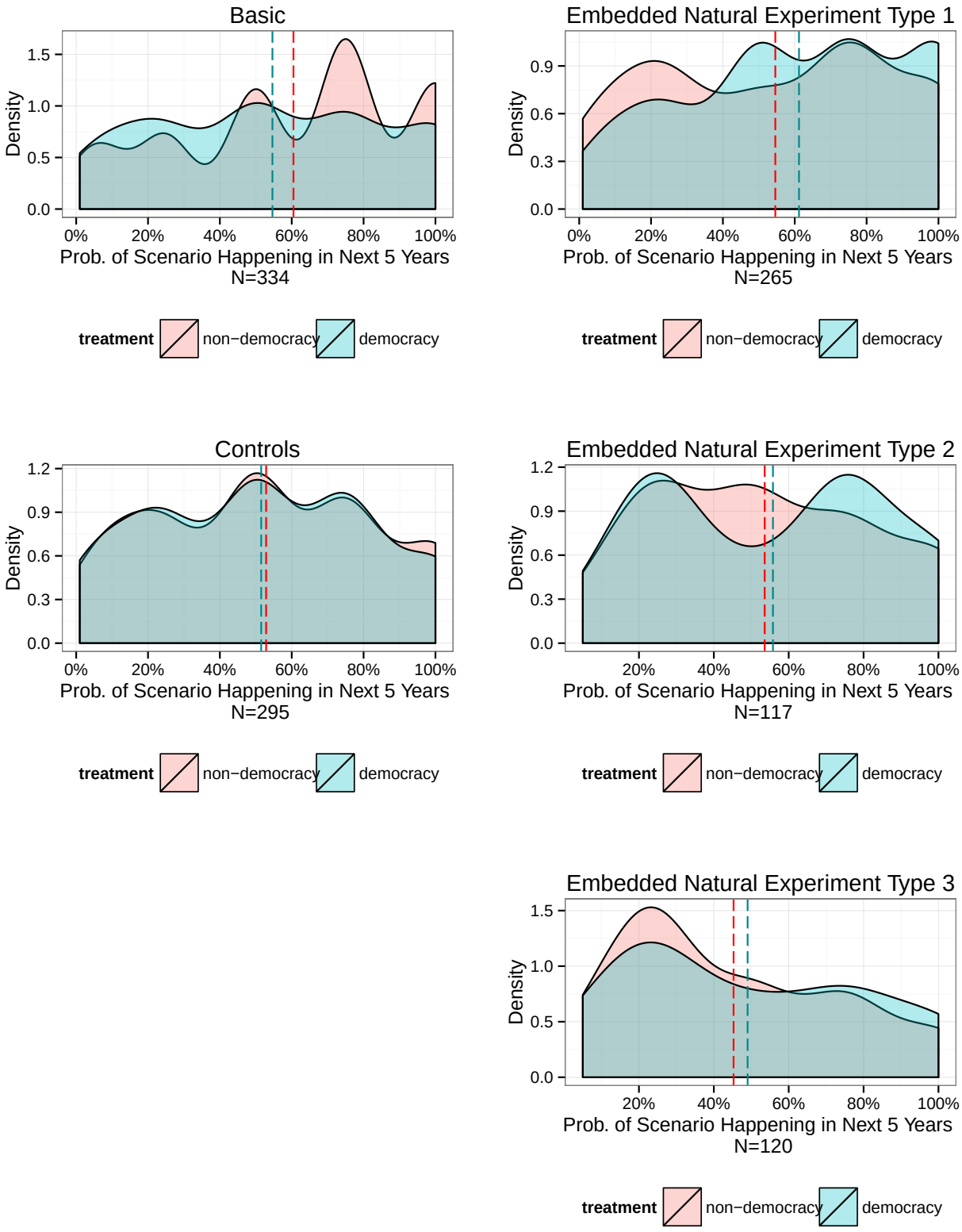
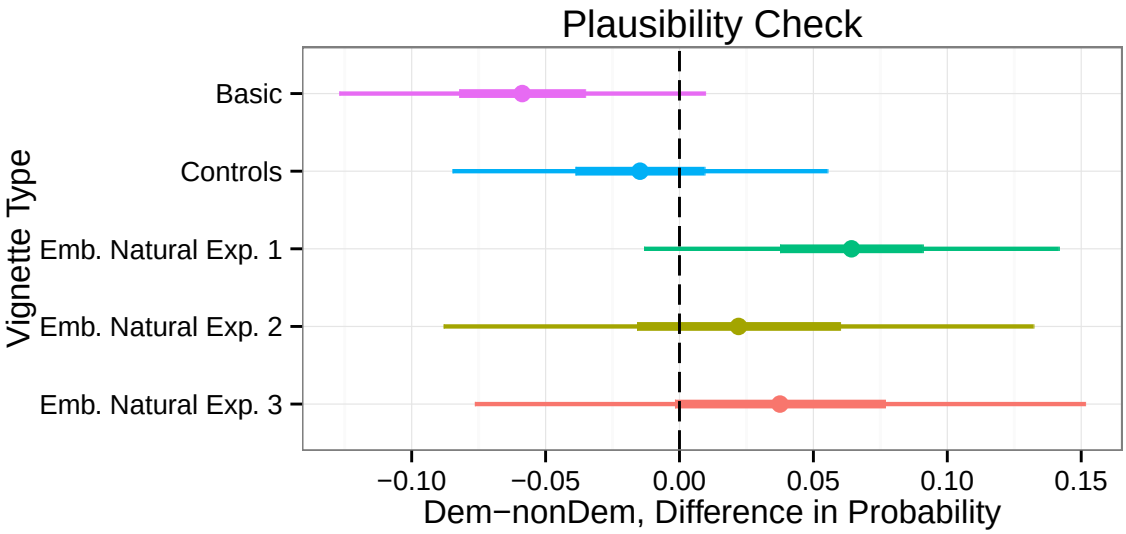


Figure 20: Coefficient Plot: Plausibility Measure



6 Explaining the Democratic Peace

In the March Wave of our survey, we not only evaluated the vignettes for potential confounding but we also sought to determine whether subjects are less likely to support using military action against the democratic aggressor compared with the non-democratic aggressor. Furthermore, we analyze possible mechanisms that might cause subjects to decrease support for going to war against a democracy. Our results demonstrate that the aggressor's being a democracy slightly, but insignificantly, reduced respondents' willingness to launch a military strike. Furthermore, mediators such as perceptions of threat, cost, and morality do not seem to be affected by the treatment. Our null results could have resulted from the small sample size we used.

6.1 The Effect of Democracy

The main outcome measure of our survey experiment asks subjects whether they "favor or oppose using U.S. armed forces to attack the nuclear development sites" of the aggressor country. Subjects are given three answer choices: favor, oppose, or I don't know. For our analysis, we replicate Tomz and Weeks's (2013) methodology. The authors estimated the effect of the aggressor's being a democracy on the drop in percentage points of respondents who favor military action. Table 7 summarizes the effect of democracy on support for striking the aggressor country. We present both Tomz and Weeks's results as well as our own broken down by vignette types.

Table 7: The Effect of Democracy on Willingness to Strike

Vignette Types	Non-Dem	Dem	Effect of Dem	95% CI Lower Bound	95% CI Upper Bound	N
Tomz and Weeks: U.K.	34.2	20.9	-13.3	-19.6	-6.9	762
Tomz and Weeks: U.S. (between)	53.3	41.9	-11.4	-17	-5.9	1273
Tomz and Weeks: U.S. (within)	50	38.5	-11.5	-14.7	-8.3	972
Basic	32.4	21.9	-10.6	-25.5	4.4	149
Controls	29.1	19.0	-10.1	-24.2	4.1	152
Emb. Natural Exp. 1	35.7	23.6	-12.1	-29.2	5.1	122
Emb. Natural Exp. 2	35.0	35.7	0.7	-17.0	18.4	125
Emb. Natural Exp. 3	40.0	26.7	-13.3	-30.9	4.3	123

The table gives the percentage of respondents who supported military strikes when the target was a democracy and when it was not. The difference is the estimated effect of democracy. Tomz and Weeks (2013) measured both differences between subjects and differences within subjects in their U.S. survey.

In the original experiment, Tomz and Weeks found that democracy significantly reduced the percentage of subjects who support war by 11.5 and 13.3 points. While we found similar drops in support for war, the effects were less significant.[22]

[22]Our less significant results likely occurred because of our smaller sample sizes. We also note that the baseline level of militarism was much lower in our survey than it was in Tomz and Weeks's U.S. surveys. This suggests that

One interesting anomaly is the fact that the manipulation in Type 3 of the Embedded Natural Experiments Vignettes had a large effect on support for war, despite having little effect on the perceived level of democracy of the country. If regime type did not affect subjects' support for war, what did? One conjecture is that subjects have norms against the use of violence to seize power. The success of the assassination attempt, rather than the subsequent regime change, might be enough to turn subjects against the non-democracy. As Tomz and Week's (2014) working paper on human rights and the democratic peace demonstrates, procedural democracy matters less than violation of human rights in turning Americans against an aggressor. Perhaps assassinating an elected official could be considered a violation of human rights. A second conjecture is that while respondents' perception of the level of democracy of the country was not perceptibly different across conditions, respondents perceived a difference in whether the country was heading in the "right" or "wrong" direction, depending on the outcome of the assassination, and were more likely to support a country heading in a democratic direction.

6.2 Mediator Variables

Tomz and Weeks (2013) examine mechanisms behind the democratic peace. The three main sets of mechanisms they test include perceptions of threat, cost, success, and morality. We perform implicit mediation analysis by treating subjects' response to the meditation questions as additional outcome measures.

We begin by examining the effect of democracy on perceptions of threat, as shown in Table 8. The first column shows what respondents expected when the scenario involved a non-democracy. The second column shows how respondents' expectations changed given an identical scenario involving a democracy. Note that in the original survey experiment, Tomz and Weeks examined the effect within subjects who received both treatment and control scenarios in separate survey waves. In contrast, our effects are estimated between subjects in treatment and control. Tomz and Weeks noted large, negative and significant effects of democracy on perceptions of the aggressor threatening to use nuclear weapons against other countries, the U.S., and its allies. In contrast, we only observe these significant effects of democracy in the Basic Vignettes: a 27 percentage point drop for threatening to use nukes versus another country and a 18 percentage point drop for threatening to use nukes versus the U.S. or an U.S. ally. For the other vignette types, the effects are either very small or too heterogeneous to produce a significant difference between treatment and control. Whereas Tomz and Weeks estimated effects significant at the 95 percent level for all of the mediator variables, we could not do so far any of our vignettes types. One reason for the disparity between their finding and ours is that our sample size is relatively small (less than 200 per vignette type) compared to theirs (an effective sample size of more than 1,800).

Similar to the Tomz and Weeks findings, our results suggest that the effect of democracy on the perception of cost is not significant. Nevertheless, we report one anomaly: in the Embedded Natural Experiment Vignettes Type 3, the aggressor being a democracy increased the percentage

the subjects of the original survey may be different than our respondents. First, YouGov panelists might be different from Mechanical Turk workers. Second, these differences may reflect changes in world politics since 2010 when Tomz and Weeks conducted their surveys. For instance, in November 2013, Iran and the P5+1 countries signed the Geneva interim agreement that consists of a short-term freeze of portions of Iran's nuclear program in exchange for decreased economic sanctions on Iran. This recent development might have made our subjects more willing to consider diplomatic, rather than military, options when dealing with an aggressor engaged in nuclear proliferation.

Table 8: The Effect of Democracy on Perceptions of Threat

If the U.S. did not attack, the country would...	Belief if Autocracy	Effect of Democracy	Significant at $\alpha = 0.05$? (1=yes, 0=no)
Build nuclear weapons			
Tomz and Weeks	75	-3	1
Basic	81	-11	0
Controls	88	-3	0
Emb. Natural Exp. 1	77	6	0
Emb. Natural Exp. 2	82	9	0
Emb. Natural Exp. 3	76	11	0
Threaten to use nukes vs. another country			
Tomz and Weeks	52	-14	1
Basic	58	-27	1
Controls	55	-14	0
Emb. Natural Exp. 1	68	1	0
Emb. Natural Exp. 2	66	-13	0
Emb. Natural Exp. 3	70	0	0
Threaten to use nukes vs. U.S. or U.S. ally			
Tomz and Weeks	45	-11	1
Basic	42	-18	1
Controls	30	-13	0
Emb. Natural Exp. 1	46	-5	0
Emb. Natural Exp. 2	48	-6	0
Emb. Natural Exp. 3	46	-8	0
Launch a nuclear attack vs. another country			
Tomz and Weeks	34	-8	1
Basic	13	-5	0
Controls	15	-12	1
Emb. Natural Exp. 1	20	-22	1
Emb. Natural Exp. 2	20	-8	0
Emb. Natural Exp. 3	17	-10	0
Launch a nuclear attack vs. U.S. or U.S. ally			
Tomz and Weeks	30	-6	1
Basic	12	-10	0
Controls	8	-4	0
Emb. Natural Exp. 1	13	-6	0
Emb. Natural Exp. 2	14	-3	0
Emb. Natural Exp. 3	13	-7	0

The first column gives the percentage of respondents who thought the event had more than a 50% chance of happening when the country was an autocracy. The second column shows how much that percentage changed when the respondents considered an identical scenario involving a democracy. Tomz and Weeks (2013)'s effects are estimated within subjects using 972 respondents. Our effects are estimated between subjects.

of people who think that there is more than a 50 percent chance the U.S. economy would suffer by 26 points.

In terms of the effect of democracy on the perceptions of success, Tomz and Weeks reported that democracy had a small (5 percentage points) but significant effect on people's expectations about preventing nuclear proliferation in the short and long run. We only observe these significant effects in the Basic and Controls Vignettes. Nevertheless, when we do observe them, the effects are actually much greater than those in the original survey experiment. For instance, within the Controls Vignettes, the aggressor's being a democracy reduced the percentage of subjects who think there is more than a 50 percent change a U.S. military strike would prevent nuclear proliferation in the future by 15 points. Similarly, within the Basic Vignettes, the effect of democracy on perceptions of preventing nuclear wapons in the long run is -13 percentage points. But we failed to notice these large and significant effects in the Embedded Natural Experiment Vignettes.

For the fourth mediator, Tomz and Weeks observed a substantial effect of democracy on perceptions of morality. When told the aggressor is a democracy, the percentage of subjects who think it would be immoral to attack the country increased by 7 percentage points. In contrast, we do not observe this significant increase in our data. Depending on vignette types, the effect of democracy may be positive (Controls, Embedded Natural Experiment Type 1), negative (Basic, Embedded Natural Experiment Type 2), or zero (Embedded Natural Experiment Type 3).

In contrast to Tomz and Weeks's results, which speak rather clearly to which mechanisms are at work, our results do not offer clear evidence. We believe this uncertainty is a consequence of our smaller sample size. In a subsequent wave we will increase the sample size appropriately.

Table 9: The Effect of Democracy on Perceptions of Cost

If the U.S. did attack...	Belief if Autocracy	Effect of Democracy	Significant at $\alpha = 0.05$? (1=yes, 0=no)
The country would attack the U.S. or U.S. ally			
Tomz and Weeks	39	0	0
Basic	52	-5	0
Controls	51	-1	0
Emb. Natural Exp. 1	41	-5	0
Emb. Natural Exp. 2	49	-14	0
Emb. Natural Exp. 3	46	2	0
The U.S. military would suffer many casualties			
Tomz and Weeks	32	1	0
Basic	49	3	0
Controls	46	-1	0
Emb. Natural Exp. 1	43	3	0
Emb. Natural Exp. 2	50	-1	0
Emb. Natural Exp. 3	45	7	0
The U.S. economy would suffer			
Tomz and Weeks	21	0	0
Basic	43	10	0
Controls	39	7	0
Emb. Natural Exp. 1	37	6	0
Emb. Natural Exp. 2	40	7	0
Emb. Natural Exp. 3	46	26	1
U.S. relations with other countries would suffer			
Tomz and Weeks	49	4	1
Basic	66	4	0
Controls	66	3	0
Emb. Natural Exp. 1	57	7	0
Emb. Natural Exp. 2	66	-7	0
Emb. Natural Exp. 3	53	11	0

The first column gives the percentage of respondents who thought the event had more than a 50% chance of happening when the country was an autocracy. The second column shows how much that percentage changed when the respondents considered an identical scenario involving a democracy. Tomz and Weeks (2013)'s effects are estimated within subjects using 972 respondents. Our effects are estimated between subjects.

Table 10: The Effect of Democracy on Perceptions of Success and Morality

If the U.S. did attack...	Belief if Autocracy	Effect of Democracy	Significant at $\alpha = 0.05$? (1=yes, 0=no)
It would prevent nukes in the near future			
Tomz and Weeks	66	-5	1
Basic	66	0	0
Controls	74	-15	1
Emb. Natural Exp. 1	66	5	0
Emb. Natural Exp. 2	66	-1	0
Emb. Natural Exp. 3	69	2	0
It would prevent nukes in the long run			
Tomz and Weeks	30	-5	1
Basic	13	-13	1
Controls	18	-6	0
Emb. Natural Exp. 1	20	-12	0
Emb. Natural Exp. 2	15	-8	0
Emb. Natural Exp. 3	19	-10	0
It would be immoral to attack			
Tomz and Weeks	31	7	1
Basic	26	-11	0
Controls	19	2	0
Emb. Natural Exp. 1	26	14	0
Emb. Natural Exp. 2	25	-1	0
Emb. Natural Exp. 3	24	0	0

The first column gives the percentage of respondents who thought the event had more than a 50% chance of happening when the country was an autocracy. The second column shows how much that percentage changed when the respondents considered an identical scenario involving a democracy. Tomz and Weeks (2013)'s effects are estimated within subjects using 972 respondents. Our effects are estimated between subjects.

7 Conclusion

7.1 Learning from Survey Experiments

In this paper we present a theoretical framework for thinking about confounding in scenario-based (and framing) survey experiments. Confounding occurs when manipulating one aspect of a scenario changes subjects' beliefs about unspecified aspects of the scenario. We hypothesize that subjects "fill in" these unspecified characteristics in a manner consistent with their knowledge of the real world and the information presented in the vignette. If our Roughly Realistic Bayesian Model is correct, the confounding we observe in scenario-based survey experiments will be similar to the relationship of these variables in real life. Results from our surveys confidently reject the null that there is no confounding, and are consistent with the Roughly Realistic Bayesian Model: confounding occurred along multiple dimensions. In addition, the differences between the treatment and control groups' beliefs about the unspecified characteristics in the vignettes reflect differences between democracies and non-democracies in the real world.

Estimates of average causal effects are always local,[23] depending on the sample and the kinds of variation induced by the experiment. This is also true of scenario-based survey experiments. A change in the characteristics of the sample, in the way in which treatment is manipulated, or in other characteristics of the study will generally change the causal estimand. However, this issue not only pertains to the representativeness of the sample of respondents, but as we show, also to the kinds of variation induced in the beliefs of the respondents by the vignette. This variation may not correspond to the intended variation in the causal factor of interest, and therefore the quantity estimated may not correspond to the quantity of interest. Specifically, the literature on the democratic peace tends to theorize and operationalize the treatment level "democracy" as involving a fully institutionalized democracy, rather than a semi-democracy or even semi-autocracy. In terms of the widely used Polity score, "democracy" is typically used to refer to countries with a Polity score above 6 (from a range of -10 to 10).

And yet, in our manipulation measure *under the democracy treatment*[24], subjects thinks the "democracy" aggressor has a higher likelihood of being a semi-democracy or semi-autocracy than a democracy or full-democracy. The kinds of "democracy" respondents have in mind correspond to the democracy level of countries like South Sudan and Algeria, not France or Japan. Since it is almost certainly the case that the causal process generating the democratic peace is stronger for more fully institutionalized democracies, the estimates from these kinds of survey experiments will provide a substantial underestimate of the effect of democracy. It may be possible to correct these underestimates through the use of IV estimation, using the manipulated vignette as an instrument as we recommend and reported beliefs about the causal factor of interest as the treatment variable. However, IV estimation depends on functional form assumptions. The democratic peace may not be linear in the Polity score. In fact, the observational evidence for the democratic peace seems to be strongest at the Polity= 7 threshold, suggesting effects that are non-linear in the Polity score.

In summary, survey experiments based on vignettes remain an extremely powerful tool for studying people's beliefs, opinions, and preferences under scenarios unlikely to arise in the real world. Contrary to what one might think, however, the fact that survey experiments employ random manipulation does not mean that the causal inference is free of confounding. This is because

[23] Unless we implausibly assume constant effects.

[24] For all designs except for the Controls Vignettes.

scenario-based survey experiments randomize *characteristics of the vignette*, but the causal factor of interest is *the belief of the respondent about a characteristic of the scenario*. In order for such a causal inference to not be confounded, scenario-based survey experiments must meet the following criteria: (1) The instrument (the manipulation of an aspect of the scenario) should be random with respect to all respondent characteristics; this is guaranteed by correct administration of the experiment. (2) The instrument (the manipulated aspect of the scenario) must induce a large enough effect in the causal factor of interest (beliefs of the respondent about the causal factor) to avoid being a weak instrument; this can be empirically verified. (3) The instrument (the manipulated aspect of the scenario) must not affect the outcome (beliefs, opinions, or preferences of the respondent) except through the intended causal pathway and therefore should not change the beliefs of the respondent about other characteristics of the scenario (the exclusion restriction). One way to achieve this is by embedding a natural experiment in the scenario, so as to make the manipulated aspect of the scenario appear as-if random in the context of the scenario. (4) The instrument should have a monotonic effect on the causal factor of interest. Criteria (1) and (2) can be empirically verified and confidently established. Criterion (3) can be tested using placebo tests as we demonstrate in this paper; however, it cannot be empirically verified. Just as causal inference in observational studies must lean heavily on the unverifiable assumption of conditionally ignorable treatment assignment, so must causal inference in scenario-based survey experiments lean on the unverifiable exclusion restriction. Similarly, criterion (4) can be tested, but not verified, and must ultimately be assumed. As with observational studies, sensitivity analysis can be performed to assess the susceptibility of one's inference to possible unobserved confounding.

7.2 Next Steps

The current draft primarily discusses the problem of confounding. In our next draft we will reframe the paper so as to direct more attention to (what we are calling) the issue of local effects, possibly changing the title to "Confounding and Local Effects in Survey Experiments". The issue is that any causal effect estimate is actually a weighted average over many different causal effects; estimated effects are only general (not local) if we implausibly assume constant effects. There are multiple dimensions along which causal effects in scenario-based survey experiments will be heterogeneous: (1) effects will vary by respondent; (2) except for dichotomous causal variables, effects will vary depending on the magnitude and range of the manipulation of the causal variable; (3) other characteristics of the scenario will condition (interact with) the causal effect. As with the problem of confounding, we advocate that scholars employ the IV framework for thinking about the kinds of local effects that their design will produce. Specifically, we will emphasize the importance of a theoretically informed and empirically rich manipulation measure; we will report estimates of the LATE of democracy for our different designs using IV estimation, making clear the assumptions needed for these estimates; we will reflect more on how other features of the scenario condition the causal estimand, in particular with respect to the democratic peace.

We also plan to apply these insights to other examples, such as in the study of American politics. Our top candidate is DeSante's 2013 study "Working Twice as Hard to Get Half as Far: Race, Work Ethic, and America's Deserving Poor." In his survey experiment, DeSante presented subjects with a pair of welfare applicants and asked subjects to allocate a fixed amount of money between the applicants. In his Basic Vignettes, he only manipulated the names of the applicants to change subjects' beliefs about each applicant's race. In his Controls Vignettes, he manipulated both the

names of the applicants and their work ethics rating. We suspect, however, the use of racialized first names convey information about the applicants besides their race. For instance, respondents might infer something about the applicants' family background Bertrand and Mullaninathan (2004), their criminal record, and their parenting abilities. As Pager writes in his 2007 review essay, "The use of names to test for black-white differences...is complicated by the social con- text in which racially distinctive names are situated" (111).

For our purposes, we intend to replicate the original DeSante experiment and add placebo tests. The placebo tests asks subjects to indicate how likely they think each applicant has A) graduated from high school, B) worked in the past 12 months, C) has pending criminal charges or convictions, D) grew up in a low-income family, E) has good parenting skills, and F) will have another child in the next two years. Although some of these questions may seem insensitive, they are based on actual evaluation criteria that state governments use to assess welfare applicants. Analyzing responses to these survey questions would allow us to detect whether respondents read more into names like "Emily," "Laurie," and "Latoya" than their racial connotations.

References

Aronow, Peter and Cyrus Samii. 2013. "Does Regression Produce Representative Estimates of Causal Effects?" EPSA 2013 Annual General Conference.

Bechtel, Michael M and Kenneth F Scheve. 2013. "Mass support for global climate agreements depends on institutional design." *Proceedings of the National Academy of Sciences* 110(34):13763–13768.

Bertrand, Marianne and Sendhil Mullaninathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *The American Economic Review* 94(4):991–1013.

Bowers, Jeffrey S and Colin J Davis. 2012. "Bayesian just-so stories in psychology and neuroscience." *Psychological bulletin* 138(3):389.

Brady, Henry E. 2000. "Contributions of Survey Research to Political Science." *PS: Political Science and Politics* 33(1):47–57.

Cao, Xiaoxia. 2014. "The Effects of Narrative Perspectives and Gender Similarity to a Victim on Sympathy and Support for Aid to People in Need." *Studies in Media and Communication* 2(1):28–37.

Caughey, Devin, Allan Dafoe and Jason Seawright. 2013. "Testing Elaborate Theories in Political Science: Nonparametric Combination of Dependent Tests." Manuscript.

Cohen, Geoffrey L. 2003. "Party Over Policy: The Dominating Impact of Group Influence on Political Beliefs." *Journal of Personality and Social Psychology* 85(5):808–822.

Dafoe, Allan and Guadalupe Tunón. 2014. "Tests of Design: Using Placebo Tests to Evaluate Bias." International Studies Association Annual Convention 2014.

Desante, Christopher D. 2013. "Working Twice as Hard to Get Half as Far: Race, Work Ethic, and America's Deserving Poor." *American Journal of Political Science* 57(2):342–356.

Druckman, James N. and Kjersten R. Nelson. 2003. "Framing and Deliberation: How Citizens' Conversations Limit Elite Influence." *American Journal of Political Science* 47(4):729–745.

Gerber, Alan S and Donald P Green. 2012. *Field experiments: Design, analysis, and interpretation*. New York, NY: WW Norton.

Gibson, James L., Gregory A. Caldeira and Lester Kenyatta Spence. 2005. "Why Do People Accept Public Policies They Oppose? Testing Legitimacy Theory with a Survey-Based Experiment." *Political Research Quarterly* 58(2):187–201.

Gilens, Martin. 2001. "Political Ignorance and Collective Policy Preferences." *American Journal of Political Science* 95(2):379–396.

Gilens, Martin. 2002. *Navigating Public Opinion: Polls, Policy, and the Future of American Democracy*. Oxford: Oxford University Press chapter An Anatomy of Survey Based Experiments.

Grieco, Joseph. M., Christopher. Gelpi, Jason Reifler and Peter. D. Feaver. 2011. "Let's Get a Second Opinion: International Institutions and American Public Support for War." *International Studies Quarterly* 55(2):563–583.

Hainmueller, Jens, Daniel J Hopkins and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22(1):1–30.

Hainmueller, Jens and Michael J. Hiscox. 2010. "Attitudes toward Highly Skilled and Low-skilled Immigration: Evidence from a Survey Experiment." *American Political Science Review* 104(1):61–84.

Harley, Trevor A. 2013. *The psychology of language: From data to theory*. New York: Psychology Press.

Hernan, Miguel A. and Tyler J. VanderWeele. 2011. "Compound Treatments and Transportability of Causal Inference." *Epidemiology* 22(3):368–377.

Herrmann, Richard K., Philip E. Tetlock and Penny S. Visser. 1999. "Mass Public Decisions to Go to War: A Cognitive-Interactionist Framework." *American Political Science Review* 93(3):553–573.

Holyoak, Keith J and Patricia W Cheng. 2011. "Causal learning and inference as a rational process: The new synthesis." *Annual review of psychology* 62:135–163.

Hurwitz, Jon and Mark Peffley. 1997. "Public Perceptions of Race and Crime: The Role of Racial Stereotypes." *American Journal of Political Science* 41(2):375–401.

Imbens, Guido W and Joshua D Angrist. 1994. "Identification and estimation of local average treatment effects." *Econometrica: Journal of the Econometric Society* pp. 467–475.

Iyengar, Shanto. 2012. "Do Attitudes About Immigration Predict Willingness to Admit Individual Immigrants? A Cross-National Test of the Person-Positivity Bias." Manuscript.

Johns, Robert and Graeme A. M. Davies. 2012. "Democratic Peace or Clash of Civilizations? Target States and Support for War in Britain and the United States." *Journal of Politics* 74(4):1038–1052.

Jones, Benjamin F. and Banjamin A. Olken. 2009. "Hit or Miss? The Effect of Assassinations on Insitutions and War." *American Economic Journal: Macroeconomics* 1(2):55–87.

Kent, Sherman. 1964. "Words of estimative probability." *Studies in Intelligence* 8(4):49–65.

King, Gary and Langche Zeng. 2006. "The dangers of extreme counterfactuals." *Political Analysis* 14(2):131–159.

Kiralyfalvi, Bela. 1990. "The Aesthetic Effect: A Search for Common Grounds Between Brecht and Lukacs." *Journal of Dramatic Theory and Criticism* 4(2):19–30.

Mintz, Alex and Nehemia Geva. 1993. "Why Don't Democracies Fight Each Other? An Experimental Study." *The Journal of Conflict Resolution* 37(3):484–503.

Morgan, Stephen L. and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge, UK: Cambridge University Press.

Mutz, Diana C. 2011. *Population-Based Survey Experiments*. Princeton, New Jersey: Princeton University Press.

Pager, Devah. 2007. "The Use of Field Experiments for Studies of Employment Discrimination: Contributions, Critiques, and Directions for the Future." *The ANNALS of the American Academy of Political and Social Science* 609:104–134.

Peffley, Mark, Jon Hurwitz and Paul M. Sniderman. 1997. "Racial Stereotypes and Whites' Political Views of Blacks in the Context of Welfare and Crime." *American Journal of Political Science* 41(1):30–60.

Perfors, Amy, Joshua B Tenenbaum, Thomas L Griffiths and Fei Xu. 2011. "A tutorial introduction to Bayesian models of cognitive development." *Cognition* 120(3):302–321.

Rosenbaum, Paul R. 2002. *Observational studies*. New York, NY: Springer.

Rousseau, David L. 2005. *Democracy and War: Insitutions, Norms, and the Evolution of International Conflict*. Stanford, California: Stanford University Press.

Sanbonmatsu, Kira. 2002. "Gender stereotypes and vote choice." *American Journal of Political Science* 46(1):20–34.

Schildkraut, Deborah J. 2009. "The Dynamics of Public Opinion on Ethnic Profiling After 9/11: Results From a Survey Experiment." *American Behavioral Scientist* 53(1):61–79.

Sekhon, Jasjeet S. 2009. "Opiates for the matches: Matching methods for causal inference." *Annual Review of Political Science* 12:487–508.

Sniderman Paul, M and G Carmines Edward. 1997. *Reaching beyond race*. Cambridge, MA: Harvard University Press.

Sniderman, Paul M., Louk Hagendoorn and Markus Prior. 2004. "Predisposing Factors and Situational Triggers: Exclusionary Reactions to Immigrant Minorities." *American Political Science Review* 98(1):35–49.

Tomz, Michael. 2007. "Domestic Audience Costs in International Relations: An Experimental Approach." *International Organization* 61(4):821–840.

Tomz, Michael and Jessica L. Weeks. 2012. "Human Rights, Democracy, and International Conflict." Unpublished.

Tomz, Michael and Jessica L. Weeks. 2013. "Public Opinion and the Democratic Peace." *American Political Science Review* 107:849–865.

Trager, Robert F. and Lynn Vavreck. 2011. "The Political Costs of Crisis Bargaining: Presidental Rhetoric and the Role of Party." *American Journal of Political Science* 55(3):526–545.