Information Equivalence 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: March 15, 2018

Abstract

Survey experiments often manipulate the description of attributes in a hypothetical scenario, with the goal of learning about those attributes' real-world effects. Such inferences rely on an underappreciated assumption: experimental conditions must be information equivalent (IE) with respect to background features of the scenario. IE is often violated because subjects, when presented with information about one attribute, update their beliefs about others too. Labeling a country "a democracy," for example, affects subjects' beliefs about the country's geographic location. When IE is violated, the effect of the manipulation need not correspond to the quantity of interest (the effect of beliefs about the focal attribute). We formally define the IE assumption, relating it to the exclusion restriction in instrumental-variable analysis. We show how to predict IE violations ex ante and diagnose them ex post with placebo tests. We evaluate three strategies for achieving IE. Abstract encouragement is ineffective. Specifying background details reduces imbalance on the specified details and highly correlated details, but not others. Embedding a natural experiment in the scenario can reduce imbalance on all background beliefs, but raises other issues. We illustrate with four survey experiments, focusing on an extension of a prominent study of the democratic peace.

1 Introduction

The survey experiment is among the most important recent additions to the political scientist's toolbox. The defining feature of such experiments is the deliberate and typically random manipulation of some aspect of the survey protocol (Marsden and Wright 2010, 838). Early survey experiments were narrowly methodological, but after the development of computer-assisted survey technology in the 1980s, social scientists increasingly used them to investigate substantive research questions (Sniderman and Grob 1996). Survey experiments have since become a core methodological tool in political science. Their prevalence has increased rapidly in recent years, and they now appear in almost 1% of all articles published in the discipline's top journals (see Supporting Information [SI], Appendix E).

The appeal of survey experiments stems in large part from their combination of internal and external validity, which makes them a powerful tool for "inferring how public opinion works in the real world" (Gaines, Kuklinski, and Quirk 2007, 4). As the foregoing quotation suggests, substantive survey experiments are designed to shed light on the real-world effects of some attribute or factor, what Barabas and Jerit (2010) call the "natural treatment." In many survey experiments, the real-world effects of interest are *informational*—that is, they concern how people react to the content and format of information presented to them. For instance, how are attitudes towards anti-poverty programs affected by whether these programs are called

Replication files for this paper can be downloaded from Dafoe, Zhang, and Caughey (2017). Further materials can be found at allandafoe.com/ie. The main studies reported in this paper have been pre-registered and pre-analysis plans have been posted. Superscripted capital letters provide hyperlinks to relevant portions of the Online Appendix. For helpful comments, we would like to thank Peter Aronow, Cameron Ballard-Rosa, Adam Berinsky, Matthew Blackwell, David Broockman, Alex Debs, Chris Fariss, Alan Gerber, Donald Green, Sophia Hatz, Dan Hopkins, Susan Hyde, Josh Kalla, Gary King, Audrey Latura, Jason Lyall, Neil Malhotra, Elizabeth Menninga, Nuno Monteiro, Brendan Nyhan, Jonathan Renshon, Bruce Russett, Cyrus Samii, Jas Sekhon, Maya Sen, Robert Trager, Mike Tomz, Jessica Weeks, Teppei Yamamoto, Sean Zeigler, Thomas Zeitzoff, and participants of the University of North Carolina Research Series, the Yale Institution for Social and Policy Studies Experiments Workshop, the Yale International Relations Workshop, the University of Konstanz Communication, Networks and Contention Workshop, the Polmeth 2014 and 2015 Summer Methods Meetings, the Survey Experiments in Peace Science Workshop, the West Coast Experiments Conference, and the Comparative Political Economy and Conjoint Analysis workshop at the University of Zurich. For support, we acknowledge the MacMillan Institute at Yale University, and the National Science Foundation Graduate Research Fellowship Program. Yale IRB has granted exemption to the survey experiments reported in this paper under IRB Protocol # 1302011471.

"welfare" or "assistance to the poor" (Gilens 2002, 236)? How does the framing and sequencing of competing arguments influence support for anti-terrorism legislation (Chong and Druckman 2010)? How do the attributes included in profiles of immigrants affect support for granting them citizenship (Hainmueller, Hangartner, and Yamamoto 2015)? To learn about real-world informational effects such as these, survey experiments manipulate the information presented to survey subjects and compare the responses of subjects assigned to different informational conditions. Although informational survey experiments can be complicated by questions of external validity, causal mechanisms, and other issues, interpreting their results is greatly simplified by the fact that the natural treatment—the presentation of information—closely corresponds to what the experiment manipulates.

In other survey experiments, however, the relationship between the experimental manipulation and the real-world treatment is more problematic. This is particularly true of experiments studying *epistemic* effects: the effects of changing subjects' *beliefs* about some factor of interest, holding constant beliefs about background features of the scenario ("background beliefs"). In some cases, epistemic effects correspond to well-defined real-world treatments. Does passing a budget on time, for example, increase the governing party's electoral support in the next election (Butler and Powell 2014)? When epistemic effects are well defined in this way, background beliefs pertain to those factors that in the real world are not affected by treatment (e.g., in the case of an on-time budget, the party's seat share in the previous election). For other epistemic effects, the natural treatment is less clear, but it is still possible to imagine interventions that manipulate the real-world factor of interest. Examples include the skill level of potential immigrants (Hainmueller and Hiscox 2010) and, to use our running example in this paper, a country's regime type (Tomz and Weeks 2013). Finally, some epistemic effects concern non-manipulable quantities such as gender or race. Desante (2013), for example, is interested in the degree to which differences in support for black and white welfare applicants are explained by "racial animus" rather than beliefs that are the basis of "principled conservatism," such as work ethic. Though not well-defined manipulations, these sorts of epistemic effects still require holding constant some beliefs (cf. Butler and Homola 2017). Despite their differences, what unifies studies of epistemic effects is their goal of inducing different subjects to consider two alternative versions of a scenario, one in which the factor of interest is present and one in which it is absent, without affecting subjects' background beliefs.

Random assignment of survey versions, however, is not sufficient to make inferences about epistemic effects. Rather, an additional assumption is required: the assumption that the survey manipulation is *information equivalent (IE) with respect to relevant background features of the scenario* (cf. Sher and McKenzie 2006).[1] Only if the IE assumption holds can response differences between versions of the survey be attributed to differences in subjects' beliefs about the factor of interest. The problem, however, is that manipulating information about a particular attribute will generally alter respondents' beliefs about background attributes in the scenario as well, thus violating information equivalence. Manipulating whether a country is described as "a democracy" or "not a democracy," for example, is likely to affect subjects' beliefs about such background features as the country's geographic location or demographic composition. If it does, then any differences between experimental groups cannot be reliably attributed to the effects of the beliefs of interest.

Survey experimentalists recognize, of course, that the relationship between survey results and real-world phenomena is far from automatic. Indeed, seminar discussions of survey experiments frequently center on whether the estimated effects are due to the construct of interest or some other aspect of the manipulated text. Moreover, a few published works do reference the specific problem we describe.[2] These include our running example of Tomz and Weeks (2013), who note that previous survey experiments on the democratic peace failed to specify attributes of the scenario "that could confound the relationship between shared democracy and public support for war," such as whether "the country was also an ally, a major trading partner, or a

[1] Though epistemic effects entail holding all background beliefs fixed, they can be estimated if IE holds with respect to beliefs that are "relevant" in the sense that they may affect the outcome. The rest of the paper implicitly presumes that all background beliefs are relevant in this sense.

[2] Previous works have used a variety of terms for violations of information equivalence: "information leakage" (Sher and McKenzie 2006; Tomz and Weeks 2013, 853), "confounding" (Tomz and Weeks 2013, 849; Dafoe, Zhang, and Caughey 2015), "masking" and "aliasing" (Hainmueller, Hopkins, and Yamamoto 2014, 5, 25), violations of "excludability" (Butler and Homola 2017), and "bundled" or "compound" treatments.

powerful adversary" (849, 853; contrast with Mintz and Geva 1993; Johns and Davies 2012). According to our review of the survey-experimental literature, however, the IE assumption and the problems caused by its violation are not widely appreciated. Only 15% of scenario-based survey experiments in our review evince any awareness of the problem.[3] Nor has any work in political science systematically considered the issue of IE in survey experiments. Applied researchers have thus received little guidance on predicting IE violations, diagnosing them when they occur, or avoiding them in the first place.

We contribute on all these fronts. First, we provide a formal definition of the IE assumption in the context of survey experiments, noting its close connection to exclusion restrictions in instrumental variable (IV) analysis. As with IV exclusion restrictions, if the IE assumption is violated, the effect of the experimental manipulation has no necessary relationship with the effect of beliefs about the attribute of interest. We further show that the IE assumption has testable implications—that background attributes of the scenario should be balanced across experimental conditions—which can and should be evaluated using placebo tests. To predict the precise form of this imbalance, we propose (and find support for) a *realistic Bayesian* model of respondent updating, under which imbalance should roughly resemble confounding in observational studies of the real-world attribute of interest.

We also evaluate three experimental designs that may help achieve information equivalence: abstract encouragement, covariate control, and embedded natural experiments, the last of which is our own invention. We find that *abstract encouragement*, which asks subjects to consider the scenario in the abstract rather than thinking of real-world examples, is not effective at reducing imbalance on background beliefs. *Covariate control* (CC), which entails specifying the values of background attributes in order to prevent respondents from updating about them, reduces imbalance only on attributes that are explicitly or implicitly controlled. The *embedded natural experiment* (ENE) design, which constructs a scenario in which the attribute of interest is randomly or haphazardly assigned, tends to reduce imbalance on all background characteristics.

[3] The studies we identified in our review were Brader, Valentino, and Suhay (2008), Hainmueller and Hiscox (2010), Tomz and Weeks (2013), Baker (2015, 98, 103), and Kertzer and Brutger (2016, Appendices 7–9).

We draw empirical support for these conclusions from four survey experiments, focusing mainly on an extension of Tomz and Weeks's study of the democratic peace. We conclude by discussing the strengths and weaknesses of CC and ENE designs and offering recommendations for applied survey experimentalists.

2 Formal Exposition

In this section we formally define the real-world and survey quantities of interest (QOIs) and the role of information equivalence in linking them. For ease of exposition, we focus on cases where the real-world QOI is the total causal effect of a single binary treatment, but our basic conclusions also hold under more general conditions (see SI, Appendix B).

Epistemic effects should generally be defined in relation to real-world quantities. Accordingly, it is important first to clarify what those quantities are before discussing survey estimands. The types of survey experiments we focus on are typically motivated by a substantive causal question of the following form: How does some causal factor D^* affect some outcome Y^* *in the real world*? Tomz and Weeks (2013), for example, are interested in how the regime-type of other countries affects democratic publics' willingness to use force against them. Though the question of interest may be general, well-defined counterfactuals involve specification of context, either a specific state of the world, a distribution over such states, or a class of states. We denote the class of scenarios that the researcher has in mind by $\mathbb{S}$. In Tomz and Weeks (2013), for instance, $\mathbb{S}$ includes scenarios in which another country is developing nuclear weapons and has specified trade levels, alliance relationships, and military strength. For a particular scenario $s \in \mathbb{S}$, the real-world effect of interest is

$$\tau_s^* \equiv Y_s^*(D_s^* = 1) - Y_s^*(D_s^* = 0), \tag{1}$$

where $Y_s^*(D_s^* = d)$ is the potential outcome when D_s^* is set to d.

Defining the scenario class $\mathbb{S}$ as clearly as possible, at least in the researcher's own mind, is important both because τ_s^* may vary across (as well as within) scenario classes and because doing so helps identify the background conditions B_s^* that D_s^* does not affect (e.g., because they are pre-treatment).[4] In Tomz and Weeks (2013), for instance, these background conditions include characteristics such as the continent on which the target country is located, which presumably predates the country's regime-type. Because D_s^* does not affect B_s^*, the latter does not vary within the counterfactual comparison of interest in (1). That is:

$$\textbf{Equivalence of Background Features: } B_s^*(D_s^* = 1) = B_s^*(D_s^* = 0).^{5} \tag{2}$$

Note that equivalence of background features is not an additional assumption, but rather an implication of the definition of τ_s^* in (1).

To gain insight into the real-world counterfactual comparison in (1), survey experimentalists seek to evoke analogous scenarios in subjects' minds and compare their responses. To do so, they present each survey subject i with a scenario description that includes details X, with the goal of ensuring that all subjects are considering scenarios in some set $\mathbb{S}$.[6] In addition, the experiment randomly varies a textual element $Z_i \in \{0, 1\}$, which is intended to manipulate subjects' beliefs $D_i \in \{0, 1\}$ about the causal factor of interest. The survey vignette in Tomz and Weeks (2013), for example, describes the trade, alliances, and military strength of a country that is developing nuclear weapons (X) and labels the country either "a democracy" ($Z_i = 1$) or "not a democracy" ($Z_i = 0$). Let Y_i denote the outcome of interest (e.g., i's support for a preventive attack on the country), and let B_i represent beliefs about background features of the

[4] Defining B_s^* as pre-treatment simplifies the exposition, but we note that for many studies the real-world quantity of interest does hold fixed some characteristics potentially affected by the cause of interest (e.g., trade in Tomz and Weeks 2013). There is no formal problem with doing this (see SI, Appendix B), but it complicates the definition of the real-world QOI and its survey counterpart, underscoring the importance of clearly defining these QOIs.

[5] Where $B_s^*(D_s^* = d)$ is B_s^*'s potential value with D_s^* set to d.

[6] Ideally, survey experiments would ensure that all subjects consider the *same* scenario $s \in \mathbb{S}$. However, because subjects cannot be prevented from idiosyncratically "filling in" missing details of the scenario, we regard this as generally impossible. Thus the best that can be done is ensure that whatever scenario subject i considers is within a desired set $\mathbb{S}$. The goal of providing textual details X is to induce subjects to consider only scenarios in $\mathbb{S}$.

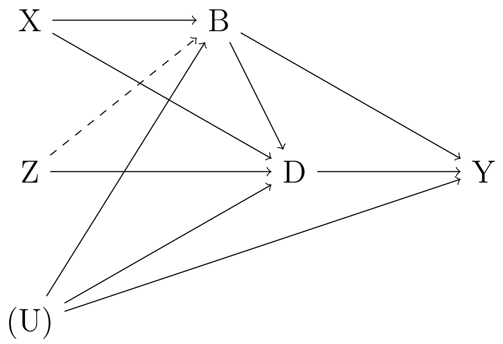


Figure 1: Graphical illustration of information equivalence in survey experiments, for the case in which B precedes D. Z denotes the survey manipulation, X other scenario details, B background beliefs, D beliefs about the causal factor of interest, Y the outcome, and U unobserved common causes of D, B, and Y. In this graph, if the dashed path $Z \to B$ is absent ($\mathcal{A}4$), then $B \perp\!\!\!\perp Z$ and the IE assumption holds.

scenario (e.g., whether i believes the country is located in Europe). For ease of exposition, we assume that beliefs about the factor of interest D do not affect background beliefs B (we relax this condition in SI, Appendix B). Under this assumption, the epistemic effect of interest is

$$\tau_i \equiv Y_i(D_i = 1) - Y_i(D_i = 0). \tag{3}$$

Suppose that instead of manipulating Z, a survey simply presented all respondents with a scenario that contained no direct information about the attribute of interest. Although respondents' beliefs D would probably still vary, comparing Y across respondents with different values of D (supposing we could measure D) would not provide a consistent estimate of D's effect because D and Y (as well as B) are likely to share unobserved common causes U (e.g., subjects' worldview and personality traits; see Figure 1). This is the motivation for survey experiments, which seek to induce exogenous variation in D by randomly varying Z. If Z_i is randomly assigned ($\mathcal{A}1$) and the stable unit treatment value assumption holds ($\mathcal{A}2$), the difference of means across experimental conditions is an unbiased estimate of the "intent-to-treat" (ITT) effect:

$$\text{ITT} \equiv \mathbb{E}[Y_i(Z_i = 1) - Y_i(Z_i = 0)]. \tag{4}$$

Unfortunately, even assumptions $\mathcal{A}1$ and $\mathcal{A}2$ are not sufficient for the ITT effect to entail conclusions about the distribution of τ_i. This inferential link is justified, however, under the standard assumptions of IV analysis (Angrist, Imbens, and Rubin 1996). First, in addition to $\mathcal{A}1$ and $\mathcal{A}2$, the effect of Z on D must be monotonic and non-zero for some subjects ($\mathcal{A}3$). This is typically plausible unless some subjects react perversely to the information provided. A more problematic assumption—and the critical one for our purposes—is the IV exclusion restriction: Z affects Y only through D ($\mathcal{A}4$), which rules out effects through B. Together, assumptions $\mathcal{A}1$–$\mathcal{A}4$ ensure that the ITT effect has the same sign as the complier average causal effect (CACE),

$$\text{CACE} \equiv \mathbb{E}[\tau_i | \{D_i(Z_i = 1) - D_i(Z_i = 0) = 1\}]. \tag{5}$$

If researchers only care whether the CACE is positive or negative, estimating the ITT is sufficient to make this inference. However, if they are also interested in the magnitude of the CACE, the latter can be estimated under assumptions $\mathcal{A}1$–$\mathcal{A}4$ as long as D is measured (without error).

The bottom line is that in order for the estimands identified by the survey experiment (the ITT and, if D is observed, the CACE) to entail conclusions about the epistemic QOI (the effect of D on Y, holding B constant), the four canonical assumptions of IV analysis (or assumptions at least as strong) must be satisfied.[7] A crucial (and testable) implication of the IV assumptions—in particular, of the exclusion restriction ($\mathcal{A}4$)—is that different versions of the survey do not affect subjects' beliefs about background characteristics:

$$\textbf{Information Equivalence of Background Features: } B_i(Z_i = 1) = B_i(Z_i = 0)\ \forall i. \tag{6}$$

If the IE assumption in (6) holds, then all background beliefs should be balanced in expectation across randomized treatment conditions. If IE fails, neither the ITT nor the estimated CACE has any necessary relationship to the epistemic QOI τ_i, let alone the real-world QOI τ_s^*.

[7] An alternative approach would be to employ mediation analysis (Imai et al. 2011; Acharya, Blackwell, and Sen 2017), as Tomz and Weeks (2013) in fact do. The problem with mediation analysis is that it requires assumptions that are typically at least as strong as the IV assumptions and more difficult to validate empirically. For further discussion, see Section 6.3.

3 Predicting and Diagnosing IE Violations

Whether the IE assumption holds depends on how subjects update in response to new information. Most existing studies of epistemic effects implicitly presume that subjects respond to the information in the experimental manipulation Z by updating their beliefs about overtly specified attributes D, but not their background beliefs B. Only if this is true are experimental conditions likely to be IE with respect to background characteristics. Unfortunately, such restricted updating is unlikely under almost any plausible model of human information processing.

Consider, for instance, a "realistic Bayesian" model of information processing. This model has two components. First, it holds that the relevant prior beliefs of survey respondents are *realistic*, in that they reflect the relationships among different attributes in the real world. For example, because democracy and European location are positively correlated in the real world, respondents should believe that a country described as "a democracy" is more likely to be in Europe than one described as "not a democracy." Second, the model holds that survey respondents are *Bayesian* updaters—that is, given their priors, they respond to new information by updating their beliefs according to the laws of conditional probability. The realistic Bayesian model thus predicts that respondents will in general react to survey manipulations by updating their beliefs about any attribute that in the real world is correlated with the information provided in the survey manipulation.[A] Only if they perceive the attribute of interest to be independent of (and thus to convey no information about) background conditions will subjects not update their background beliefs.

The realistic Bayesian model predicts not only that IE will often be violated, but also the precise form of these violations. Specifically, it predicts that the imbalance on background beliefs between experimental conditions should resemble covariate imbalance in analogous observational studies. Thus, for example, the factors that confound real-world studies of the democratic peace—trade, geography, culture—should also "confound" survey experiments on the same topic. This specificity is valuable because it enables scholars to formulate precise predictions about the probable form of IE violations and to design their survey experiment so as to diagnose and ameliorate them. As we show below, we find substantial empirical evidence that survey subjects update their beliefs in a manner consistent with the realistic Bayesian model.[B] We emphasize, however, that other plausible models of information processing, such as those emphasizing stereotypes or heuristics (e.g., Kahneman and Tversky 1973), would also predict IE violations, though of a different form.[C] Whatever updating model researchers adopt, the important thing is that it generate testable predictions about how the survey manipulation is likely to affect background beliefs.

Just as observational researchers validate their identification assumptions by conducting placebo tests of effects assumed to be zero (Sekhon 2009), so too should survey experimentalists validate the IE assumption by testing balance on background beliefs across experimental groups. Models of information updating are useful in this regard because they predict which background beliefs are likely to be imbalanced and in what direction, leading to more powerful placebo tests. Presuming that the real-world effect of interest is a well-defined causal effect, the ideal placebo belief is one that is (1) is affected by Z under plausible information-processing models, (2) affects the survey outcome Y, and (3) does not concern an attribute affected by the factor of interest in the real world (for more details, see SI, Appendix C). Since each placebo belief will likely satisfy some of these criteria better than others, we recommend conducting multiple placebo tests, each of which lies on the frontier of this criteria space.

4 Preventing IE Violations

While it is important to diagnose IE violations if they exist, it is better to prevent them to begin with. Here, we discuss three strategies for achieving IE: abstract encouragement, covariate control, and embedded natural experiments. After describing these strategies, we then move to an example in which we compare their performance.

4.1 Abstract Encouragement

Abstract encouragement is our term for asking respondents to consider the scenario or vignette in abstract terms, using a statement such as the following: "For scientific validity the situation is general, and is not about a specific country in the news today" (Tomz and Weeks 2013, 853). The primary argument in favor of abstract designs has been that they can yield more externally valid or generalizable results (Mutz 2011, 158; Tomz and Weeks 2013, 860). But researchers might also expect abstract designs to reduce imbalance on background attributes by encouraging respondents to avoid using real-world data to inform their beliefs about the scenario. Based on the realistic Bayesian model, however, we anticipate that abstract encouragement will not systematically improve balance on background beliefs.

4.2 Covariate Control

The second strategy we consider is what we call *covariate control* (CC), which is both more common and more explicitly aimed at IE than abstract encouragement. To the extent that survey-experimental studies have recognized the importance of IE, they have mainly addressed it through this strategy. In a CC design, the survey vignette includes additional details designed to fix respondents' beliefs about background characteristics that might be correlated with beliefs about the factor of interest. In some studies, the additional details are identical across experimental conditions, but in others the main survey manipulation is crossed with variation in the controls. An especially elaborate form of the latter kind of covariate control is conjoint analysis (Hainmueller, Hopkins, and Yamamoto 2014), a high-dimensional factorial experiment that varies many attributes of the vignette simultaneously.[D]

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

4.3 Embedded Natural Experiments

The third strategy is to employ an *embedded natural experiment* (ENE). This strategy is motivated by the realistic Bayesian model, which predicts that the survey manipulation will influence respondents' beliefs about background attributes unless they perceive the content of the manipulation to be statistically independent of—and thus to convey no information about—those attributes. In other words, a Bayesian will not update their beliefs about background features of the scenario if and only if they believe the causal factor of interest was as good as randomly assigned in the scenario world.

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

The most straightforward ENEs involve a lottery or other form of transparent random process. Consider, for example, a survey experiment that examines whether subsidizing childcare increases employees' willingness to take a time-consuming promotion (Latura 2015). Simply manipulating whether a hypothetical firm is described as subsidizing childcare will probably not isolate the effect of interest because respondents know that some kinds of firms (e.g., ones with a family-friendly culture) are more likely to offer this policy, and these inferences may affect their decision whether to accept the promotion. In the ENE version of this experiment, which we discuss later, the firm is described as having a limited number of subsidized childcare slots that are assigned by a random lottery; the survey manipulation is whether the respondent wins the lottery. Assuming respondents perceive the lottery outcome to be truly random, they should not update their inferences about the background attributes of the firm (but see Section 6.2 for complications that arise in practice).

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

The critical criterion for evaluating ENE designs is not whether the ENE is strictly random, but whether respondents perceive it to be independent of background attributes and update their beliefs accordingly (i.e., information equivalence). As we have described, the IE assumption can be tested empirically using placebo tests. In general, we expect that well-designed ENEs will exhibit less evidence of IE violations than abstract encouragement or CC designs. Unlike CC designs, which should be expected to balance only explicitly controlled attributes and their close relatives, ENEs should balance beliefs about *all* background attributes, regardless of whether they are explicitly controlled. This, of course, is the signal advantage of design-based observational studies over "selection-on-observables" identification strategies. ENE designs, however, are not always easy or even possible to construct. Moreover, the description of the natural experiment may change the treatment and estimand in ways that raise questions of interpretation and generalizability. We discuss these issues further below, but we first turn to empirical examples of IE in survey experiments.

5 An Application to the Democratic Peace

We evaluate evidence of IE violations, and the effectiveness of the various strategies for mitigating them, using several applications.[8] The first and most elaborate is a replication and extension of Tomz and Weeks' (2013) survey experiment on the mass basis for the democratic peace. Using placebo tests, we show that randomly manipulating whether a target country is described as democratic is not sufficient to prevent respondents from updating their beliefs about the background attributes of the country, potentially biasing the effect of interest. We further demonstrate that abstract encouragement does little to mitigate these violations of IE, and that covariate control does so only on attributes explicitly or indirectly controlled in the vignette. An ENE design is most effective at achieving IE. Appendix D in the SI demonstrates the use of an IV estimator to estimate the CACE in this study, and discusses relevant assumptions and issues of interpretation.

5.1 Survey Design

On July 1–3, 2015, we used the Qualtrics survey platform to survey 3,080 Americans recruited through Amazon's Mechanical Turk (MTurk).[9] The basic setup of our survey experiment hewed closely to Tomz and Weeks (2013). We presented respondents with a vignette in which a country is developing nuclear weapons, randomly manipulated whether the country is described as a democracy, and asked whether respondents supported using military force against the country (among other questions). In addition to the main manipulation (democracy/non-democracy), we also varied experimental conditions on two other dimensions designed to assess the effectiveness of different strategies for improving balance on background beliefs. The first dimension was whether respondents were assigned to receive abstract encouragement. The second dimension consisted of three versions of the vignette: a basic vignette that provided

[8] Replication files for all the analyses in this paper can be downloaded from Dafoe, Zhang, and Caughey (2017).

[9] See SI, Appendix F for a complete description of the survey design and SI, Appendix G for the full summary of our analysis. Our study pre-registration and pre-analysis plan can be found at EGAP (http://e-gap.org/design-registration/registered-designs/).

respondents with little information about the country besides the democracy manipulation; a CC vignette that included details about the target country; and an ENE vignette that described an assassination attempt as a source of as-if random variation in regime-type.

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

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

Respondents then read a description of the country's regime-type, randomly manipulated to be democratic or non-democratic:

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

Finally, respondents read the conclusion of the scenario:

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

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

The ENE design began with a description in which the regime type varied as follows:

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

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

After reading the vignette, respondents were asked about their support for using force against the target country, as well as demographic questions and questions related to the placebos, potential mechanisms, and the treatment. We randomized the order of all these questions (we did not detect any relevant question-order effects; see SI, Appendix F). In order to conduct placebo tests, we asked respondents about their beliefs regarding the following background attributes of the target country: *region*, *GDP*, *religion*, *race*, *oil reserves*, *alliance with the US*, *trade with the US*, *joint military exercise with the US*, *FDI in the US*, and *military spending.* All of these variables except the last were selected based on the criteria described in SI, Appendix C: all are at least partly pre-treatment, are correlated with regime-type in the real world, and plausibly affect public support for military action.[F] To minimize the risk of respondents' thinking that these attributes could be affected by democracy in the real-world, the questions asked subjects about the attributes' values ten years in the past.

5.2 Placebo Tests

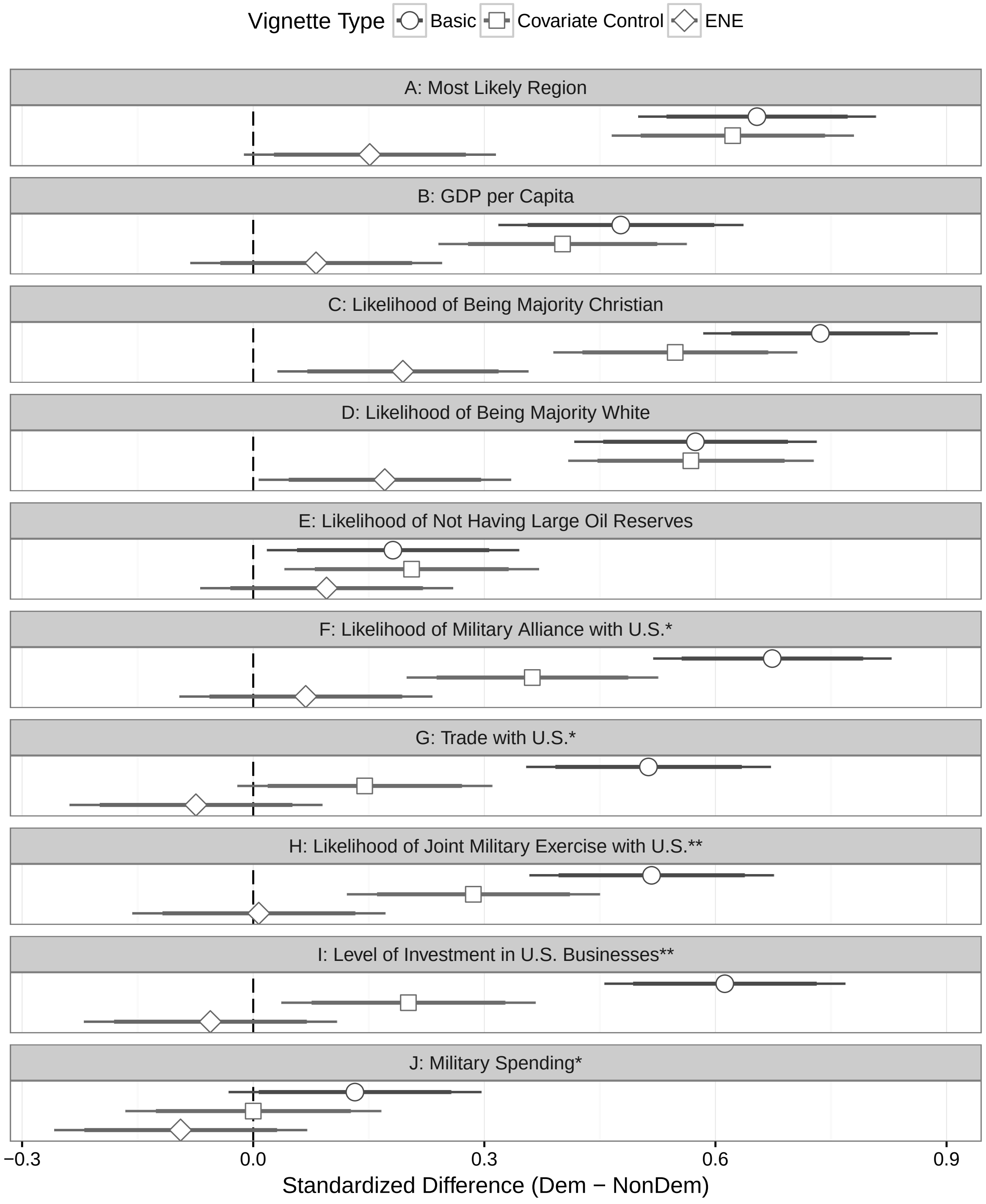
Figure 2 summarizes the main results for the placebo tests (for more details, see SI, Appendix G). They reveal clear evidence of imbalance on attributes, in a manner consistent with the realistic Bayesian model. The imbalance is most pervasive in the basic design: for every placebo variable, mean equality between the two experimental conditions can be rejected at the 5% level, in every case in the direction predicted by the realistic Bayesian model. Subjects who are told that the country is a democracy are more likely to perceive it as having the characteristics associated with democracies in the real world, such as being more likely to have higher GDP per capita, to have populations that are majority Christian and white, to not have large oil reserves, to have an alliance with the US and have conducted a joint military exercise with the US, and to trade with and invest in the US. Across all vignette versions, subjects assigned to receive abstract encouragement exhibited similar imbalance, suggesting that as implemented abstract encouragement is ineffective at achieving IE.

Like the basic design, the CC design exhibits large imbalances on placebo attributes that were not controlled (*region*, *GDP*, *religion*, *race*, and *oil reserves*). On attributes that were explicitly (*alliance* and *trade*) or indirectly (*joint military exercise* and *FDI*) controlled, the imbalance is less extreme, but it was almost never completely eliminated. The CC design did succeed in eliminating imbalance on *military spending*, but even in the basic design this was the least-imbalanced attribute, probably because (as we predicted ex ante) democracy has no clear real-world relationship with military spending.

The ENE design was by far the most effective at reducing imbalance on placebo attributes. For most placebos, the imbalance is much less severe than for the other two designs, and in no case was it detectably worse. Strikingly, even attributes that were *explicitly controlled* in the CC design were more balanced in the ENE design. This result is not a symptom of a weak manipulation, as the ENE manipulation's effect on perceived regime was nearly as great as the other designs (SI, Appendix D). Overall, the results suggest that just as natural experiments, when truly as-if random, tend to yield plausible causal inferences in observational studies, so too are embedded natural experiments a potentially effective strategy for credible causal inference in survey experiments.

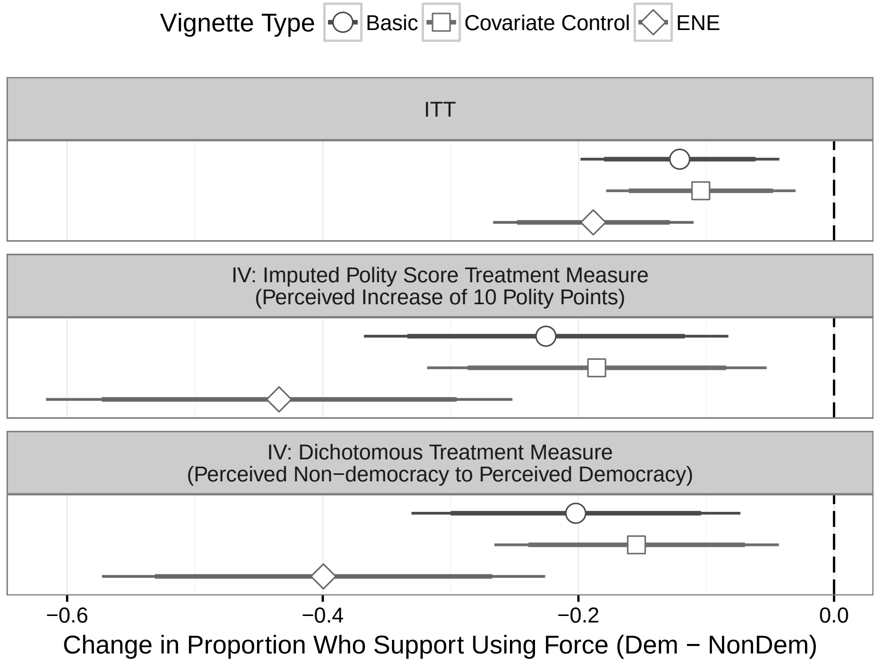
As it happens, in this case an ENE design does not lead to qualitatively different inferences from either a Basic design of the sort employed by Mintz and Geva (1993) or a CC design like that of Tomz and Weeks (2013). As Figure 3 shows, the estimated ITT effect on support for war is slightly larger in the ENE design than the other two designs, and the estimated CACE is even more clearly so. Regardless of the design used, then, the results suggest that believing a

Figure 2: Placebo Tests by Vignette Type



Hollow points indicate the standardized average difference between the democratic and non-democratic treatment conditions. In this coefficient plot and all following ones, we report the 95% and 99% confidence intervals estimated using heteroscedasticity-robust standard errors. Background attributes that were explicitly mentioned in the CC design are indicated with ∗, and ∗∗ indicates attributes implicitly controlled.

Figure 3: Effect Estimates from Different Versions of the Democratic Peace Experiment



hypothetical opponent to be a democracy causes citizens to be less supportive of using military force against it (for more details, see SI, Appendix G).

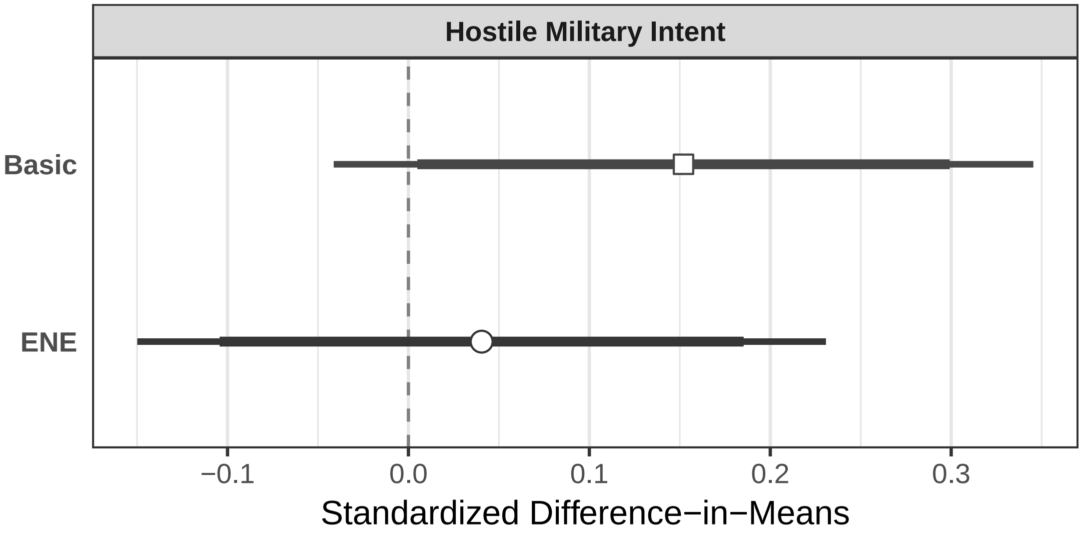
6 Extensions to Other Studies

We have extended these methods to several other studies, three of which we summarize here.

6.1 Effects of Coercive Harm

Dafoe and Weiss (2016) investigate whether harm experienced in a coercive context provokes resolve and desire for retaliation, through survey experiments fielded in China and the United States (SI, Appendix H). The scenario depicts China and the United States engaged in a tense dispute in the East China Sea. In the (American version of the) basic design ($n = 705$), the control describes the dispute, the treatment also describes China shooting down a US military plane for trespassing in Chinese airspace. In the ENE version ($n = 731$) the US plane is made

Figure 4: Balance on Hostile Military Intent (95% and 99% Confidence Intervals)



to crash (or not) in an as-if random way, that is nevertheless part of the coercive context: "the Chinese plane was flying dangerous maneuvers around the American plane, making several close passes. On the third pass the planes [almost collided/collided]. . . ." Figure 4 shows how perceptions of hostile military intent were imbalanced in the basic design, but become balanced in the ENE design. The ENE design thus has a claim to better isolating the causal factor of interest: harm in a coercive context.

6.2 Effects of Subsidized Childcare

Latura (2015) examines whether people are more likely to accept a time-consuming promotion if their firm provides subsidized high-quality extended hours childcare. With Latura, we compared a basic to ENE design (see SI, Appendix I). In the basic design ($n = 771$), after reading about other aspects of their situation and the firm, some subjects were informed that "the company you work at subsidizes the cost of high-quality, extended-hours childcare for employees." The ENE design ($n = 1003$) informed all respondents that their firm operates an "on-site, high-quality, extended-hours day-care center open from 6:00 AM to 10:00 PM on weekdays. The center is free for employees, but slots are allocated via random lottery." The control group was then told that they did not win a day-care slot; the treatment group that they did.[G] Figure 5 shows that all placebo variables are imbalanced in the basic design. The ENE design reduces imbalance relative to the basic one but does not fully eliminate it.

The imbalance in the ENE design suggests that subjects either updated their beliefs in non-Bayesian fashion or did not believe that the lottery was random with the same probability. One subtle possibility is that since we did not specify the subject of winning the lottery, a respondent in control could reasonably infer that there were only a few spots allocated by lottery (e.g., the lottery was a public relations stunt), whereas a respondent in treatment could infer that many spots were allocated. If this was the problem, then it reveals how careful one must be in constructing an ENE to make sure the respondents not only perceive treatment to be as-if random, but as-if random with the same probability across conditions.

Figure 5: Placebo Test Results from Latura (2015)

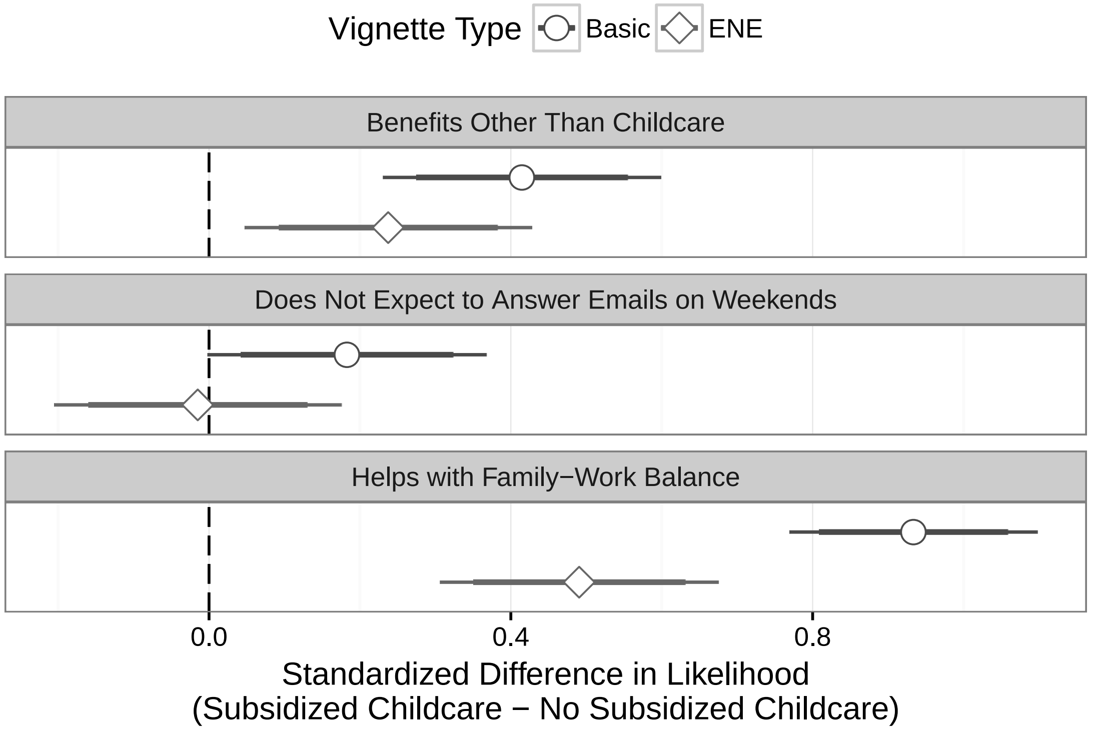
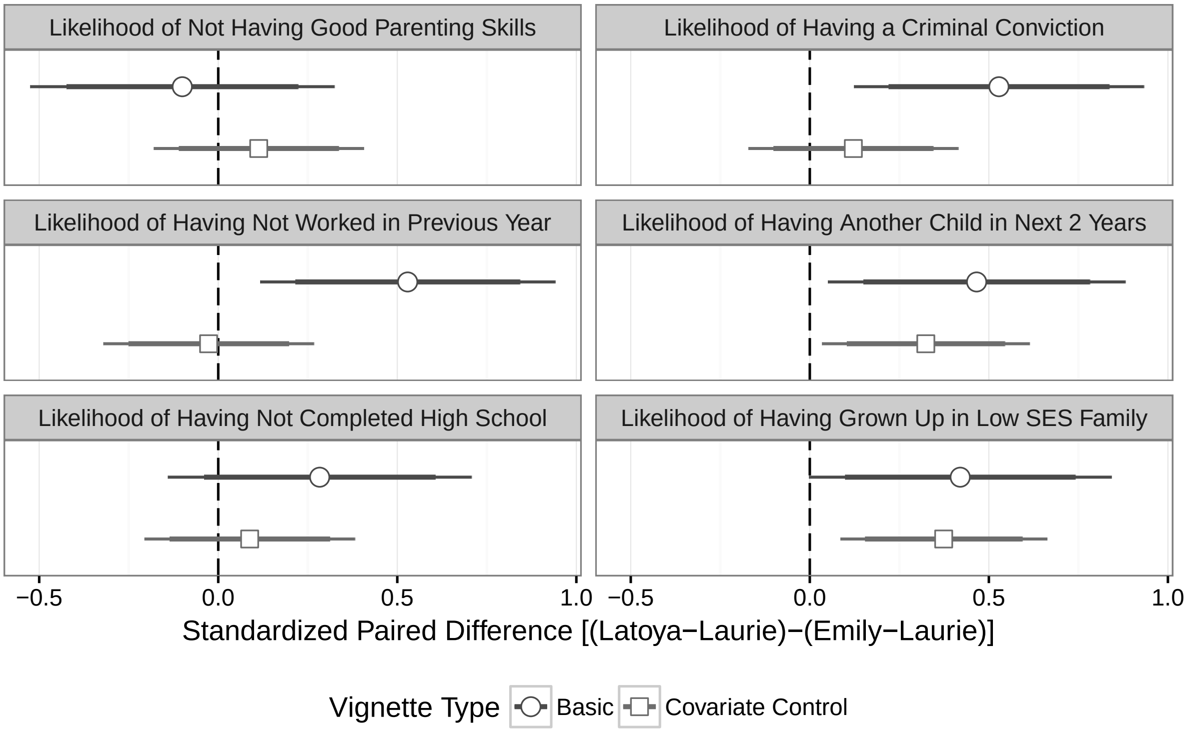


Figure 6: Placebo Test Results from Replication and Expansion of DeSante (2013)



6.3 Why is Latoya Discriminated Against?

Finally, we replicated and extended Desante's (2013) study of whether and why Americans are more willing to support welfare for people who are white than black. Our basic design manipulates the name of the welfare applicant (e.g., Emily vs. Latoya), and holds constant the number and age of the applicant's children. Following Desante (2013), our CC design additionally includes a "Worker Quality Assessment" (with values of "Poor" or "Excellent"). In so doing, the CC design hopes to rule out "principled conservative" reasons for discrimination, leaving only "racial animus" as the basis for discrimination. For placebo questions, we sought characteristics that would be the basis for "principled conservative" discrimination, which led us to a set of questions from the North Carolina welfare agency.[H] We also included two additional questions: whether the applicant grew up in a low socioeconomic status (SES) family and whether the applicant is likely to have another child in the next two years.[I]

Figure 6 (full details in SI, Appendix J) shows how the CC design ($n = 312$) reduced imbalance relative to the basic design ($n = 156$), though there was still imbalance on their socio-economic status as a child and probability of having a child in the future. These results suggest that DeSante's control strategy successfully reduced imbalance on most characteristics that a "principled conservative" might discriminate on (prior work experience, criminal conviction), but not on all. Thus, while the results in Desante (2013) do provide insight into the reasons for racial discrimination, caution is still required before accepting this as definitive evidence of racial animus.

Studies of racial cues raise subtle issues about the causal estimand (Sen and Wasow 2016). This is revealed by asking what would an ENE design look like if we wanted to manipulate subjects' perception of someone's race. It is difficult to imagine a process that as-if randomly assigns race, independent of "background characteristics," in large part because race is not a clearly defined manipulable trait.[J] One alternative way forward is to define the treatment as an informational cue that signals a person's race and attempt to decompose the mechanisms by which this cue affects the outcome (Acharya, Blackwell, and Sen 2017). This approach entails providing information regarding potential mediators, with the goal of fixing those mediators and identifying the controlled direct effect of the cue (e.g., the portion not mediated through a principled conservative basis for discrimination). While this strategy is potentially promising, Acharya, Blackwell, and Sen (2017, 9) note that estimating the controlled direct effect relies on exclusion restrictions similar to the IE assumption in that they require informational manipulations to affect only the mediator of interest.

7 Limitations of Different Designs

The evidence from the preceding studies suggests that ENE designs, when feasible, are generally the best strategy for promoting information equivalence. Nevertheless, it is also clear that neither the CC nor the ENE design is a panacea, and the most effective design depends on the study and its QOIs. Below, we discuss the limitations of each design in turn.

7.1 Limitations of Covariate Control

While covariate control did not achieve IE in our examples, it did reduce imbalance on those background features that were specified. Is the solution then simply to devise extremely detailed scenarios that specify every possible background feature? One potential problem with this strategy is respondent exhaustion or satisficing (Krosnick 1999; but see Bansak et al. 2017). A more fundamental limit, however, is what might termed the *plausibility constraint*: as the number of controls increases, so too does the probability of a vignette that is implausible to respondents. As in observational studies, the more variables we control for, the more likely it is for a counterfactual to go beyond the support of the data (King and Zeng 2006). There is, for example, simply no empirical referent for a Western European democracy that uses Sharia law for criminal proceedings. Researchers can prune away implausible vignettes (Hainmueller, Hopkins, and Yamamoto 2014, 20), but as the number of control variables increases the subset of plausible combinations will tend to become smaller.

A related strategy is to use a *proper-noun vignette*: specifying a real-world referent in the scenario, thus implicitly controlling for an almost infinite number of background attributes. For example, in another survey experiment we provided selective information about a country's past foreign-policy behavior; in one version the country was identified as Iran. We found that naming the country reduced imbalance on background attributes such as the country's regime type, but it did not eliminate it. This is likely to always be the case since the vignette is a hypothetical, allowing for the possibility of other unspecified changes in background covariates. Violations of IE will likely be more severe the less respondents know about the real-world referent, since then respondents will infer more about background features from the treatment prompt (Z). A variant of this strategy that avoids hypotheticals, which we label a *selective-history design*, entails selectively informing or reminding the respondent about certain facts of a historical episode. Such a strategy has promise, but it is limited by the kinds of scenarios generated by the real world and can still induce IE violations if respondents are not perfectly informed about history.

Covariate control can create or amplify bias from IE violations as well as reduce it. The most obvious way it can do so is, for a realistic Bayesian respondent, by unintentionally controlling for real-world consequences of the factor of interest, leading to biases akin to selection bias in observational studies. Further, as in observational studies (Middleton et al. 2016), controlling for even pre-treatment background characteristics can amplify bias in survey-experimental estimands. For intuition on this point, consider a covariate-control version of the democratic peace experiment that specifies that the scenario takes place in the Middle East. For realistic Bayesian subjects, this geographic control would increase imbalance on beliefs regarding religion because the negative correlation between democracy and being majority-Muslim is even stronger in the Middle East than in the world as a whole. In short, covariate control provides no general solution to the problem of IE violations.

7.2 Limitations of Embedded Natural Experiments

ENEs have their own limitations. First, just as valid real-world natural experiments are hard to find, so is it hard to construct plausible ENEs that generate large enough effects on treatment (IV bias being larger for weak instruments). For example, we could not think of a plausible strong natural experiment for which the "democracy" level would be a country like Belgium and the "non-democracy" level a country like Egypt (let alone North Korea), because the real world has not produced and is not likely to produce such interventions. This limitation can be understood as an instance of the plausibility constraint.

A second concern is that ENE designs only allow us to estimate a narrow estimand—the effects for a narrowly defined set of scenarios—and not the general causal estimand the researchers may have had in mind. In the democratic peace study, our ENE only allows us to estimate the effect for the kinds of countries and manipulations that fit the ENE scenario. This is a *local* causal effect because it captures the average effect among only those countries that fit the ENE. Consistent with this concern, our respondents report perceiving the ENE scenario to be much more typical of the Middle East and North Africa than Western Europe (see SI, Appendix G, Figure 12). Relatedly, the effects identified by ENE may be specific to the particular manipulation. For example, if Americans are especially concerned about leaders of fragile democracies, our democratic peace ENE estimand (the effect of democracy in countries prone to coup attempts) may be distinctly different from that in other kinds of countries. It is even possible for ENEs to introduce their own IE violations if the cause of interest in the scenario is "bundled" with other causes. For example, it may be that surviving an assassination attempt makes the surviving leader more sympathetic to subjects, in addition to changing respondents' beliefs about the country's regime type.

One way to mitigate the limitations of ENE designs is to employ several distinct ENEs. For example, to address the above concern that sympathy or a related mechanism (independent of regime type) accounts for the assassination results, we produced a version of the ENE in which the assassination attempt was against a dictator, and when successful led to democratization, inverting the effect of assassination on regime type. The effect of Z on Y similarly flipped, leading to a similar (slightly larger) estimate of the effect of D (the CACE; see SI, Appendix G, Figure 32). In any case, researchers employing ENEs should explicitly discuss how the distinctiveness of the ENE manipulation and the "localness" of the estimand qualifies the interpretation of their findings.

Although CC designs may seem to avoid the localness limitations of ENEs, they probably do not. After controlling for sufficient details and retaining only plausible combinations, a superficially general scenario will, in fact, be restricted to a limited region of the covariate space, namely the space for which there is variation in the treatment (cf. Aronow and Samii 2016). If the covariates are sufficient to identify the effect of D, then the scenario will be limited to comparisons in which the causal factor of interest is independent of background causes of Y. We see this in our study: as in the ENE design, respondents in the basic and CC designs found the scenario much more typical of the Middle East and North Africa than Western Europe (SI, Appendix G, Figure 12). Though CC scenarios may seem abstract and general, if they have enough detail to control important background characteristics then respondents are likely to be drawing strong inferences about the kinds of units in the scenario. The same problem arises if we use proper nouns ("Iran") in our CC design since we will be restricted to those proper nouns for which both counterfactuals are somewhat plausible. Localness, in short, may be a fundamental feature of survey experiments.

8 Conclusion and Recommendations

A well-implemented experiment allows us to identify the causal effect of that which was randomly assigned. But we usually want to go beyond that to identify the effect of some specific causal factor: the active drug in a medicine, not a placebo effect induced by the pill itself. To do so we must *assume* that the experimental effect only operates through the intended causal channel. Assumptions, however, can and should be tested. The results of these placebo tests can then be used to improve experimental design.

In this paper we did this for scenario-based survey experiments: articulating the necessary assumptions, theorizing how they are likely to be violated, examining their testable implications, and evaluating the performance of several experimental designs. We found that IE violations are common. Further, we showed how respondent updating has a specific structure which we can use to anticipate and prevent IE violations. In some respects, the nature of the problem and solutions bears a close similarity to the problem of and solutions for confounding in observational studies. Best practice for survey experiments accordingly resembles best practice for observational studies. Specifically, we recommend the following:

1. **State your QOI and theorize about information equivalence.** Think clearly about what real-world counterfactual you are trying to reproduce. What set of background characteristics need to be held fixed for this to succeed? What background characteristics are correlated in the real-world with treatment, and thus are most at risk of being influenced by your survey manipulation?

2. **Measure your causal factor.** This can be used to evaluate the assumption of a monotonic first stage, to estimate complier average treatment effects, and to understand the kinds of variation in D that are informing your estimates.

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

4. **Control covariates.** If you can't employ an embedded natural experiment, employ covariate control designs to rule of at least some sources of IE violations.

5. **Diagnose violations of IE.** Employ placebo tests to evaluate whether IE seems plausible, and if not, why not.

6. **Theorize the bias.** Think through, informally or formally, the direction and size of biases likely to come from the violations of IE. A causal estimate will be more compelling if you can persuasively argue that the bias is likely to be small or in the opposite direction as your prediction.

7. **Qualify your inferences.** Acknowledge the remaining risk of violations of IE. Recognize that your estimated causal effects are local to the kinds of scenarios that you presented and the respondents' inferences about the context of the scenario.

Survey experiments are valuable tools for social science. They permit the study of important causal questions that are otherwise elusive. But random assignment alone does not free scholars from the need to think carefully about identifying their quantities of interest.

References

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2017. "Analyzing Causal Mechanisms in Survey Experiments." Unpublished manuscript, March 30. Accessed July 20, 2017. http://www.mattblackwell.org/files/papers/survey-experiments.pdf.

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

Aronow, Peter M., and Cyrus Samii. 2016. “Does Regression Produce Representative Estimates of Causal Effects?” *American Journal of Political Science* 60 (1): 250–267.

Baker, Andy. 2015. “Race, Paternalism, and Foreign Aid: Evidence from US Public Opinion.” *American Political Science Review* 109 (1): 93–109.

Bansak, Kirk, Jens Hainmueller, Daniel J. Hopkins, and Teppei Yamamoto. 2017. “Beyond the Breaking Point? Survey Satisficing in Conjoint Experiments.” MIT Political Science Research Paper No. 2017-16, April 26. http://dx.doi.org/10.2139/ssrn.2959146.

Barabas, Jason, and Jennifer Jerit. 2010. “Are Survey Experiments Externally Valid?” *American Political Science Review* 104 (2): 226–242.

Brader, Ted, Nicholas A Valentino, and Elizabeth Suhay. 2008. “What Triggers Public Opposition to Immigration? Anxiety, Group Cues, and Immigration Threat.” *American Journal of Political Science* 52 (4): 959–978.

Butler, Daniel M., and Jonathan Homola. 2017. “An Empirical Justification for the Use of Racially Distinctive Names to Signal Race in Experiments.” *Political Analysis* 25 (1): 122–130.

Butler, Daniel M., and Eleanor Neff Powell. 2014. “Understanding the Party Brand: Experimental Evidence on the Role of Valence.” *Journal of Politics* 76 (2): 492–505.

Chong, Dennis, and James N. Druckman. 2010. “Dynamic Public Opinion: Communication Effects over Time.” *American Political Science Review* 104 (4): 663–680.

Dafoe, Allan, and Jessica Chen Weiss. 2016. “Provocation, Nationalist Sentiment, and Crisis Escalation: Evidence from China.” Unpublished working paper, June 6. https://www.dropbox.com/s/tb9pgzzkjkef6wg/16-06-06_provocation.pdf?dl=1.

Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2015. “Confounding in Survey Experiments.” Paper presented at the Annual Meeting of The Society for Political Methodology, University of Rochester, Rochester, NY, July 23.

———. 2017. “Replication Data for: ‘Information Equivalence in Survey Experiments’.” Harvard Dataverse. doi:10.7910/DVN/KVZXE8.

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.

Dunning, Thad. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach.* New York: Cambridge.

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

Gilens, Martin. 2002. “An Anatomy of Survey-Based Experiments.” In *Navigating Public Opinion: Polls, Policy, and the Future of American Democracy,* edited by Jeff Manza, Fay Lomax Cook, and Benjamin I. Page, 232–250. New York: Oxford.

Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto. 2015. “Validating Vignette and Conjoint Survey Experiments Against Real-World Behavior.” *Proceedings of the National Academy of Sciences* 112 (8): 2395–2400.

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.

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.

Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011. “Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies.” *American Political Science Review* 105 (4): 765–789.

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.

Kahneman, Daniel, and Amos Tversky. 1973. “On the Psychology of Prediction.” *Psychological Review* 80 (4): 237–251.

Kertzer, Joshua D., and Ryan Brutger. 2016. “Decomposing Audience Costs: Bringing the Audience Back Into Audience Cost Theory.” *American Journal of Political Science* 60 (1): 234–249.

King, Gary, and Langche Zeng. 2006. “The Dangers of Extreme Counterfactuals.” *Political Analysis* 14 (2): 131–159.

Krosnick, Jon A. 1999. “Survey Research.” *Annual Review of Psychology* 50 (1): 537–567.

Latura, Audrey. 2015. “Material and Normative Factors in Women’s Professional Advancement: Experimental Evidence from a Childcare Policy Intervention.” Paper presented at the American Politics Research Workshop, Harvard University, April 28. http://lists.fas.harvard.edu/pipermail/gov3004-list/attachments/20150427/ea95d274/attachment-0001.pdf.

Marsden, Peter V., and James D. Wright. 2010. *Handbook of Survey Research.* Bingley, UK: Emerald Group Publishing.

Middleton, Joel A., Marc A. Scott, Ronli Diakow, and Jennifer L. Hill. 2016. “Bias Amplification and Bias Unmasking.” *Political Analysis* 24 (4): 307–323.

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

Mutz, Diana C. 2011. *Population-Based Survey Experiments.* Princeton, NJ: Princeton.

Sekhon, Jasjeet S. 2009. “Opiates for the Matches: Matching Methods for Causal Inference.” *Annual Review of Political Science* 12 (1): 487–508.

Sen, Maya, and Omar Wasow. 2016. “Race as a ‘Bundle of Sticks’: Designs that Estimate Effects of Seemingly Immutable Characteristics.” *Annual Review of Political Science* 19:499–522.

Sher, Shlomi, and Craig R. M. McKenzie. 2006. “Information Leakage from Logically Equivalent Frames.” *Cognition* 101 (3): 467–494.

Sniderman, Paul M., and Douglas B. Grob. 1996. “Innovations in Experimental Design in Attitude Surveys.” *Annual Review of Sociology* 22:377–399.

Tomz, Michael, and Jessica L. Weeks. 2013. “Public Opinion and the Democratic Peace.” *American Political Science Review* 107 (4): 849–865.

Supplementary Information for:
Allan Dafoe, Baobao Zhang, and Devin Caughey,
"Information Equivalence in Survey Experiments"

Contents

A Appendix Notes

This section contains the notes referenced in the main text with superscript capital letters.

[A] Formally our model of Bayesian respondents implies:

$$p(D = d_k, B = B_k|Z = z) = \frac{p(Z = z|D = d_k, B = b_k)p(D = d_k, B = b_k)}{p(Z = z)}$$

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

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

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

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

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

[F] The exception is military spending, which is not significantly related to regime-type in the real-world; we tested this variable because the vignettes in Tomz and Weeks (2013) control for nonnuclear military capabilities.

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

[H] See North Carolina’s Work First Family Assessment of Strengths and Needs Form (DSS-5298, rev. 05/13, Economic and Family Services). http://info.dhhs.state.nc.us/olm/forms/dss/dss-5298-ia.pdf.

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

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

B The Survey Manipulation as an Instrumental Variable

Section 2 of the main text makes the simplifying assumptions that (1) the real-world background features B^* held constant across counterfactual comparisons are pre-treatment attributes unaffected by the factor of interest D^*, and (2) survey subjects' beliefs about background features B are unaffected by beliefs about the factor of interest D. As they pertain to the survey, these assumptions are encoded in the left panel of Figure 7, labeled "B precedes D." Under these assumptions, both the real-world and survey QOIs are simply the total effects of D^* on Y^* and D on Y.

In many cases, however, the real-world QOI is (implicitly, at least) not the total effect of D^* but the effect with some post-treatment attribute held constant. Tomz and Weeks (2013), for example, are interested in the real-world effect of democracy, but their scenario holds constant levels of trade with the United States, which could very well be a consequence of democracy. Similarly, Desante (2013) is interested in the real-world effect of (perceptions of) welfare applicants' race, holding constant (perceptions of) characteristics related to "principled conservatism," which may themselves be affected by (perceptions of) applicants' race. Even in cases where the real-world B^* is clearly pre-treatment, it is frequently implausible to assume that survey subjects *beliefs* about B^* (that is, B) are formed entirely before beliefs about D^* (D). In all of these cases, then, we are interested in the direct effects of the (beliefs about the) factor of interest not mediated through (beliefs about) background characteristics.

To account for the possibility of post-treatment B^* and B, we therefore generalize our notation to allow these quantities to be affected by D^* and D. That is, we write $B^*(D^* = d)$ for the potential outcome of B^* with D^* set to d and $B(D = d)$ for the potential outcome of B with D set to d. With this notation, the real-world effect of interest becomes

$$\delta_s^* \equiv Y_s^*(D_s^* = 1, B_s(D_s = d)) - Y_s^*(D_s^* = 0, B_s(D_s = d)) \tag{7}$$

for some reference level of treatment d (contrast with τ_s^* defined in equation (1) in the main text). Analogously, the epistemic effect of interest is

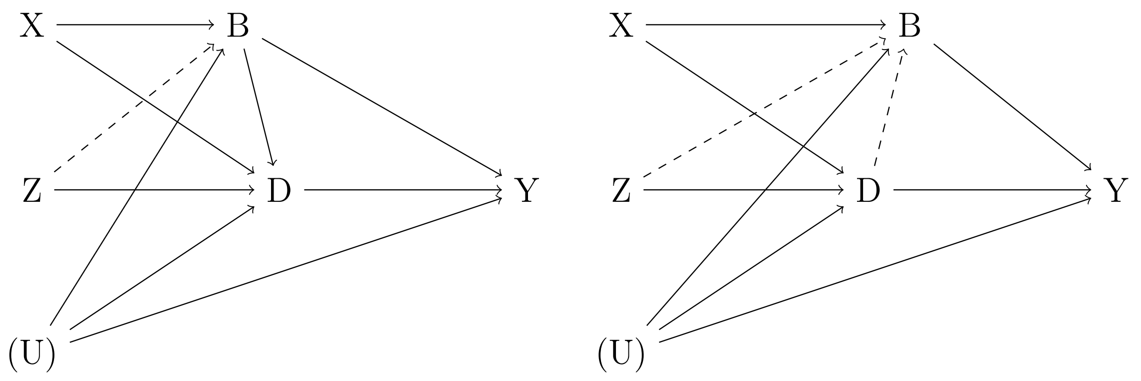
$$\delta_i \equiv Y_i(D_i = 1, B_i(D_i = d)) - Y_i(D_i = 0, B_i(D_i = d)). \tag{8}$$

Both δ_s^* and δ_i are natural direct effects (NDEs): the difference in potential outcomes across treatment and control, with the mediator for each unit set to the level it would naturally take under treatment level d.[10]

[10] It is not always clear whether researchers' QOI is the NDE or the controlled direct effect

$$\text{CDE} = Y_s^*(D_s^* = 1, B_s^* = b) - Y_s^*(D_s^* = 0, B_s^* = b),$$

which differs from the NDE in that it fixes the mediator to a specific value (VanderWeele 2015; Acharya, Blackwell, and Sen 2017). The CDE is the portion of the treatment effect due neither to mediation nor interaction with the mediator. The assumptions under which the CDE is identified, though weaker than those required for identification of the NDE, are in our view unlikely to be satisfied in survey experiments where the IE assumption (which identifies both the NDE and CDE) fails.



(a) B precedes D

(b) D precedes B

Figure 7: Information equivalence when B is pre-treatment (left) and post-treatment (right). Dashed paths indicate causal relationships ruled out by the IE assumption. Under either graph, if the dashed lines are absent, then $B \perp\!\!\!\perp Z$.

As discussed in the main text, under the standard IV assumptions (see Figure 7, left), including especially the exclusion restriction that Z affects Y only through D,[11] the complier average causal effect (CACE) is consistently estimated by the Wald estimator

$$\widehat{CACE} \equiv \frac{(\bar{Y}|Z=1) - (\bar{Y}|Z=0)}{(\bar{D}|Z=1) - (\bar{D}|Z=0)}. \tag{9}$$

Despite the exclusion restriction, the CACE does not necessarily hold B_i fixed across potential outcomes because D may affect B. Thus, if B cannot be assumed to temporally precede D, we require an alternative assumption—that D, even though it precedes B, does not affect B (see Figure 7, right). If this assumption holds in addition to the standard IV assumptions, the complier average NDE (CANDE; Frölich and Huber 2017) is equal to the CACE and can thus be estimated with standard IV methods:

$$\begin{aligned}
\text{CANDE} &\equiv \text{E}[\delta_i | D_i(Z_i=1) - D_i(Z_i=0) = 1] \\
&= \text{E}[(Y_i(D_i=1, B_i(D_i=d)) - Y_i(D_i=0, B_i(D_i=d))) \\
&\qquad\quad |D_i(Z_i=1) - D_i(Z_i=0) = 1] \\
&= \text{E}[(Y_i(D_i=1) - Y_i(D_i=0)) \\
&\qquad\quad |D_i(Z_i=1) - D_i(Z_i=0) = 1] \\
&\equiv \text{CACE}.
\end{aligned} \tag{10}$$

Further, if we allow for multivalued treatments, then $\widehat{CACE}$ is a consistent estimator of what Angrist and Imbens (1995) call the "average causal response" (ACR), which is a compliance-weighted average of subject-level effects of D on Y. Finally, even if D is unobserved, we can still make inferences about the sign of the CANDE because it is guaranteed to have the same

[11] As in the main text, we presume throughout that B may possibly affect Y not through D. If such effects can be assumed away, information equivalence is not necessary for IV estimation, but in general, this assumption is neither plausible nor easily testable.

sign as the intent-to-treat (ITT) effect,

$$\text{ITT} \equiv \mathbb{E}[Y_i(Z_i = 1) - Y_i(Z_i = 0)]. \tag{11}$$

In other words, the standard IV assumptions plus the additional assumption that D does not affect B ensure that the $CACE$ equals $\text{E}[\delta_i]$, the NDE of D not mediated through B. This, in turn, justifies making inferences about the distribution of δ_i, and by extension its real-world counterpart δ_s^*, from the estimable quantities $\widehat{ITT}$, $\widehat{CACE}$, and $\widehat{ACR}$. Together, the restrictions that Z does not affect B, and that D does not affect B, jointly imply information equivalence of Z with respect to background features of the scenario, which has the testable implication $B \perp\!\!\!\perp Z$ (both of the causal graphs in Figure 7 have this implication).

C Placebo Tests

A good placebo attribute satisfies three criteria:

- $\mathcal{C}_1$: **The respondent does not believe that the placebo attribute is affected by the factor of interest.** If the respondent believes that the placebo attribute is influenced by the factor of interest, then rejection of the placebo test does not necessarily indicate violation of the exclusion restriction. Rather, it could reflect the fact that the placebo attribute is part of the causal pathway of interest. The easiest way to satisfy this criterion is for the placebo attribute to be "pre-treatment" *in scenario time* since we assume that respondents know that causality only flows forward in time.

- $\mathcal{C}_2$: **If information equivalence is violated, the manipulation Z will affect respondent beliefs about the placebo attribute.** $\mathcal{C}_2$ states that if the experimental manipulation affects the survey responses Y through causal pathways that don't include D, then Z should affect the placebo. The easiest way to satisfy this criterion is to find attributes on the confounding causal pathways. Under our model of Realistic Bayesian beliefs, these will be factors that are correlated with the factor of interest in the real world. Specifically, we screened our candidate placebo characteristics based on whether they are correlated with regime-type in the real world (see Table 3 in the Supplementary Appendix).[12]

- $\mathcal{C}_3$: **Beliefs about the placebo attribute affect the subject's response Y.** The exclusion restriction stipulates that the instrument affects the outcome only through the causal factor of interest. A placebo attribute that does not affect relevant outcomes cannot confound the treatment effect, even if beliefs about the attribute are indeed influenced by the manipulation, not through treatment. Absent an alternative causal pathway between Z and Y, Z remains a valid instrument for D (given assumptions $\mathcal{A}_1$–$\mathcal{A}_3$).

A placebo test based on an attribute that satisfies criterion $\mathcal{C}_1$ will be statistically valid: when there is information equivalence it will not reject more than the size of the test. A test based on an attribute that satisfies criteria $\mathcal{C}_2$ and $\mathcal{C}_3$ will be statistically *powerful*: when there is information equivalence violation it will be likely to reject. The most informative placebo tests will be both valid and powerful.

For example, many of the most salient real-world confounders of the democratic peace, such as countries' wealth, alliances, and trade relationships, are at least potentially affected by regime-type ($\mathcal{C}_1$). On the other hand, some attributes that are clearly not influenced by regime-type, such as geographic region, may not have a strong effect on public support for war ($\mathcal{C}_3$). And some causes of support for war, such as the target country's military spending, are not strongly associated with democracy ($\mathcal{C}_2$). Negotiating the trade-offs between these criteria requires careful *ex ante* theorizing as well as suitable caution in interpreting the results. Nevertheless, to the extent that they satisfy these conditions, placebo tests are a powerful tool for diagnosing whether the exclusion restriction is violated and thus inferences are confounded.

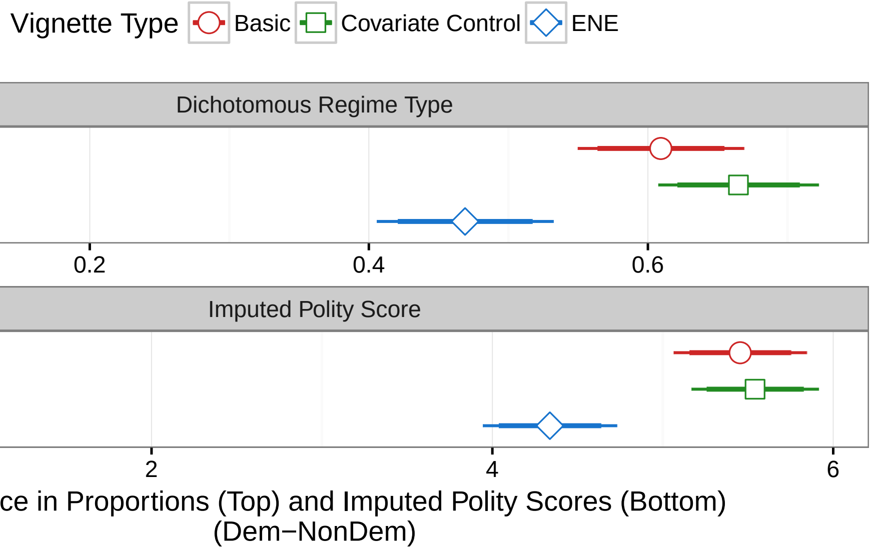
[12] The URL for the online-only Supplementary Appendix is https://github.com/yaleirsurveys/information_equivalence.

D Local Average Treatment Effects

The three vignette manipulations had broadly similar effects on support for the use of force: in all three, respondents told that a country was democratic were 10 to 20 percentage points less likely to support a military attack on the target country (Figure 9, top panel). Under the IV assumptions, we know that this ITT estimate has the same sign as the local average treatment effect among those whose beliefs were influenced by the survey manipulation, a.k.a., the complier average causal effect (CACE). Setting aside the foregoing concerns about the exclusion restriction, we now show how to estimate the CACE directly.

Unbiased estimation of the CACE requires that the treatment be measured without error, a critical and often non-trivial assumption. Since democracy is a complex and nuanced concept, we used a rich set of questions to measure respondents' beliefs about it. One subset of these questions asked respondents how likely it was that the target country fell into each of five Polity scale categories, ranging from fully non-democratic to fully democratic. We summarize these questions in two ways: with a dichotomous indicator D_D for whether the respondent assigned greater probability to democratic than non-democratic categories, and with a continuous measure D_C consisting of the probability-weighted average of the mean Polity score of countries in each discrete category. We also included a second battery of questions asking whether respondents thought there was "more than a 50 percent chance" that the country exhibits various components or indicators of democracy, such as an elected government, a free press, or legal opposition parties (Appendix F.7, G.5.2).

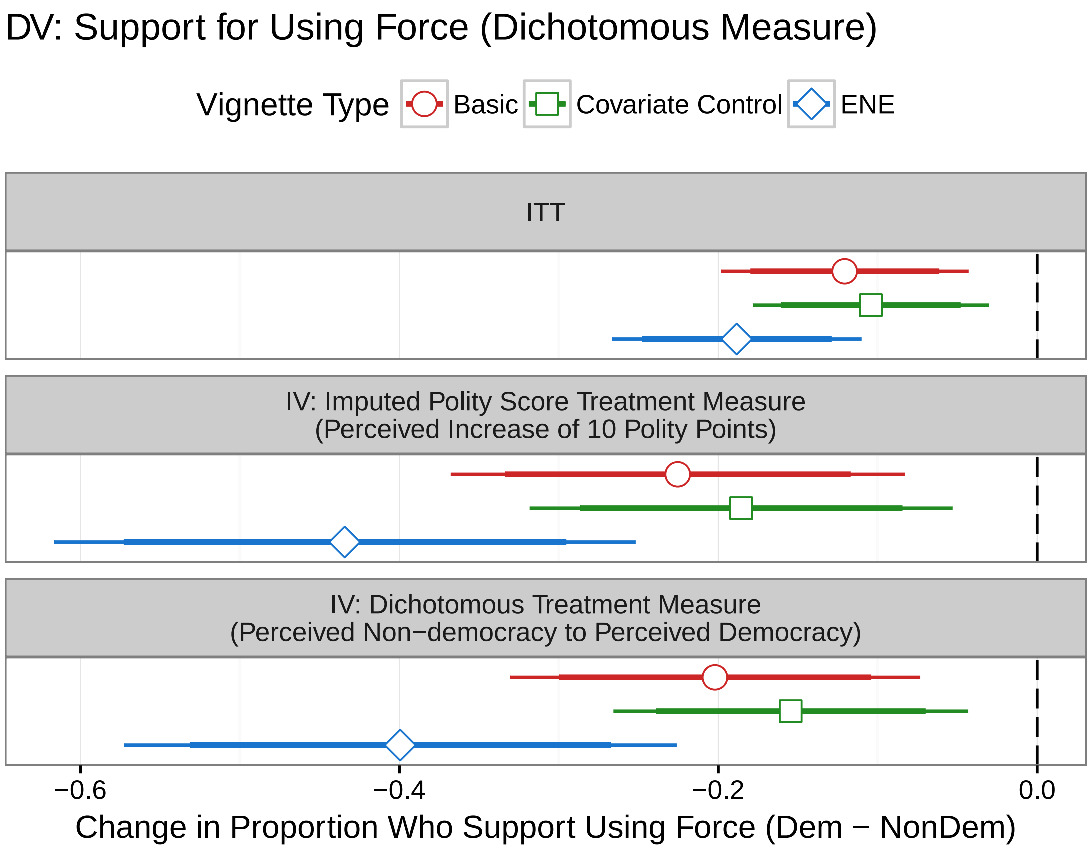
Figure 8: First Stage of the IV Analysis



The top panel estimates the survey manipulation's effect on a dichotomous democracy measure (D_D), and the bottom estimates the effect on the imputed Polity score (D_C).

Regardless of how D is measured, Z had a strong "first stage" effect (Figure 8 and Appendix G.5.2). This effect, however, is substantially smaller than a literal interpretation of the regime-type prompt would suggest, largely because respondents in the "democracy" condition did not perceive the target country to be especially democratic. On average, respondents in this condition gave the country an imputed Polity score of around $+3$ on a -10 to $+10$ scale, a Polity score representative of countries like Russia and Iraq. It is likely the case that respondents perceived the "democracy" to be not very democratic because their beliefs were influenced by the other information in the scenario, such as the fact that the country was developing nuclear weapons in a threatening manner. As a consequence, stating that the target "is a democracy and shows every sign that it will remain a democracy" increased the probability of $D_D = 1$ by only around 50 percentage points and D_C by around 5 points on the 21-point Polity scale, in both cases likely much less than the intended counterfactual.

Figure 9: ITT and CACE Estimates



The dependent variable is a dichotomous measure for support for using force. The ITT estimate is the average effect of assignment to the democracy condition. The CACE estimate is the average effect of perceiving the target country to be a democracy, defined continuously (middle panel) or dichotomously (bottom).

As a consequence of respondents' incomplete "compliance" with their assigned regime-type, the estimated (complier average) causal effect (CACE) is about twice as large as the ITT for all three designs. The CACE for the ENE design is largest: among those affected by the manipulation, believing the country was democratic ($D_D = 1$) decreased respondents' probability of supporting use of force by an estimated 40 percentage points; we estimate a similarly large CACE for a change in the perceived democracy level by 10 points on the Polity scale. The treatment effect estimates for the basic and covariate control designs are significantly smaller, about 20 percentage points, but it is unclear whether this reflects true differences in the LATE or the contaminating influence of information equivalence violation.

E Literature Review

To read our extensive literature review of survey experiments in top political science journals, please see Section A of our Supplementary Appendix.

F “Democratic Peace” Survey Experiment Details

F.1 Outline of the Survey

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

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

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

We experimentally vary the order of the five blocks in the Survey Questions section:

A Placebo Test: Open-ended response

B Placebo Tests: Multiple choice

C Treatment Measure

D Plausibility Check

E Support for Using Force, Mediation Questions

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

F.2 Three Vignette Types

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

F.2.1 Basic

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

[The country is **not a democracy** and shows no sign of becoming a democracy./The country **is a democracy** and shows every sign that it will remain a democracy.]

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

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

F.2.2 Covariate Control

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

- The country [is **not a democracy** and shows no sign of becoming a democracy./**is a democracy** and shows every sign that it will remain a democracy.]
- The country [**has not/has**] signed a **military alliance** with the U.S.
- The country has [**low/high**] levels of **trade** with the U.S.
- The country's nonnuclear military forces are **half as strong** as the U.S.'s nonnuclear forces.

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

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

F.2.3 Embedded Natural Experiment

Embedded Natural Experiment Fragile Democracy (ENEd)

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

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

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

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

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

Embedded Natural Experiment Fragile Non-democracy (ENEn)

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

Two years ago at a public event, a pro-democracy rebel shot at the dictator of Country A. [**The dictator was hit in the head and did not survive the attack.** In the political vacuum that followed, pro-democracy protestors took to the streets and forced those in the former dictator's government to resign. **Soon after Country A held national elections and it is still a democracy today.**/The dictator was hit in the shoulder and survived the attack. **The dictator's regime survived the political turmoil. Today, Country A is still a dictatorship.**]

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

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

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

F.3 Support for Force and Mediation Questions

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

A Support for Using Military Force

B Mechanisms 1: consequences if military action is taken

C Mechanisms 2: consequences if military action is not taken

D Mechanism 3: the morality of military action

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

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

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

1. Support for Force, GDP per Capita, Religion, Regions, All Other Placebo Test Questions
2. GDP per Capita, Support for Force, All Other Placebo Test Questions
3. Regions, Support for Force, All Other Placebo Test Questions
4. Religion, Support for Force, All Other Placebo Test Questions
5. GDP per Capita, Regions, Support for Force, All Other Placebo Test Questions
6. GDP per Capita, Religion, Support for Force, All Other Placebo Test Questions
7. Regions, Religion, Support for Force, All Other Placebo Test Questions
8. GDP per Capita, Regions, Religion, Support for Force, All Other Placebo Test Questions

F.5 Survey Questions

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

F.6 Placebo Test Questions

F.6.1 Justifications for Placebo Test Questions

To read about our justifications for the placebo test questions used, please see Section B of our Supplementary Appendix.

F.6.2 Notes on Placebo Test Questions

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

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

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

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

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

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

F.6.3 Text of Placebo Test Questions

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

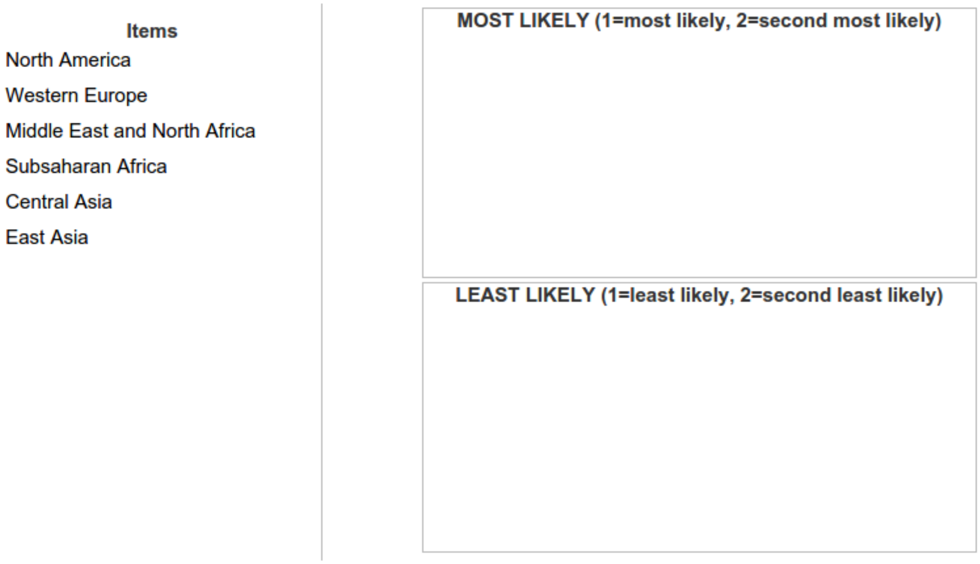
[Textbox]

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

[Textbox]

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

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



Items

North America

Western Europe

Middle East and North Africa

Subsaharan Africa

Central Asia

East Asia

MOST LIKELY (1=most likely, 2=second most likely)

LEAST LIKELY (1=least likely, 2=second least likely)

D How wealthy do you think the country was in terms of GDP per capita? (GDP per capita is often considered an indicator of a country's standard of living.)
We provide you with two example countries in each category.

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

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

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

F How likely do you think it is that the country had large oil reserves?

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

- Very Likely (81-100% chance)

G How likely do you it is that the majority of the country's population was white (Caucasian)?

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

H How much do you think the country spent annually on its military?[13]

- Very Little (less than $30 million)
- A Little ($30 to $120 million)
- About Average ($120 million to $600 million)
- A Large Amount ($600 million to $3.5 billion)
- A Very Large Amount (greater than $3.5 billion)

I How likely do you think it is that the country had been a U.S. military ally since World War II?

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

J What do you think was the total volume of import and export between the country and the U.S.?[14]

- A Very Small Amount (less than $100 million)
- A Small Amount ($100 million to $350 million)
- An Average Amount ($350 million to $1.5 billion)
- A Large Amount ($1.5 billion to $10 billion)
- A Very Large Amount (greater than $10 billion)

K How likely do you think it is that the country had carried out a joint military exercise with the U.S.?

- Very Unlikely (0-20% chance)

[13] The intervals are based on quintiles of countries's military expenditure in 2005.

[14] The intervals are based on quintiles of total volume of trade between the U.S. and other countries in 2005.

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

L Do you think the country had high levels or low levels of investment in U.S. businesses?

- Very high levels of investment in U.S. businesses
- High levels of investment in U.S. businesses
- Medium levels of investment in U.S. businesses
- Low levels of investment in U.S. businesses
- Very low levels of investment in U.S. businesses

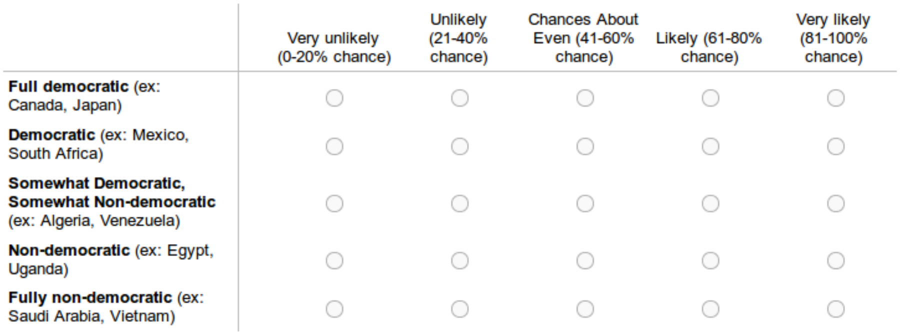
F.7 Treatment Measures

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

Treatment Measure 1: Probability of Being in Each Regime Type

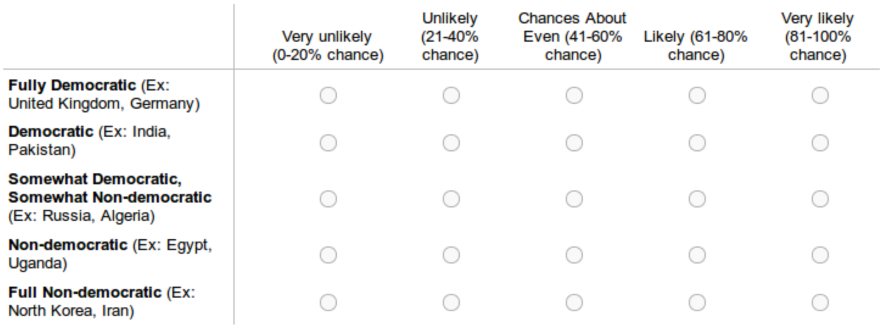
Think about the country described in the scenario. We would like to know how you would characterize its government. How likely do you think it is that the country has the following types of government?

For each government type, we provide you with two reference countries.[15]

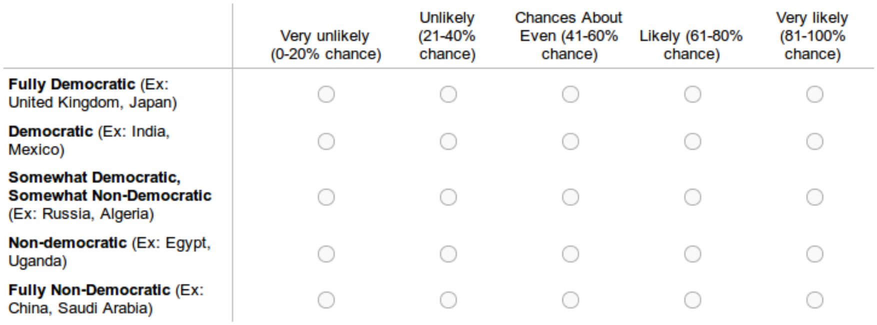


	Very unlikely (0-20% chance)	Unlikely (21-40% chance)	Chances About Even (41-60% chance)	Likely (61-80% chance)	Very likely (81-100% chance)
Full democratic (ex: Canada, Japan)	○	○	○	○	○
Democratic (ex: Mexico, South Africa)	○	○	○	○	○
Somewhat Democratic, Somewhat Non-democratic (ex: Algeria, Venezuela)	○	○	○	○	○
Non-democratic (ex: Egypt, Uganda)	○	○	○	○	○
Fully non-democratic (ex: Saudi Arabia, Vietnam)	○	○	○	○	○

[15] Each respondent input her answers using one of the three following matrices randomly assigned to him or her. We do this to make sure responses are not driven by the example countries we provide.



	Very unlikely (0-20% chance)	Unlikely (21-40% chance)	Chances About Even (41-60% chance)	Likely (61-80% chance)	Very likely (81-100% chance)
Fully Democratic (Ex: United Kingdom, Germany)	○	○	○	○	○
Democratic (Ex: India, Pakistan)	○	○	○	○	○
Somewhat Democratic, Somewhat Non-democratic (Ex: Russia, Algeria)	○	○	○	○	○
Non-democratic (Ex: Egypt, Uganda)	○	○	○	○	○
Full Non-democratic (Ex: North Korea, Iran)	○	○	○	○	○



	Very unlikely (0-20% chance)	Unlikely (21-40% chance)	Chances About Even (41-60% chance)	Likely (61-80% chance)	Very likely (81-100% chance)
Fully Democratic (Ex: United Kingdom, Japan)	○	○	○	○	○
Democratic (Ex: India, Mexico)	○	○	○	○	○
Somewhat Democratic, Somewhat Non-Democratic (Ex: Russia, Algeria)	○	○	○	○	○
Non-democratic (Ex: Egypt, Uganda)	○	○	○	○	○
Fully Non-Democratic (Ex: China, Saudi Arabia)	○	○	○	○	○

Treatment Measure 2: Characteristics of Democracies

Think about the country described in the scenario.

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

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

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

F.8 Support for Military Action

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

Support for Using Force Question

Think about the scenario you read.

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

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

- Favor strongly
- Favor somewhat
- Neither favor nor oppose
- Oppose somewhat
- Oppose strongly
- I don't know

[Textbox]

F.9 Mediation Questions

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

Open-ended Mediation Question

Why did you select that answer choice in the previous question?[16]

F.9.1 If the U.S. attacked...

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

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

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

[16] This question was not be asked in the Separated-Placebos Design because we did not want to increase the complexity of an already complex design.

F.9.2 If the U.S. did not attack...

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

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

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

F.9.3 Morality of Using Force

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

- It is morally wrong.
- It is not morally wrong.
- I don't know.

F.10 Demographics Questions

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

F.10.1 Education

What is the highest level of education you have completed?

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

F.10.2 Political Party

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

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

F.10.3 Age

What is your age?
[Drop-down menu: 18 to older than 100]

F.10.4 Sex

What is your sex?

- Female
- Male
- Other

F.10.5 Political Ideology

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

G Full Summary of "Democratic Peace" Survey Results

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

G.1 Survey Procedure

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

G.2 Balance Tests

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

Table 1: Number of Respondents by Experimental Condition

Treatment Assignment	Vignette Type	*N*
Non-democracy	Basic	513
Democracy	Basic	517
Non-democracy	Covariate Control	512
Democracy	Covariate Control	513
Non-democracy	ENE	516
Democracy	ENE	509

Figure 10: Demographic Variables by Experimental Condition

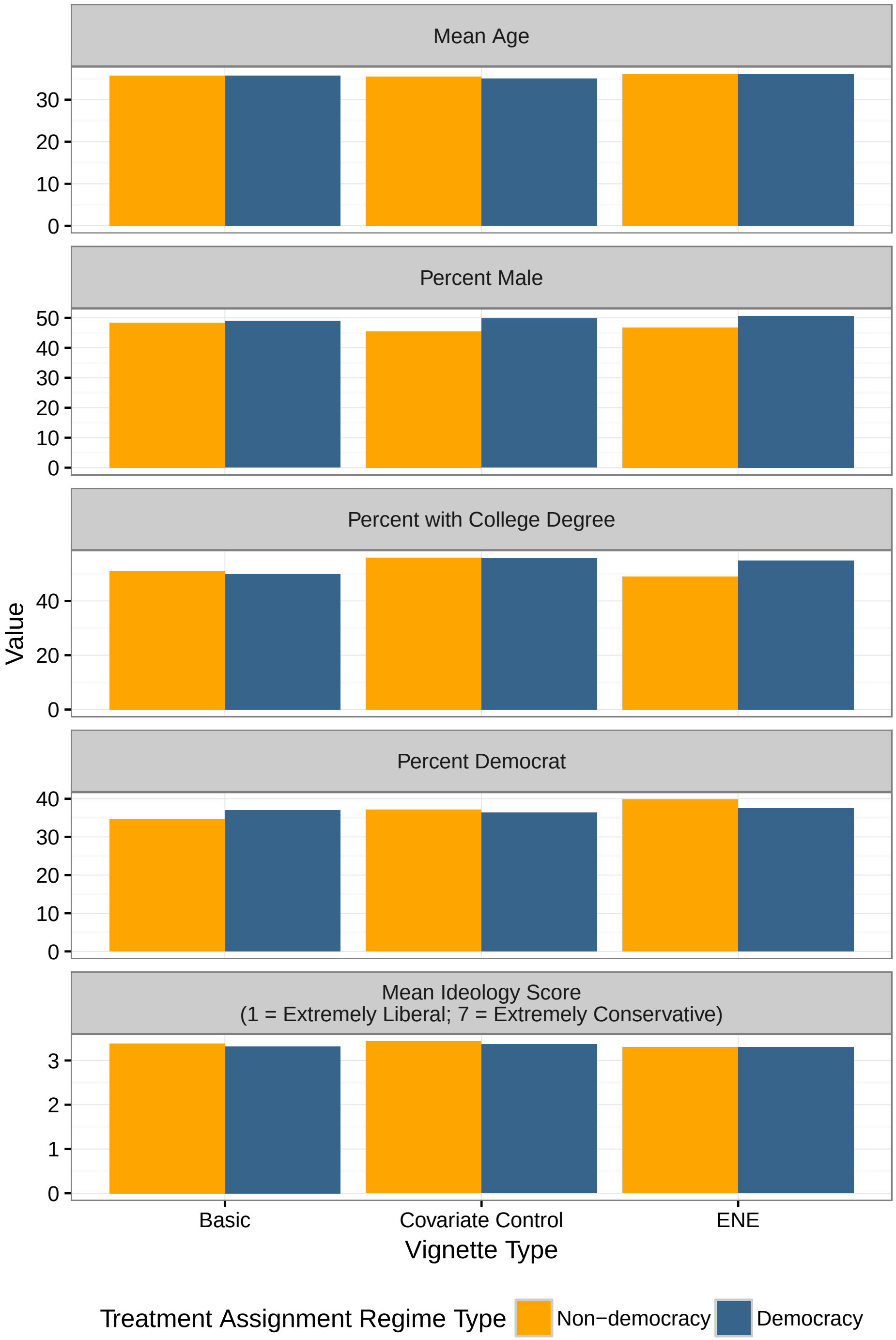


Figure 11: Balance Tests on Demographic Variables

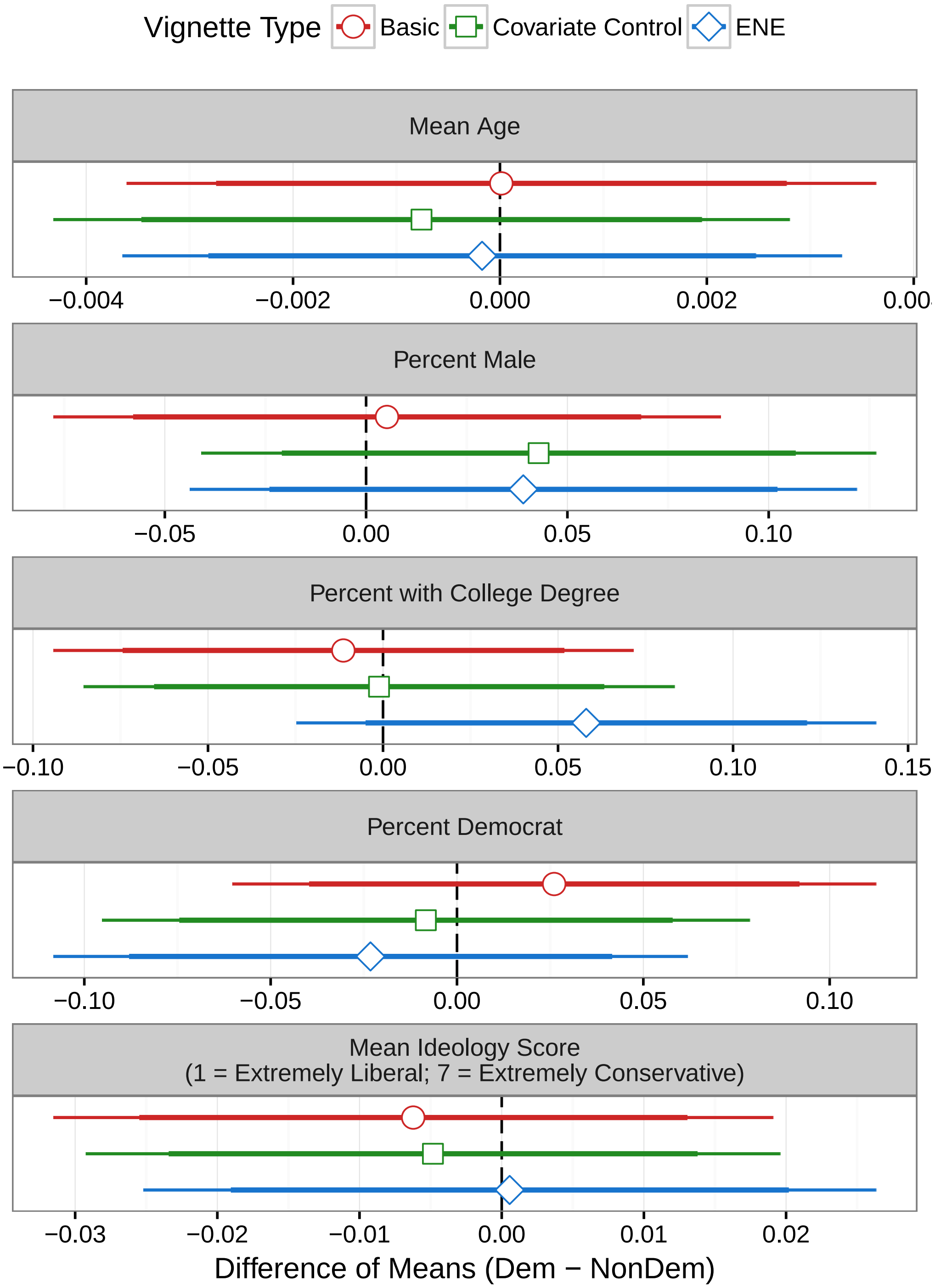


Table 2: Balance Test: Results from Joint F-test Using All Five Demographics Variables to Predict Treatment Assignment

Vignette Type	F-statistic	p-value
Basic	$F(5{,}965) = 0.19$	0.968
Covariate Control	$F(5{,}942) = 0.69$	0.632
ENE	$F(5{,}964) = 1.12$	0.349

G.3 Coding Placebo Test Results

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

A: Regions of the World

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

B: GDP per Capita

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

C: Religion

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

D: Oil Reserves

We define Y^N_D as one minus the subjects' response to the oil reserves placebo test question: $Y^N_{i,D}$ equals one minus the mean of the probability interval subject i selects.[17]

E: Race

We define Y^N_E as subjects' response to the race placebo test question; we scale the responses so that $Y^N_{i,E}$ equals the mean of the probability interval subject i selects.

[17] Because we hypothesize that subjects think the democratic country is less likely to have had large oil reserves, we invert the responses so the direction of the IE violation is the same as the direction in the other placebo tests.

F: Military Alliance

We define Y_F^N as subjects' response to the military alliance placebo test question; we scale the responses so that $Y_{i,F}^N$ equals the mean of the probability interval subject i selects.

G: Trade with the U.S.

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

H: Joint Military Exercise

We define Y_H^N as subjects' response to the joint military exercise placebo test question; we scale the responses so that $Y_{i,H}^N$ equals the mean of the probability interval subject i selects.

I: Foreign Direct Investment

We define Y_I^N as subjects' response to the FDI test question; we scale the responses so that $Y_{i,I}^N$ corresponds to an ordinal scale with "very high levels of investment" being a 4 and "very low levels of investment" being a 0.

J: Military Capability

We define Y_J^N as subjects' response to the military capability placebo test question; we scale the responses so that $Y_{i,J}^N$ equals the real-world median of the military expenditure interval subject i selects. For instance, in 2005, there are 36 countries in the "A Very Large Amount (greater than $3.5 billion)" interval; the median value among them was $9.1815 billion. This means that $Y_{i,J}^N = 9181500000$ when subject i selects "greater than $3.5 billion." As a robustness check, we also scale the responses ordinally so that $Y_{i,J}^N = 0$ when subject i selects "Very Little" and $Y_{i,J}^N = 4$ when she selects "A Very Large Amount."

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

G.4 Placebo Test Questions

G.4.1 Analysis of Placebo Test Responses

In our analyses, we use both non-standardized and standardized versions of the placebo response variable. Unless noted, we analyze the results for the Basic, Covariate Control, and the ENE designs separately. Define $Y_{i,j}^N$ as subject i's non-standardized response to placebo test question j.

There are a total of Q respondents; each respondent i is assigned to vignette type V_i. The mean non-standardized placebo response within each vignette type v is $\bar{Y}^N_{j,v} = \frac{1}{Q} \sum_{i=1}^{Q} \left[Y^N_{i,j} \mathbf{1}(V_i = v) \right]$.

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

$$Y_{i,j} = \frac{Y^N_{i,j} - \bar{Y}^N_{j,v}}{\sqrt{\frac{1}{Q} \sum_{i=1}^{Q} \left(Y^N_{i,j} - \bar{Y}^N_{j,v} \right)^2}} \tag{12}$$

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

We predict the difference-in-means will be largest (furthest from 0 in the positive direction) for the Basic Vignettes, medium for the Control Vignettes, and smallest (closest to 0) for the ENE Vignettes.[18]

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

G.4.2 Placebo Test Results

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

[18] To simplify the presentation, all placebo variables are coded so that more positive values correspond to the values more common in democracies.

Figure 12: Placebo A: Most Likely Regions Distribution of Responses

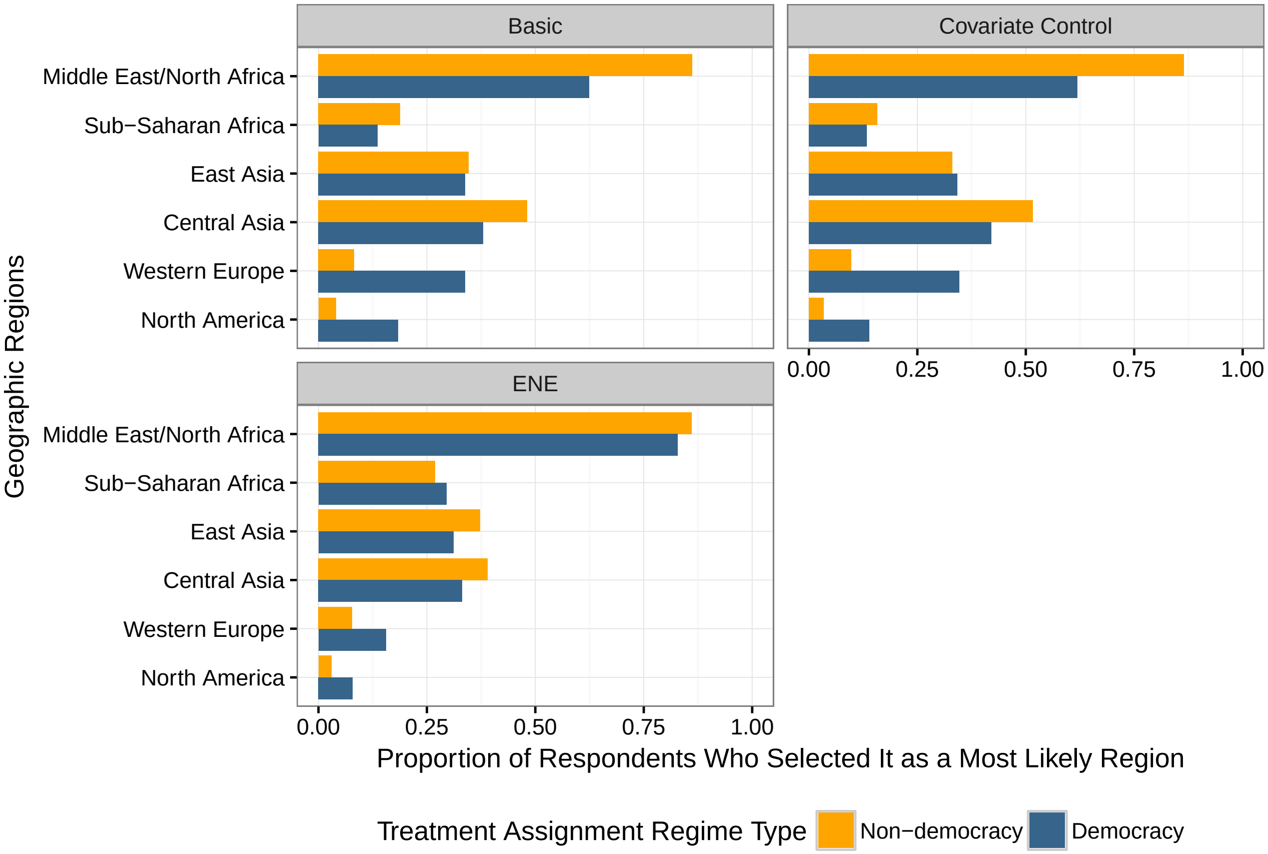


Figure 13: Placebo B: GDP per Capita Distribution of Responses

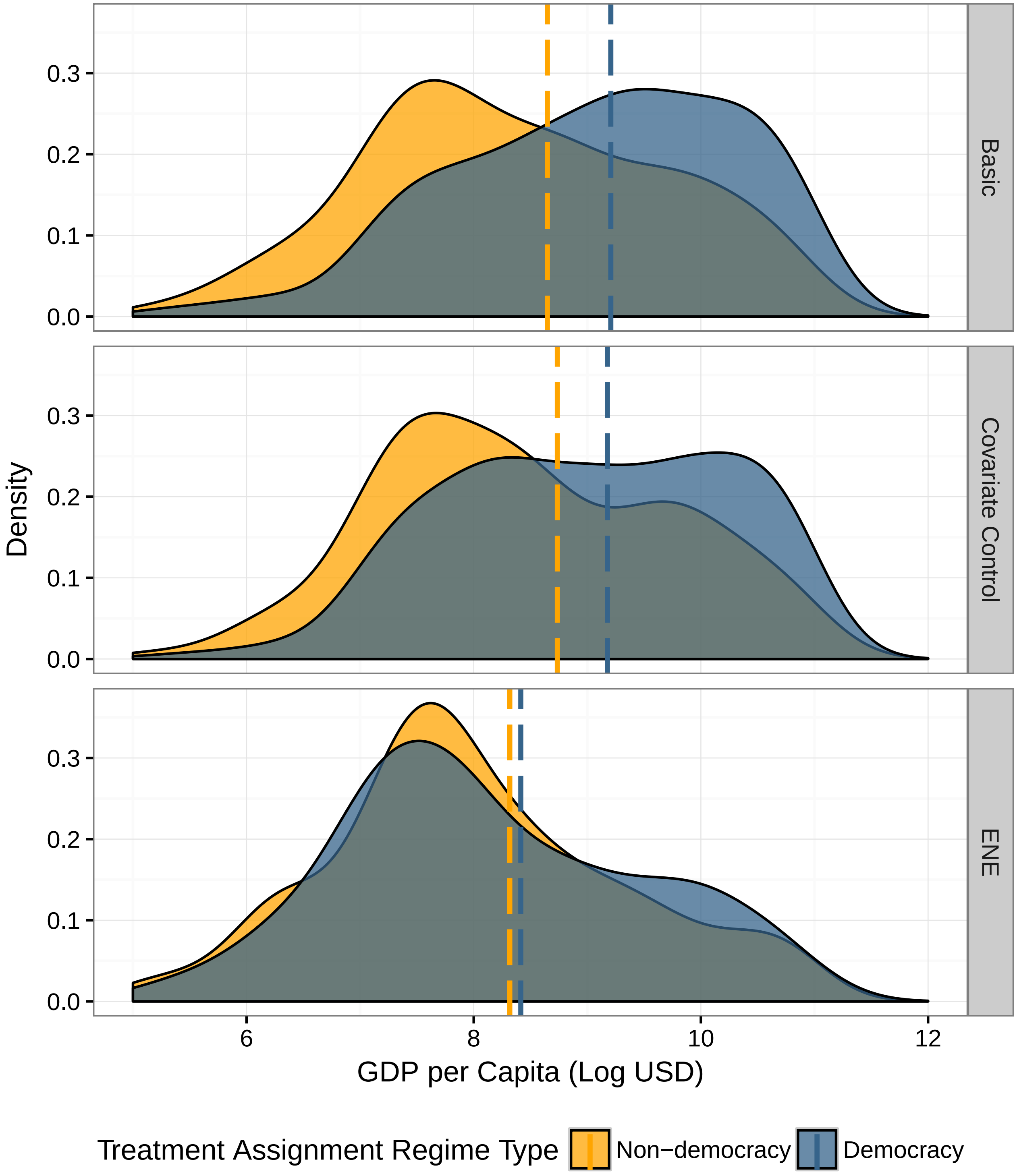


Figure 14: Placebo C: Likelihood of Being Majority Christian Distribution of Responses

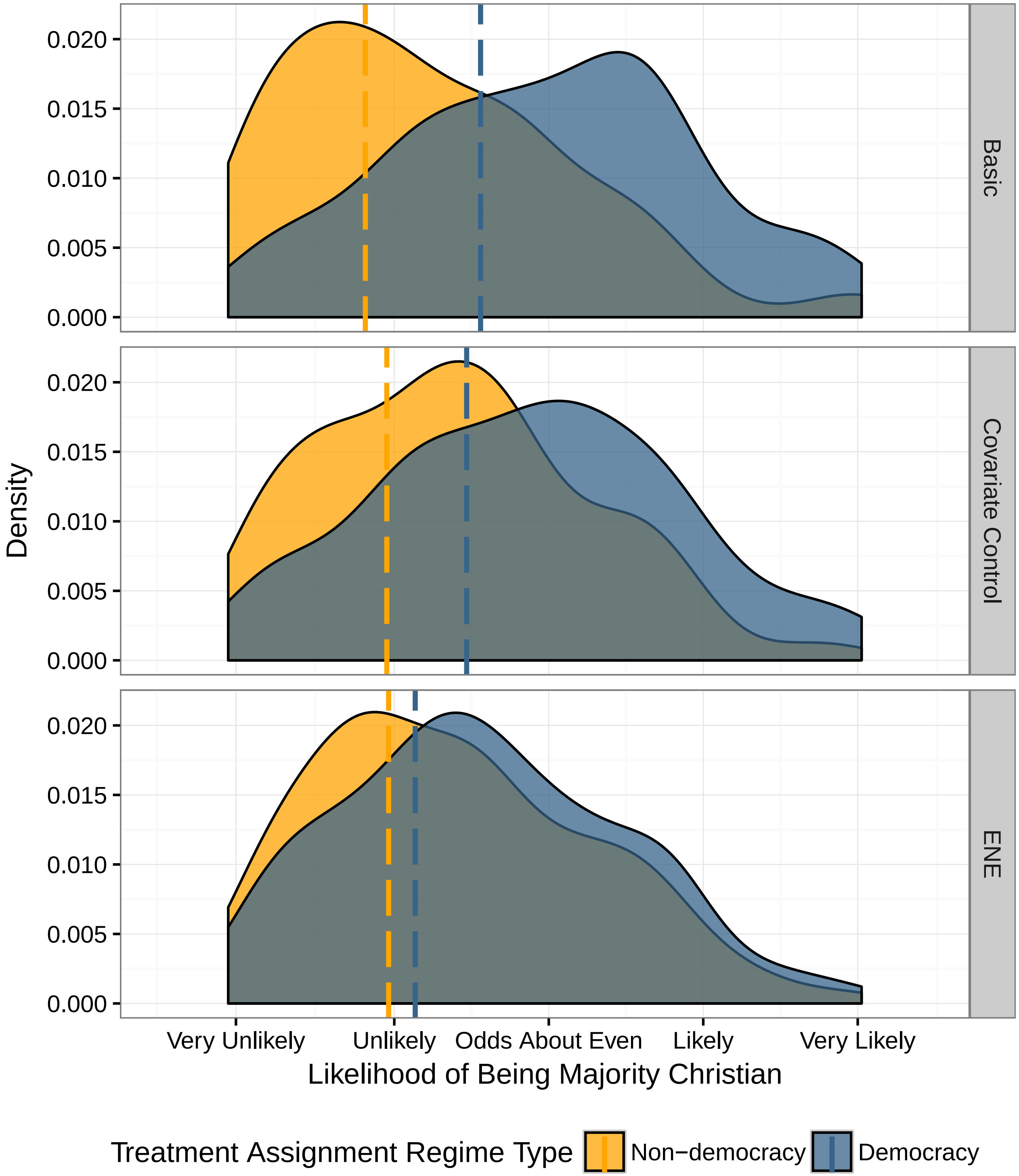


Figure 15: Placebo D: Likelihood of Not Having Large Oil Reserves Distribution of Responses

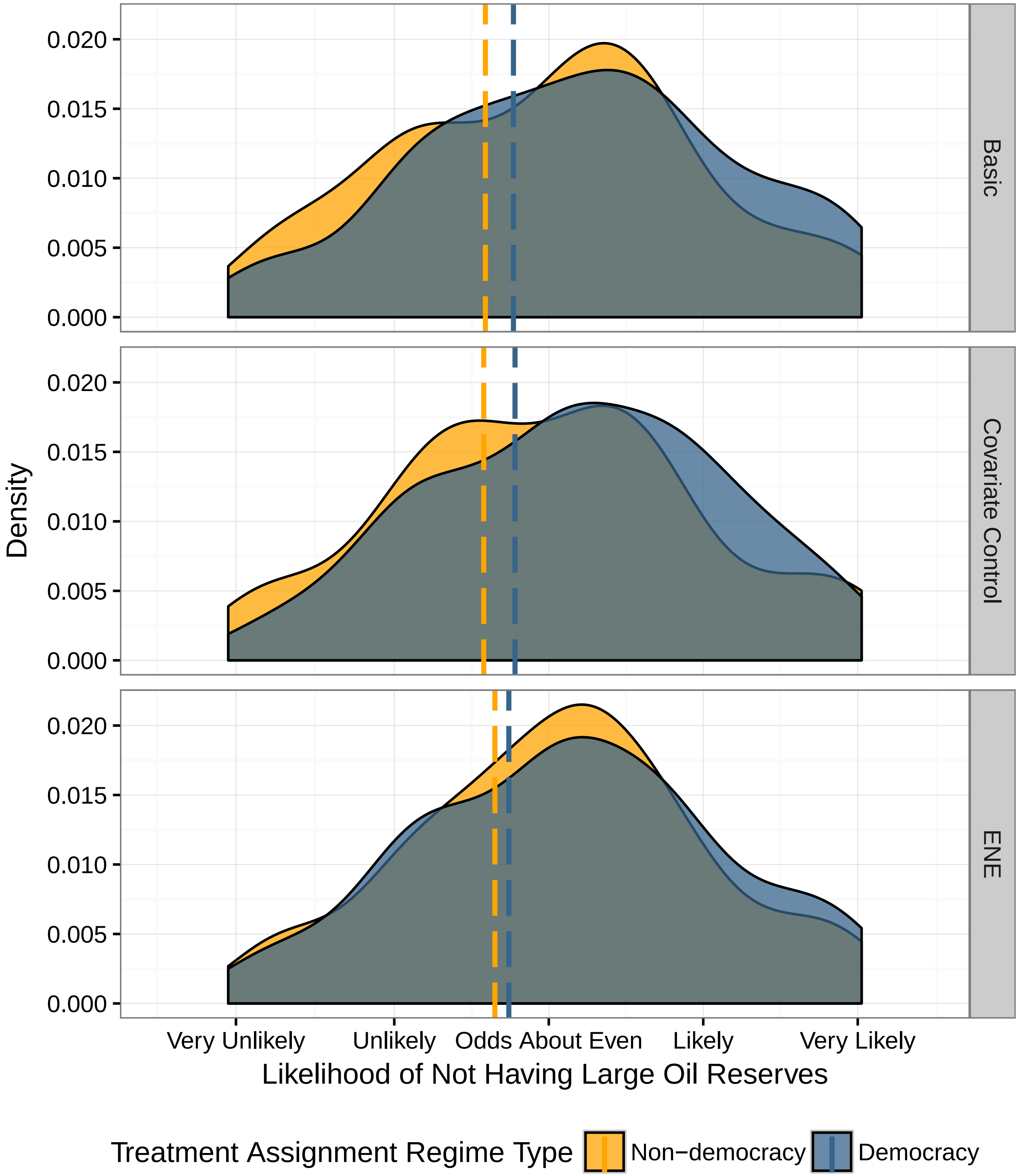


Figure 16: Placebo E: Likelihood of Being Majority White Distribution of Responses

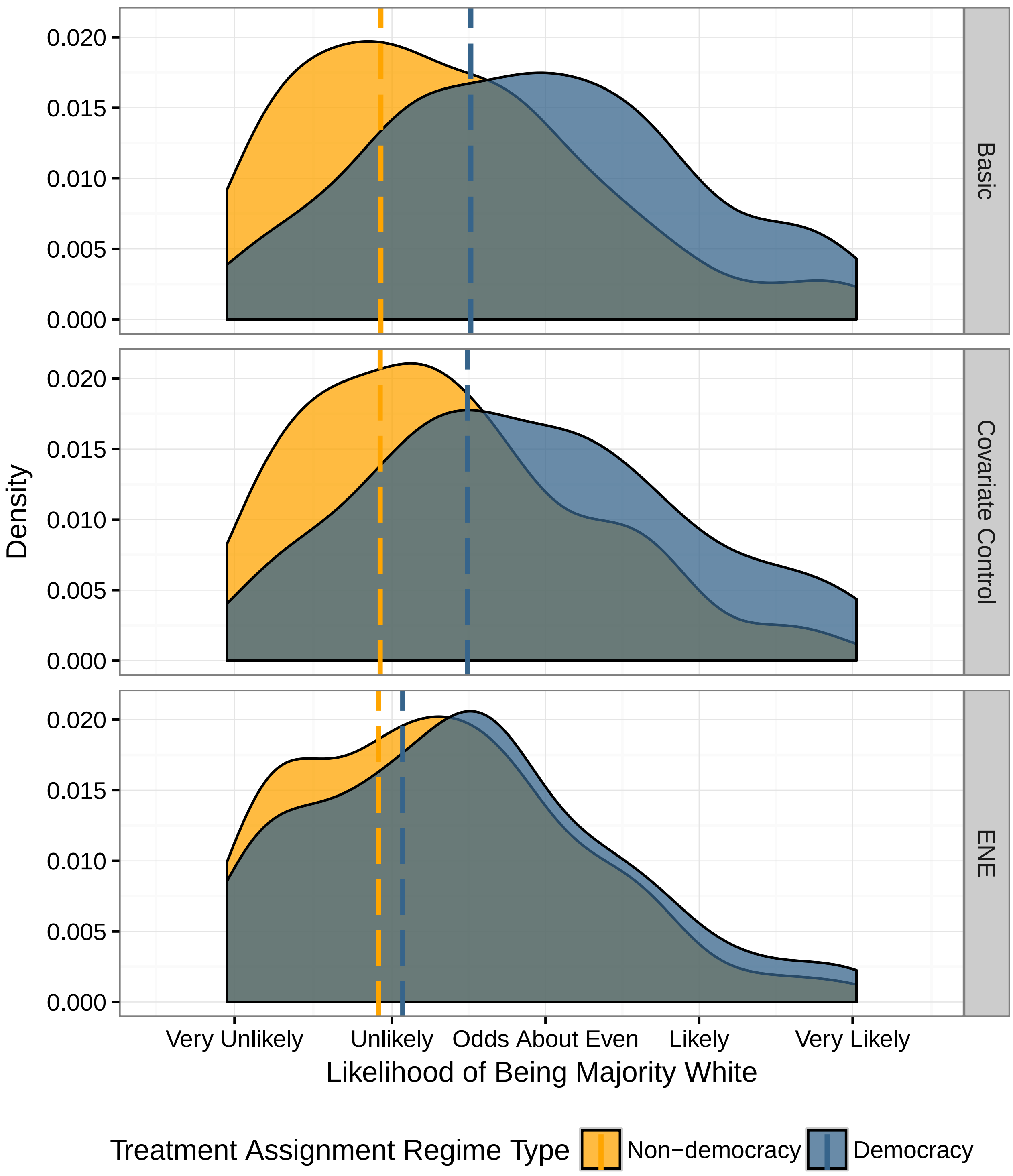


Figure 17: Placebo F: Likelihood of Military Alliance with the U.S. since World War II Distribution of Responses

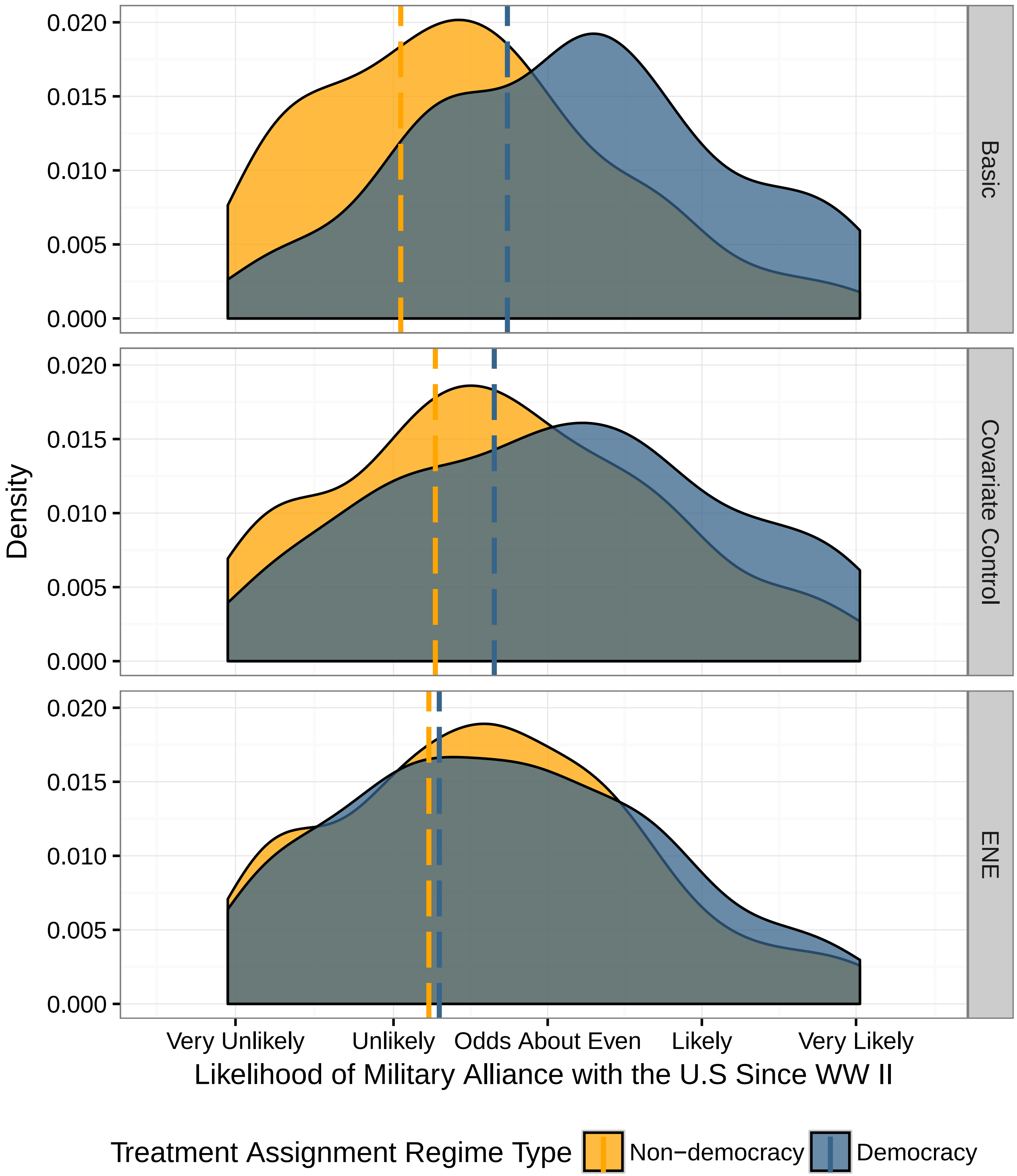


Figure 18: Placebo G: Level of Trade with the U.S. Distribution of Responses

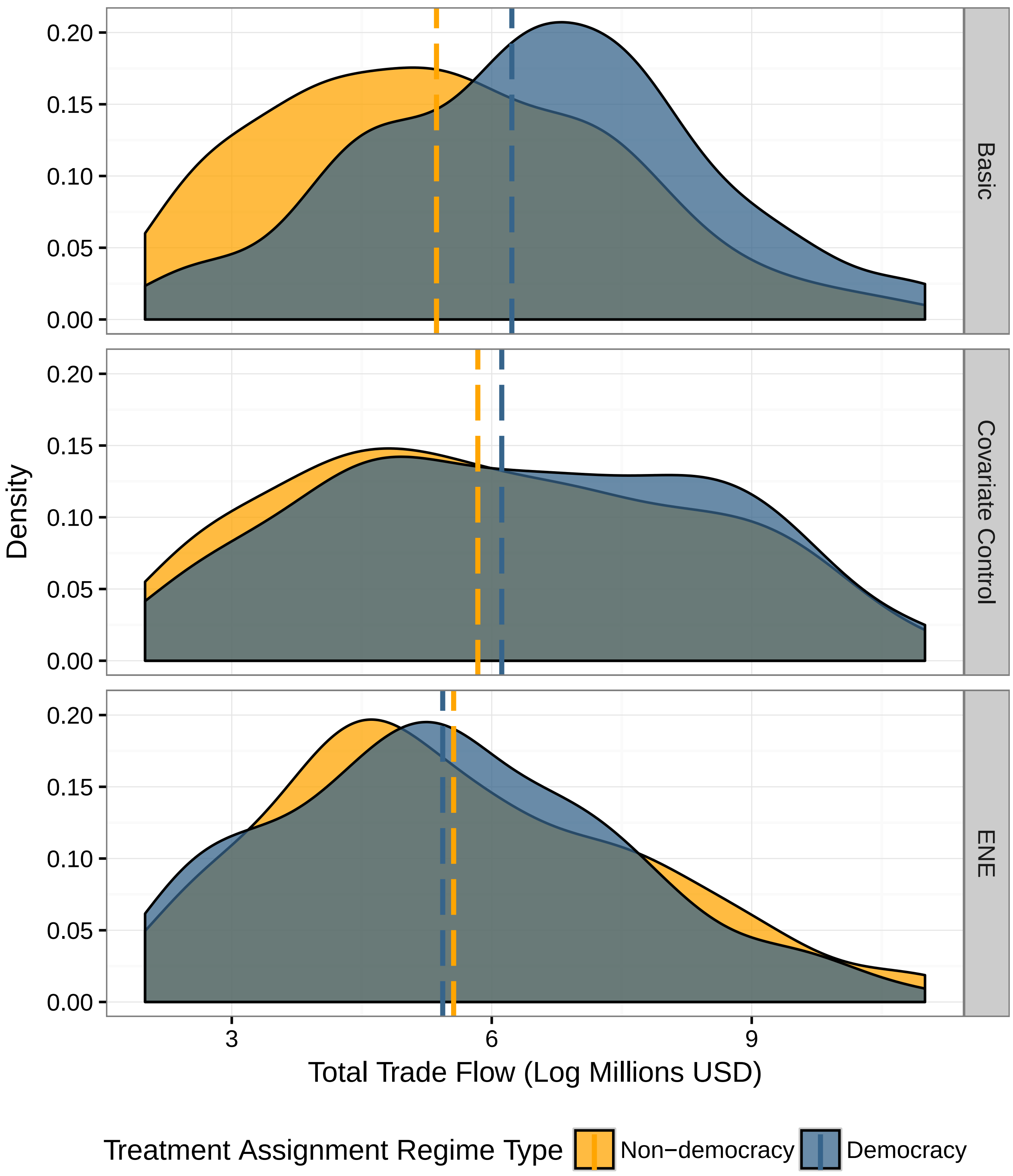


Figure 19: Placebo H: Likelihood of Joint Military Exercise with the U.S. Distribution of Responses

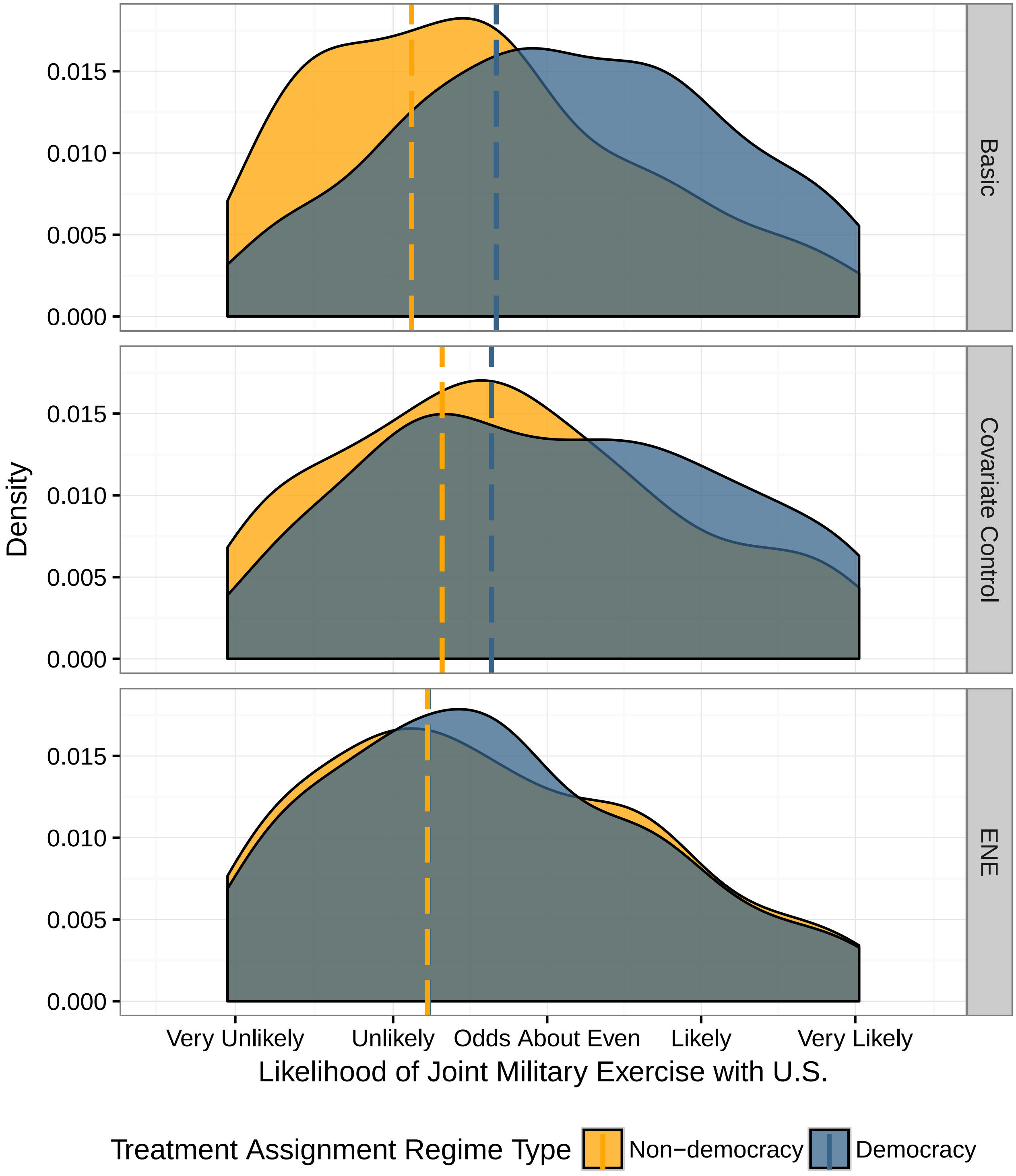


Figure 20: Placebo I: Level of Investment in U.S. Businesses Distribution of Responses

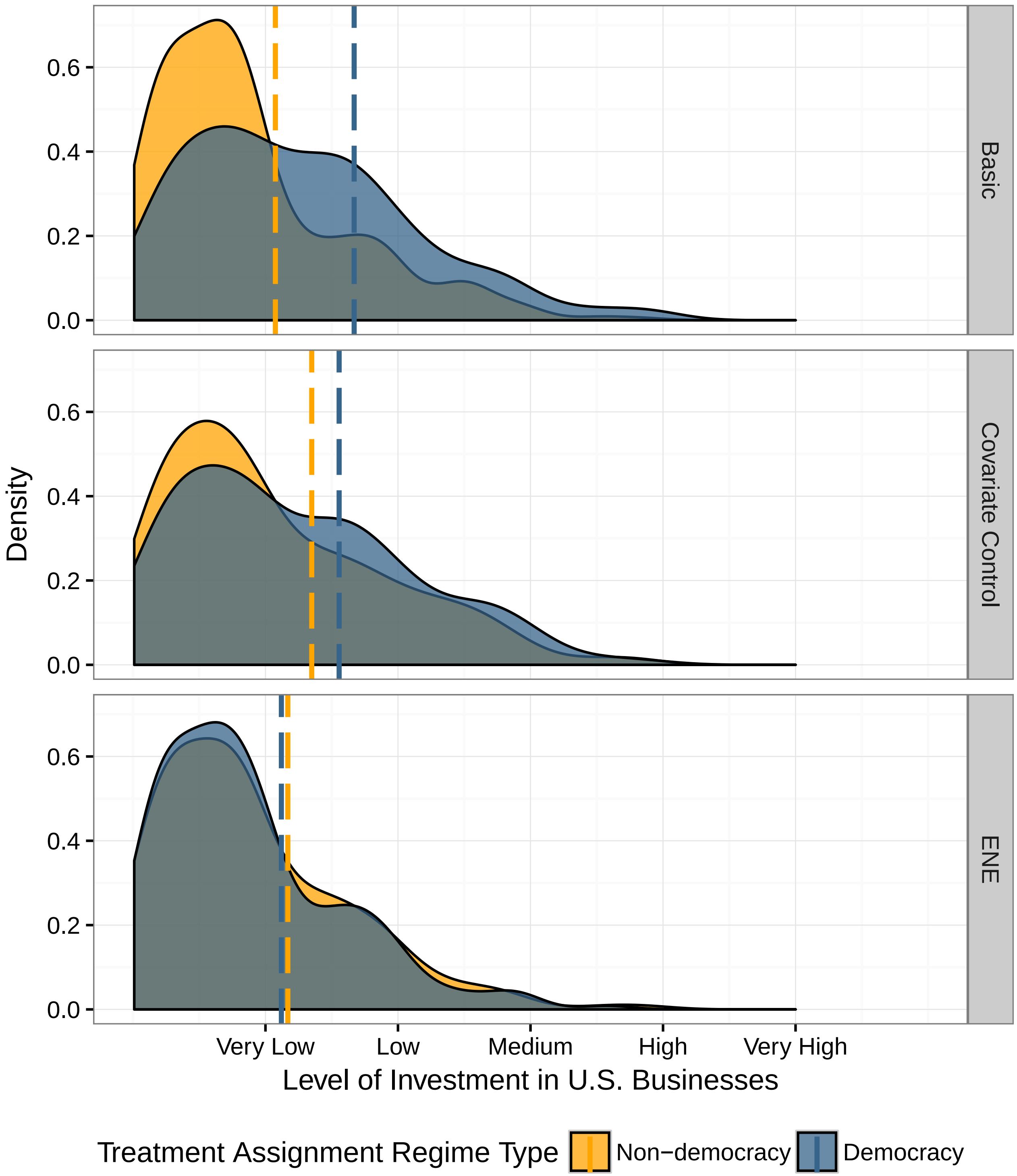


Figure 21: Placebo J: Military Spending Distribution of Responses

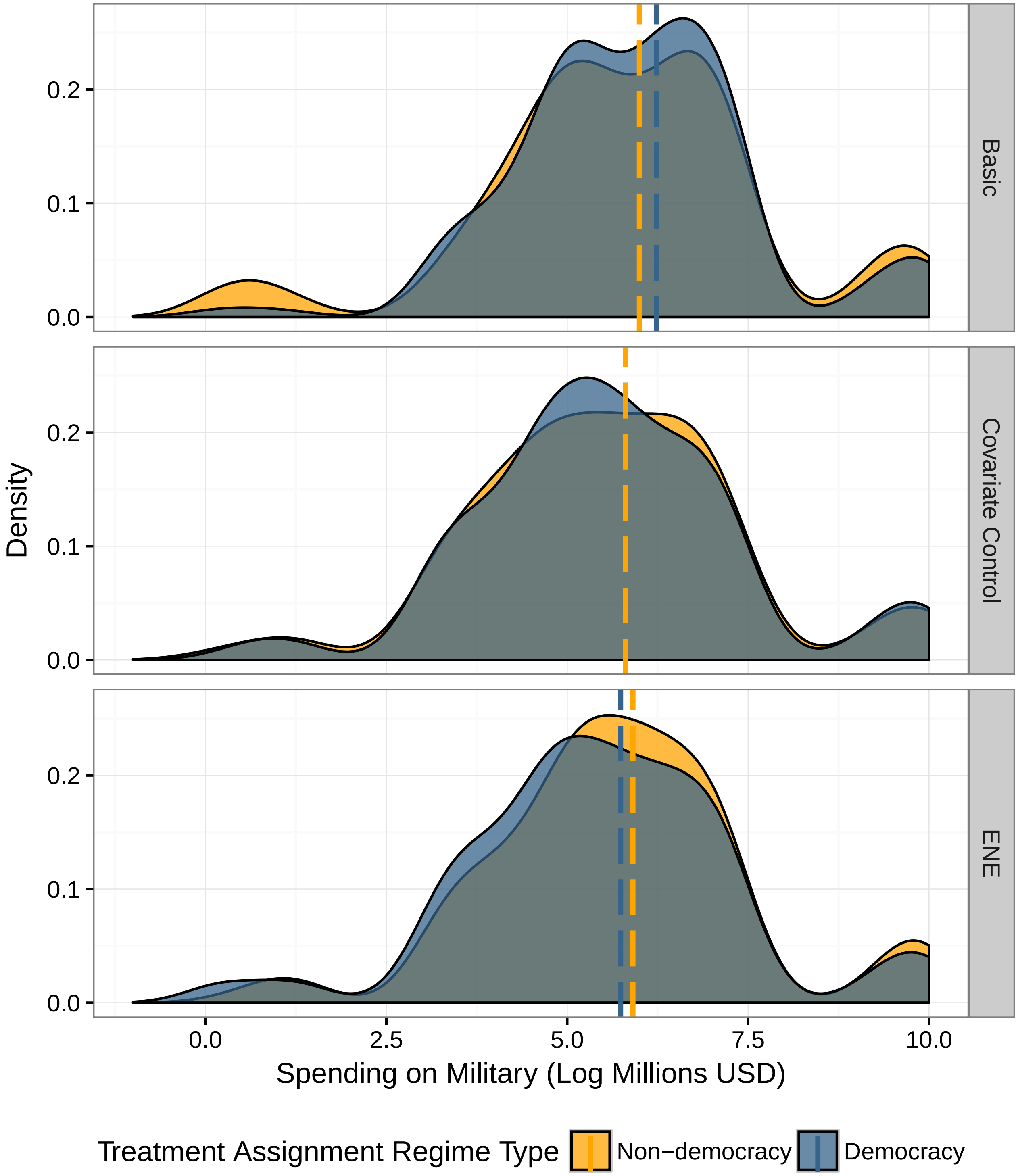


Figure 22: Placebo K: Most Likely Countries Distribution of Responses

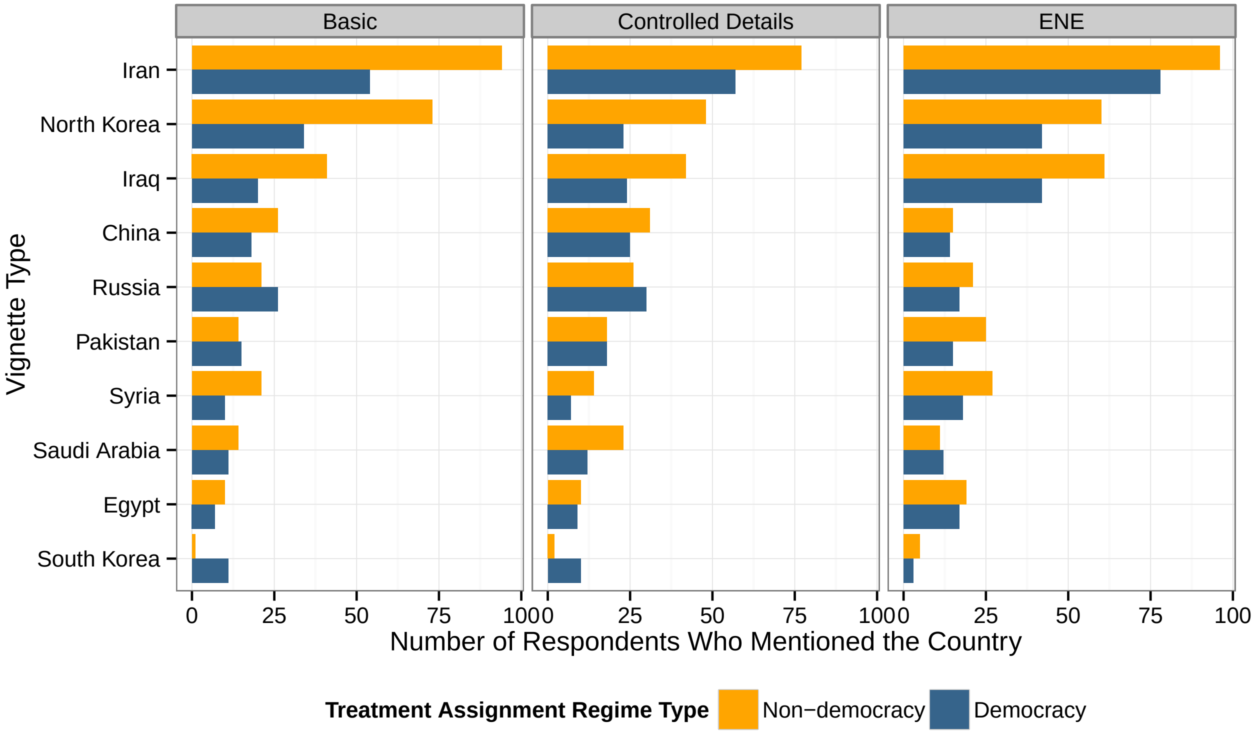
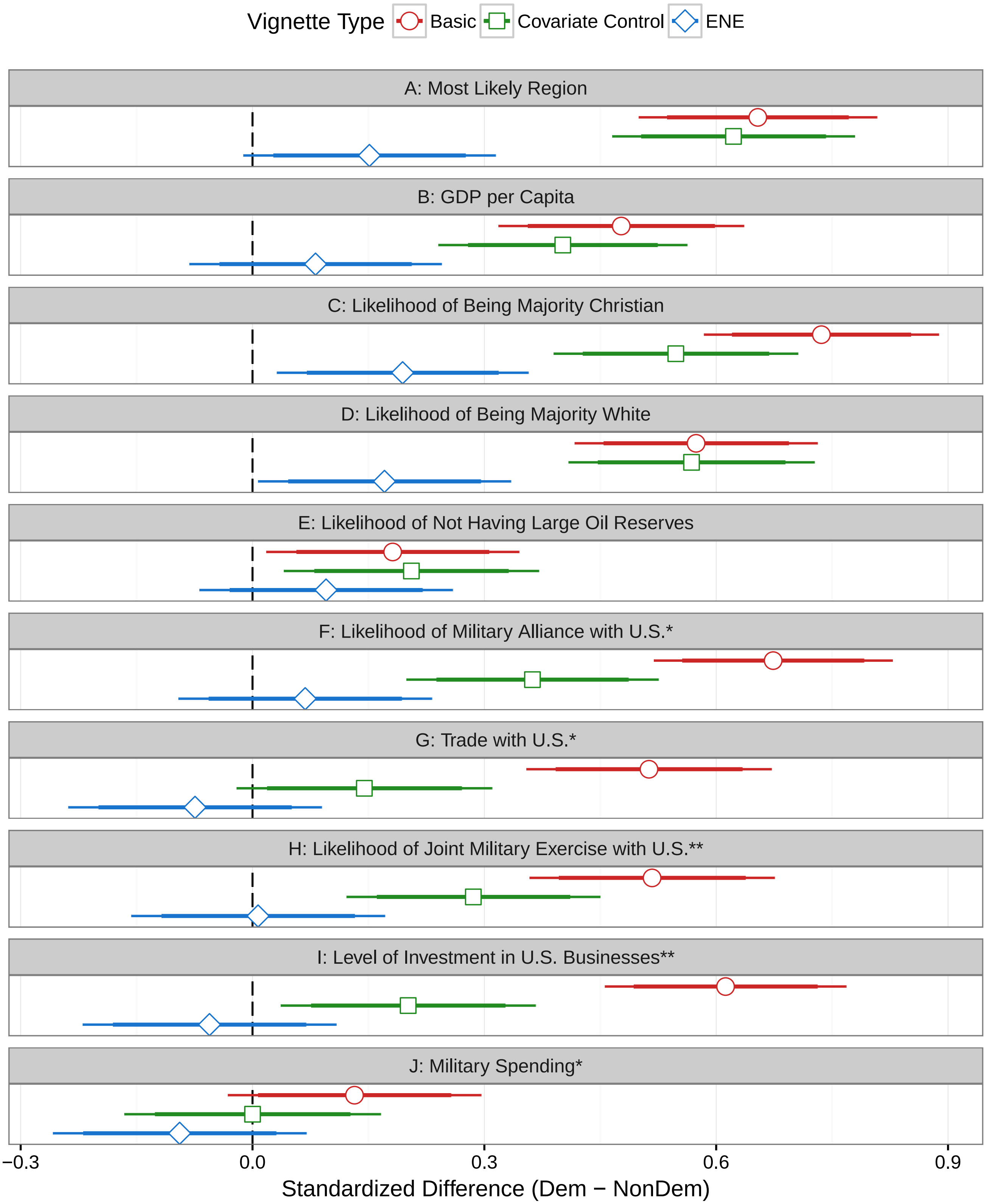
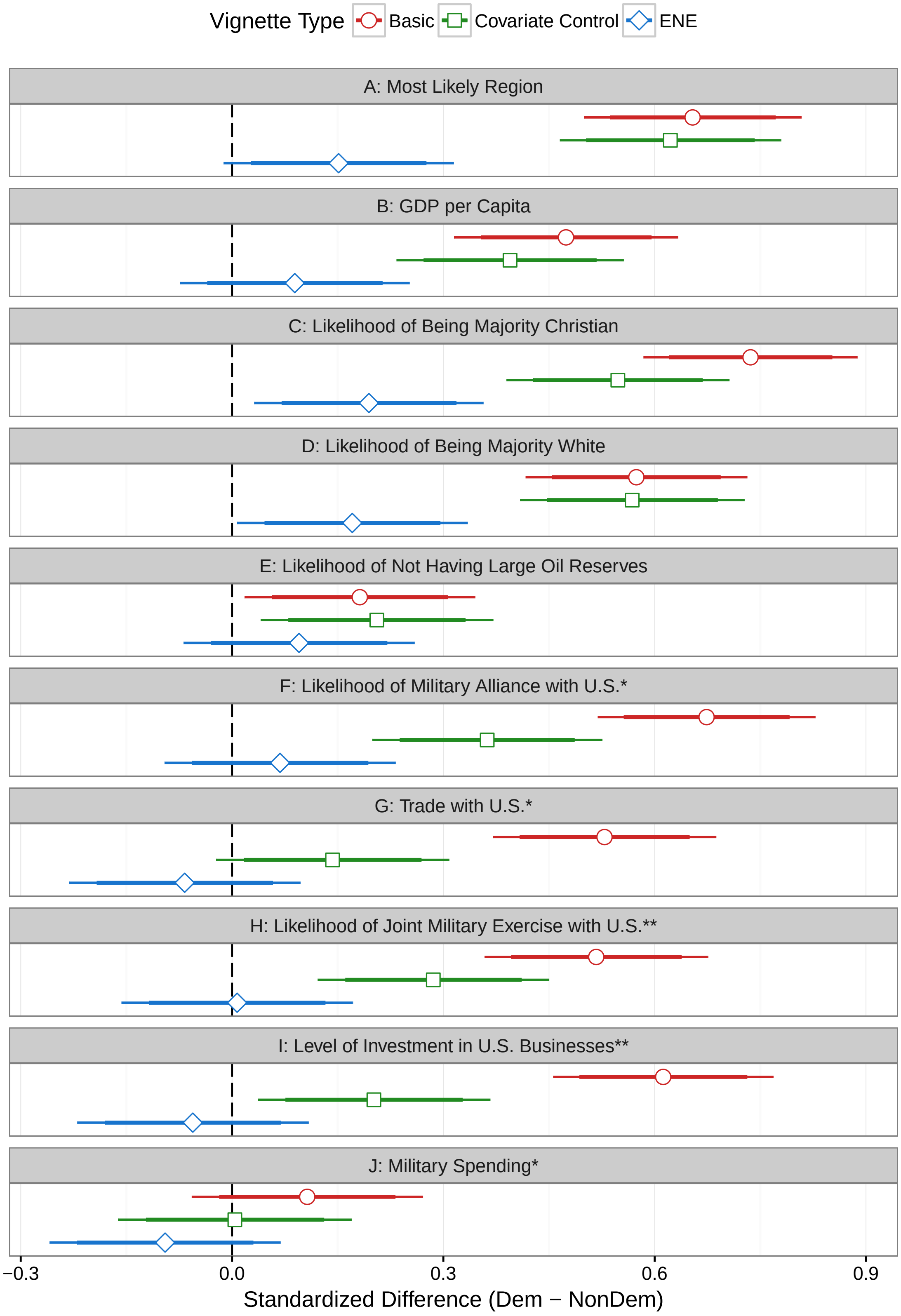


Figure 23: Placebo Test Questions Results (Standardized) including Military Spending



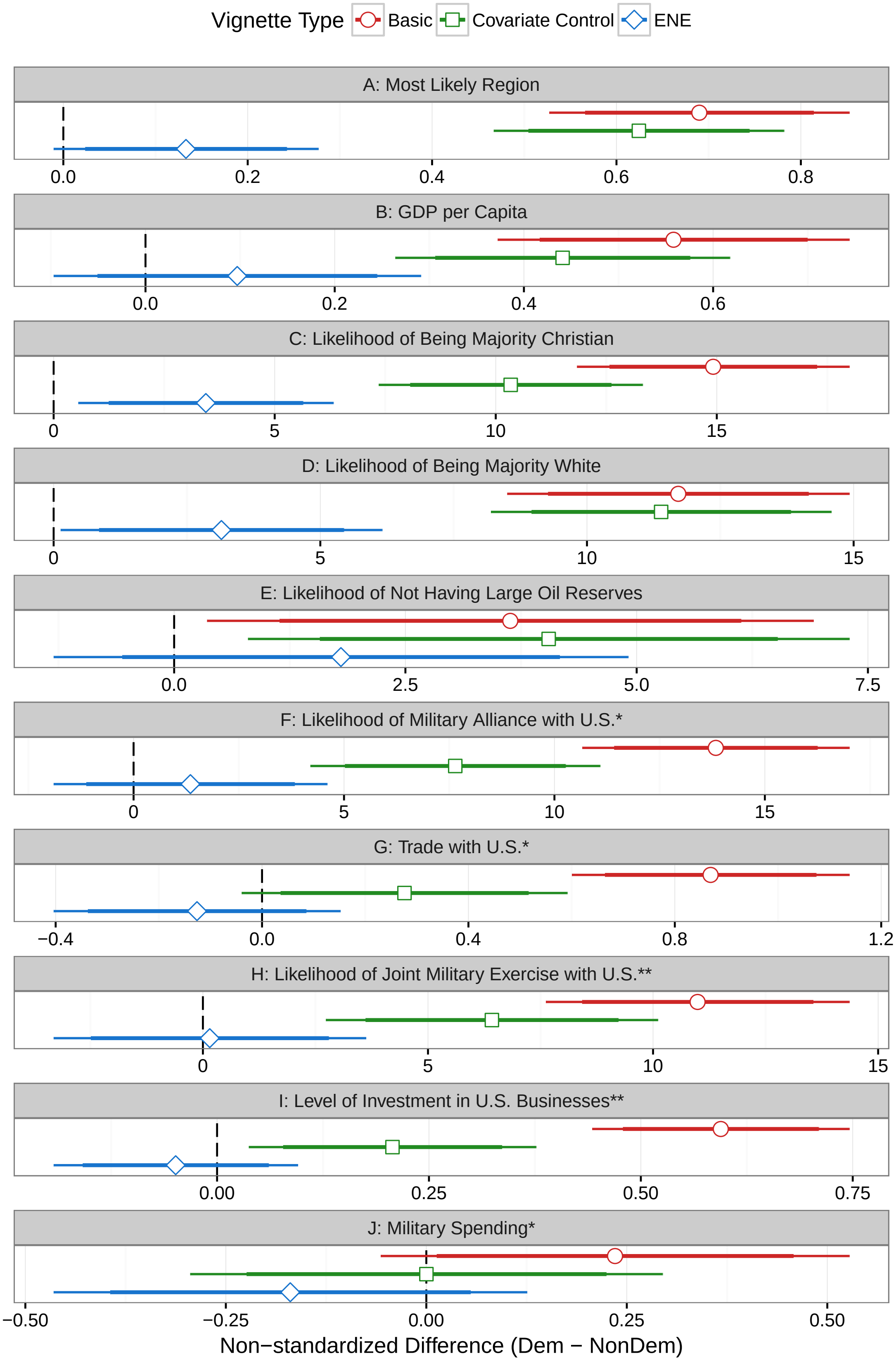
For Placebo Outcomes B, G, and J, we took the natural log of the non-standardized USD outcome before standardizing within vignette type.

Figure 24: Placebo Test Questions Results (Standardized) including Military Spending



For Placebo Outcomes B, G, and J, we converted the non-standardized USD outcomes to ordinal values (0 to 6 for Placebo Outcome B; 0 to 4 for Placebo Outcomes G and J) before standardization.

Figure 25: Placebo Test Questions Results (Non-standardized) including Military Spending



For Placebo Tests B, G, and J, the outcomes are in their original USD values.

G.4.3 Joint Test of Placebo Outcomes Using NPC

We will jointly test if there exists imbalances among the placebo test responses A through I using the non-parametric combination test (NPC).[19] Note that the null hypothesis we assume for NPC is different from those we assumed for the previous hypothesis tests. For NPC, we assume the global sharp null, that is the regime type of the country in the scenario has no effect for every subject's response to every placebo test question.

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

1. Calculate a vector of observed test statistics $\mathbf{T^{obs}} = (T_A^{obs}, T_B^{obs}, ..., T_m^{obs}, ..., T_I^{obs})$ corresponding to the nine partial placebo tests. We use the difference-of-means between the treatment and control group as our test statistic: $T_m^{obs} = \mathbb{E}(Y_m|Z_i = 1) - \mathbb{E}(Y_m|Z_i = 0)$ for each placebo test m.

2. Repeat the following Q times:
a) Randomly permute the group labels Z of units that are exchangeable under the sharp null.
b) In each permutation $q \in \{1, ..., Q\}$, calculate the vector $\mathbf{T_q^*} = (T_{Aq}^*, T_{Bq}^*, ..., T_{mq}^*, ..., T_{Iq}^*)$ of values of the nine test statistics.

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

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

   a) Fisher's: $\psi_a = -\sum_m \log(p_m)$
   b) Liptak's: $\psi_b = -\sum_m \Phi^{-1}(p_m)$, where Φ^{-1} is the inverse of the normal CDF
   c) Tippett's: $\psi_c = -\min_m(p_m)$

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

In Table 3, we report the NPC results by vignette type. We use all three combining functions and use $Q = 10000$ permutation for each analysis. As the results show, the global p-values are orders of magnitudes smaller for the Basic and Covariate Control design than for the ENE design. This suggests that the imbalance in placebo outcomes, as a whole, are far worse in those first two designs than in the ENE design.

[19] See Caughey, Dafoe, and Seawright (2017) for a detailed explanation of NPC.

Table 3: p-values from Joint Test of Placebo Outcomes Using NPC

Combining Function Used	Basic	Covariate Control	ENE
Fisher's	< 0.0001	< 0.0001	0.0055
Liptak's	< 0.0001	< 0.0001	0.0191
Tippett's	< 0.0001	< 0.0001	0.0101

G.5 Treatment Measures

G.5.1 Coding and Analysis of Treatment Measure Results

Type Question 1: Probability of Being in Each Regime Type Our Treatment Measure 1 measures subjects' beliefs about how democratic the target country is. We call this latent variable D_i, which we proxy using our imputed measure $R1_i$ based on subject i's response to Treatment Measure 1.

Define $K_{i,j}$ as subject i's response to regime type category $j \in \{1, 2, ..., 5\}$. Using these responses, we impute $R1_i$, which ranges from -10 to 10 — much like the Polity score. The procedure for imputing $R1_i$ is:

1. First, we impute $K_{i,j}$, the probability subject i assigned to regime type category j. Recall that in the survey, each subject i selected a probability interval $[K^a_{i,j}, K^b_{i,j}]$ for each regime type category j. We reduce the dimensionality of each subject's responses by defining $K_{i,j}$ as the mean of the probability interval $[K^a_{i,j}, K^b_{i,j}]$.
$K_{i,j} = (K^a_{i,j} + K^b_{i,j})/2$

2. Let $R1_{i,j}$ be the normalized probability subject i assigns to regime type category j. For $j \in \{1, 2, ..., 5\}$, we normalize $K_{i,j}$ so that $\sum_j K_{i,j} = 1$, meaning that the probabilities each subject assigned to the regime type categories will sum to one.
$R1_{i,j} = \frac{K_{i,j}}{\sum_j K_{i,j}}$

3. Finally we impute $R1_i$. For $j \in \{1, 2, ..., 5\}$, we multiply the mean polity score of the jth regime type category O_j[20] by $R1_{i,j}$ then we sum these five products. In short, we calculate the expected value of the "Polity score" for each subject i's response.
$R1_i = \sum_j (O_j R1_{i,j})$

Define the treatment effect of the democracy condition (Z) on responses to this treatment measure question as $\tau_{i,R1} = R1_i(Z_i = 1) - R1_i(Z_i = 0)$. For each vignette type, we estimate $\mathbb{E}(\tau_{i,R1})$ using $\hat{\beta}_{1,R1}$ from the regression: $\mathbb{E}(R1_i|Z_i) = \beta_{0,R1} + \beta_{1,R1} Z_i$.

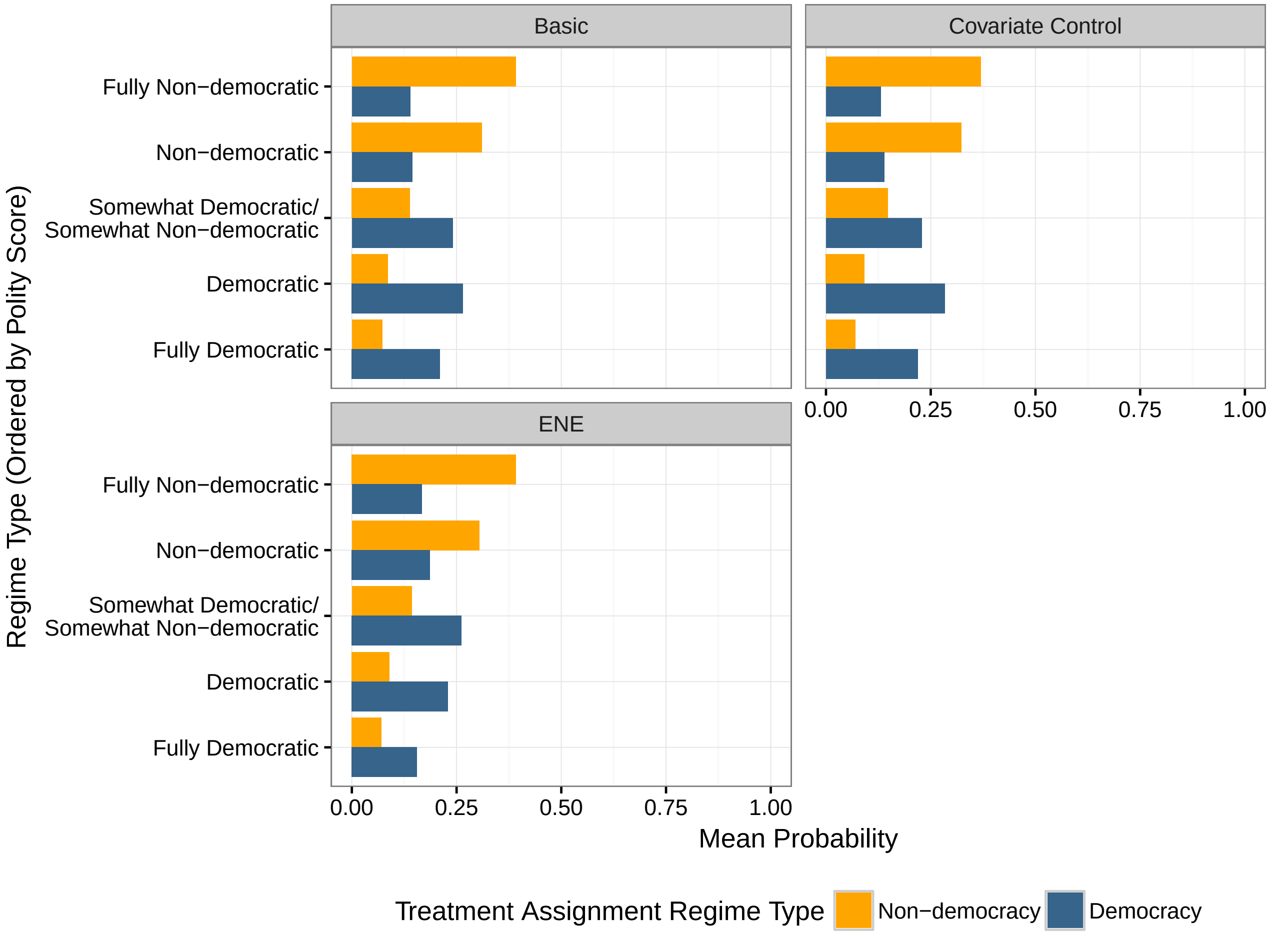
Regime Type Question 2: Characteristics of Democracies We define $R2_i$ as the number of democratic characteristics respondent i selected, which serves as a proxy for how democratic respondent i thought the target country is.

[20] The five regime types we present in our survey are fully democratic, democratic, somewhat democratic/somewhat non-democratic, non-democratic, and fully non-democratic. These correspond to the following Polity 4 regime types: full democracy (10), democracy (6 to 9), open anocracy (1 to 5), closed anocracy (-5 to 0), and autocracy (-10 to -6). We choose not to use the Polity 4 terms because they are too specialized for our respondents to understand.

Define the treatment effect of the democracy condition (Z) on responses to this treatment measure question as $\tau_{i,R2} = R2_i(Z_i = 1) - R2_i(Z_i = 0)$. For each vignette type, we estimate $\mathbb{E}(\tau_{i,R2})$ using $\hat{\beta}_{1,R2}$ from the regression: $\mathbb{E}(R2_i|Z_i) = \beta_{0,R2} + \beta_{1,R2}Z_i$.

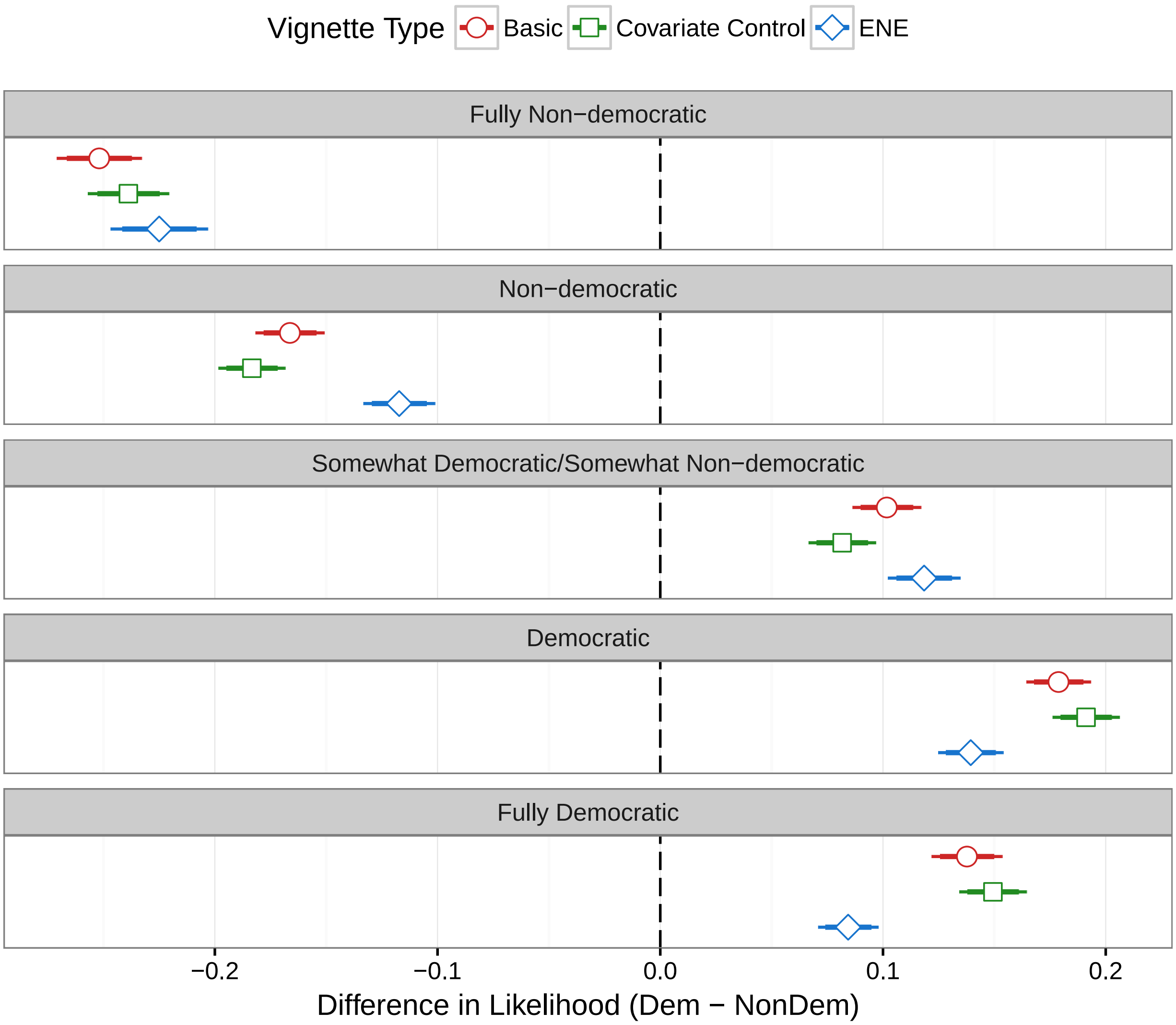
G.5.2 Treatment Measures Results

Figure 26: Treatment Measure: Probability of Each Regime Type



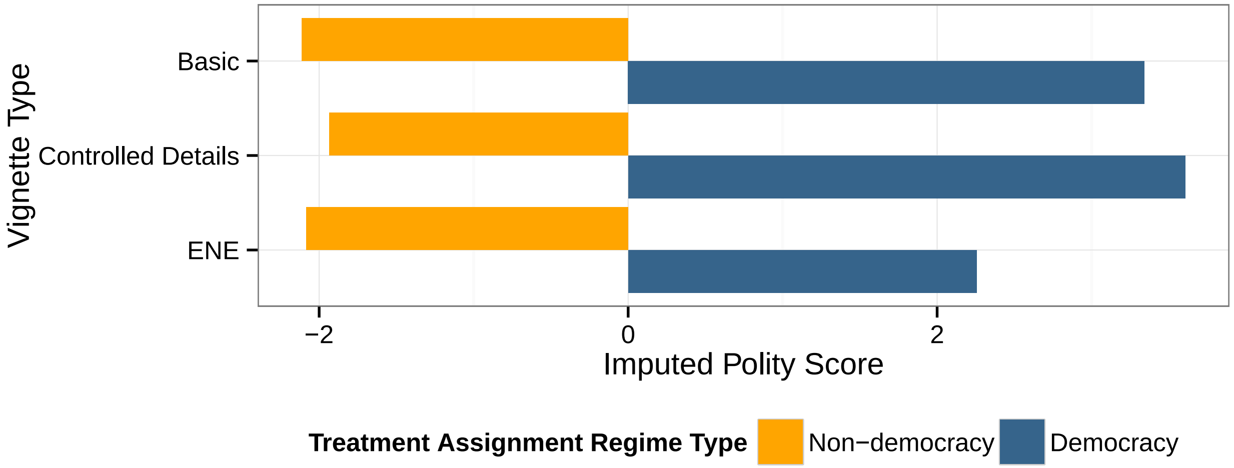
We compare the mean probability subjects assigned to each regime type between those who received the Democracy vignette and those who received the Non-democracy vignette. For each subject, we normalize the probability she assigned to each regime type so that her probabilities sum up to 100 percent.

Figure 27: Treatment Measure: Probability of Each Regime Type Coefficient Plot

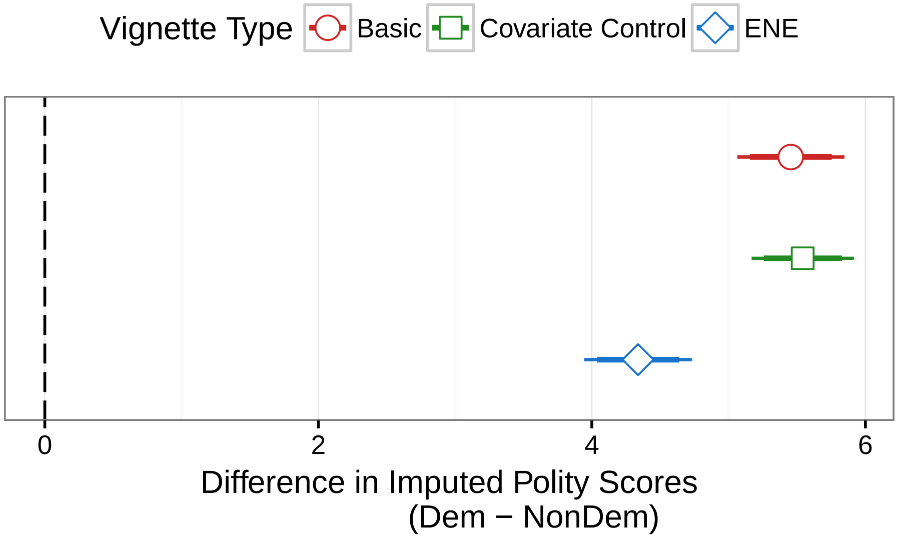


In the coefficient plot above, we estimate the difference in the likelihood of respondents assign to each regime type by vignette type.

Figure 28: Treatment Measure: Imputed “Polity Score”

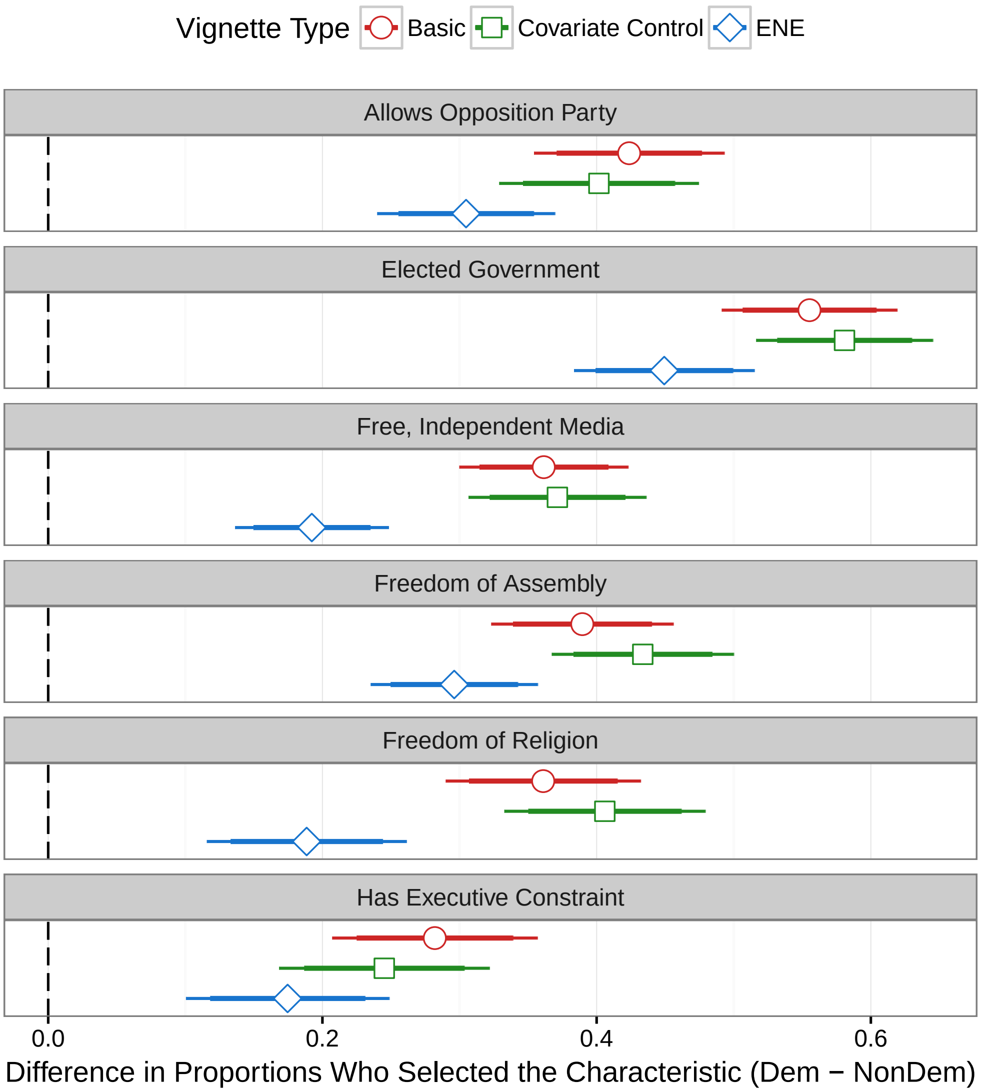


We impute “Polity scores” based on responses using the method described in Subsection G.5.1.

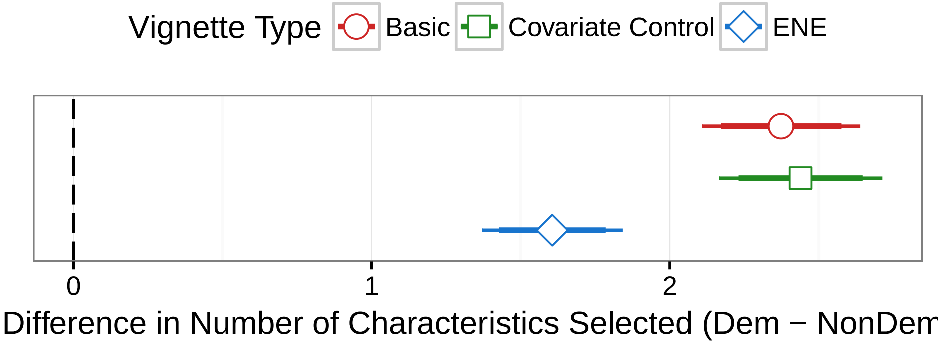


In this coefficient plot above, we present the difference in the mean imputed “Polity score” between treatment and control for each vignette type.

Figure 29: Treatment Measure: Characteristics of Democracies Coefficient Plots



In this coefficient plot above, we present the difference in the proportion of respondents who selected each characteristic of democracies by each vignette type.



In the coefficient plot above, we present the difference in the mean number of characteristics selected between treatment and control by each vignette type.

G.6 Substantive Outcome: Support for War

G.6.1 ITT and IV Analyses

ITT Estimates First, we estimate the effect of the democracy condition on support for military action against the target country. We call this the intention-to-treat (ITT) estimate. We define S_i as subject i's response to the support for using force question (the typical outcome variable used in experimental studies of the democratic peace), such that $S_i = 4$ for "strongly oppose" and $S_i = 0$ for "strongly support."[21] Define the treatment effect of the democracy condition Z on response S as $\zeta_i = S_i(Z_i = 1) - S_i(Z_i = 0)$.

For each vignette type, we estimate $\mathbb{E}(\zeta_i)$ using the coefficient estimate $\hat{\beta}_1$ from the regression: $\mathbb{E}(S_i|Z_i) = \beta_0 + \beta_1 Z_i$.

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

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

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

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

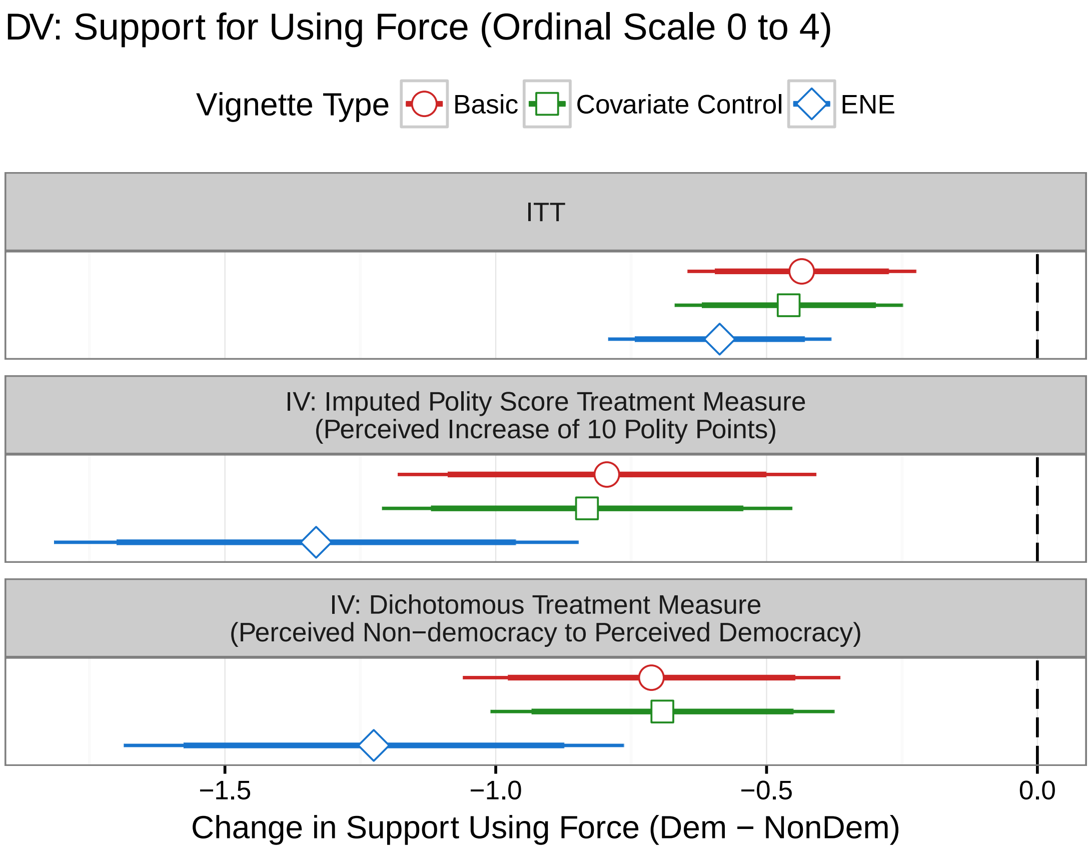
For each vignette type, we estimate ρ and calculate a confidence interval. Note that the IV estimates are likely biased for vignette types that violate the exclusion restriction. Significant imbalance on dispositive placebo variables for the Basic and Covariate Control design provides evidence against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

As a robustness check, we repeat the analysis above using the dichotomous measure of support for using force S^* and a dichotomous treatment measure $R1^*$. Let $R1_i^* = 1$ if respondent i indicates the aggressor country has a higher probability of being democratic or fully democratic than being non-democratic or fully non-democratic and let $R1_i^* = 0$ if she indicates otherwise.

[21] Subjects who answered "don't know" to the question were assigned $S = 2$.

G.6.2 ITT and IV Results

Figure 30: ITT and IV Estimates: Ordinal Measure of Support for Using Force

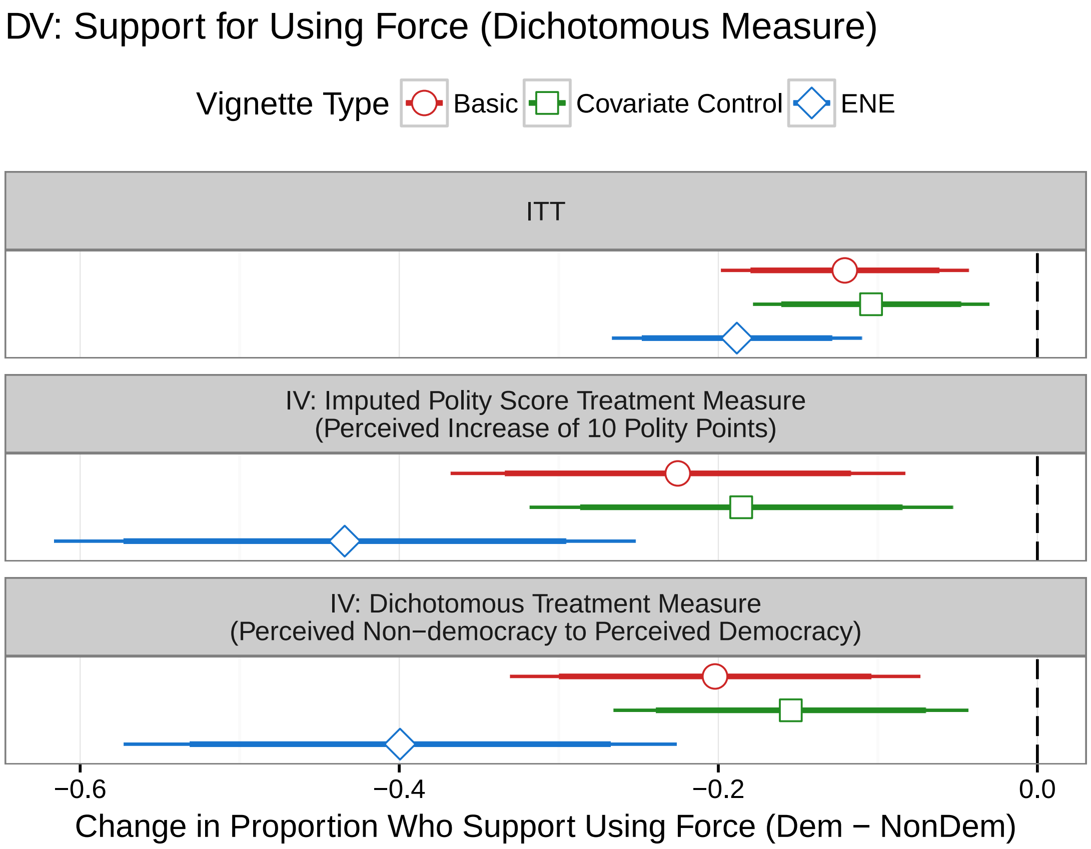


The dependent variable is support for using force measured using a 5 point ordinal scale. Those who strongly favor using force is coded 4 and those who strongly oppose using force is coded 0. Those who responded with "don't know" is coded 2.

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

For the Dichotomous Treatment Measure, we code that respondents perceive the country is a democracy when they indicate the country has a higher probability of being democratic or fully democratic than being non-democratic or fully non-democratic.

Figure 31: ITT and IV Estimates: Dichotomous Measure of Support for Using Force

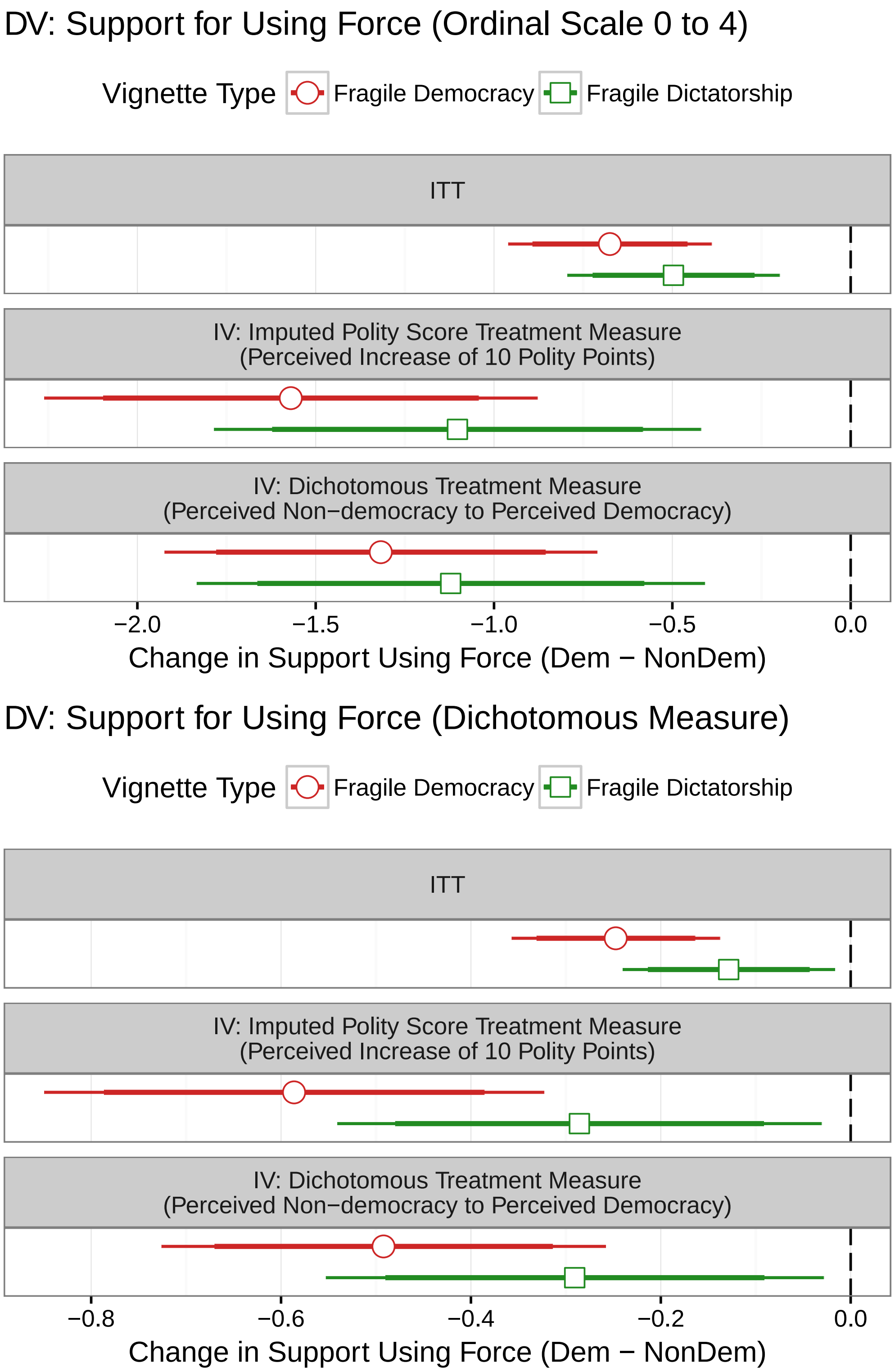


The dependent variable is dichomotous measure of support for using force. Those who “favor strongly” and “favor somewhat” using force are coded 1 and 0 otherwise.

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

For the Dichotomous Treatment Measure, we code that respondents perceive the country is a democracy when they indicate the country has a higher probability of being democratic or fully democratic than being non-democratic or fully non-democratic.

Figure 32: ITT and IV Estimates: Two Versions of the ENE Design



We perform the same type of analysis seen in Figures 30 and 31 except we examine the two different versions of the ENE Design. Recall that in one version of the ENE Design, the country started out a fragile democracy and in the other version, the country started out as a fragile dictatorship.

G.7 Additional Analyses

For results of additional analyses, please see Section C of our Supplementary Appendix.

H Coercion and Provocation Survey Experiment (2016)

H.1 Survey Overview

We employ survey experiments to evaluate whether, and under what circumstances, incidents of coercion provoke resolve, while holding constant or ruling out other possible explanations. Survey respondents read a short scenario describing a hypothetical dispute between China and the United States in the East China Sea. Our key experimental manipulation of the scenario consists of additional information regarding a recent incident in this dispute. Respondents are randomly assigned to receive information describing one of five incidents with equal probability: control, attack, non-collision, collision, or accident. The control group receives no additional information other than the background information presented in the scenario. The attack condition describes China attacking a US plane, killing the pilot. The collision and non-collision conditions describe US and Chinese planes involved in dangerous maneuvers, and either colliding or not colliding depending on the condition. The accident treatment describes a non-coercive event (a weather accident) leading to the death of a US pilot.

The substantive outcome questions asked respondents how far the U.S. should go , in terms of using military force, to assert its claims in the dispute. Our placebo test question measured respondents' perceptions of China's background hostile military intentions in the dispute.

H.2 Text of the Survey

In this subsection, we present the text of the survey related to the results shown in the paper.

H.2.1 Experimental Introduction

- The United States and China disagree about the US's right to conduct military operations in the East China Sea.
- The US claims that international law grants it the right to conduct military operations in the international waters of the East China Sea. The US wants unrestricted access to these areas, stating that this is a matter of national interest.
- China claims that most of the East China Sea is within China's Air Defense Identification Zone (ADIZ) and that the US policy of conducting military operations in the East China Sea threatens China's sovereignty. China wants to restrict US access in the East China Sea.
- The dispute between the US and China has become tense, with both countries increasing their naval and air patrols in the East China Sea.
- Experts agree that the US's military capability is [inferior/superior] to China's for a conflict in the East China Sea.

H.2.2 Scenarios

Note: Respondents in the control group are not given any additional information.

Attack Recently a Chinese military plane shot down an American military plane over the East China Sea. China claims the American plane was trespassing in China's ADIZ. The American pilot died and the plane was destroyed in the crash.

Collision Recently there was a collision between an American and a Chinese military plane. The collision occurred because the Chinese plane was flying dangerous maneuvers around the American plane, making several close passes. On the third pass, the Chinese plane collided with the American plane. The American pilot died. The Chinese pilot just barely managed to eject and survive. Both planes were destroyed.

Non-Collision Recently an American military plane and a Chinese military plane almost collided with each other. They almost collided because the Chinese plane was flying dangerous maneuvers around the American plane, making several close passes. No one was hurt.

Accident Recently an American military plane that was conducting routine operations in the East China Sea crashed in a storm. The pilot died and the plane was destroyed.

H.2.3 Placebo Test Question: Hostile Military Intent

Given the information available in this scenario, how likely do think it is that China has plans to expand its military presence and capabilities in the East China Sea?

- Very Unlikely
- Unlikely
- As Likely As Not
- Likely
- Very Likely

H.3 Results of Placebo Test Questions

H.3.1 Coding and Analyzing the Placebo Outcomes

Each respondent provided responses to the placebo test question. The non-standardized response Y_i^N to the placebo test question is along a 5-point likelihood scale. We code "Very Unlikely" as 1, "Unlikely" as 2, "As Likely As Not" as 3, "Likely" as 4, and "Very Likely" as 5. We construct the standardized response Y_i using the method in the Democratic Peace survey experiment (see G.4.1 for details). As a robustness check, we also analyze the non-standardized placebo outcomes Y_i^N.

Let Z_i be an indicator variable for whether respondent i is told that an American pilot was killed in the scenario. For each vignette type, we estimate $\mathbb{E}(\tau_i) = \mathbb{E}[Y_i(Z_i = 1) - Y_i(Z_i = 0)]$ using $\hat{\beta}_1$ from the regression $\mathbb{E}(Y_i|Z_i) = \beta_0 + \beta_1 Z_i$. We calculate heteroskedasticity-robust standard errors for our coefficients and present the 95 and 99 percent confidence intervals in our coefficient plots.

H.3.2 Placebo Test Results

In Figure 33, we plot the distribution of the placebo outcome by treatment assignment and vignette type. The distributions of placebo outcomes are more similar for the treatment and control groups in the ENE design than in the Basic design. In Figure 34, we use coefficient plots to show our estimates of the non-standardized difference-of-means in the placebo outcome by vignette type. In the Basic design, we find significant imbalance in the placebo outcome between the treatment group and control group. In the ENE design, we do not detect any statistically significant imbalance in the placebo outcome.

Figure 33: Coercion-Provocation Survey Experiment (2016): Distribution of Responses to the Placebo Test Question by Treatment Assignment and Vignette Type

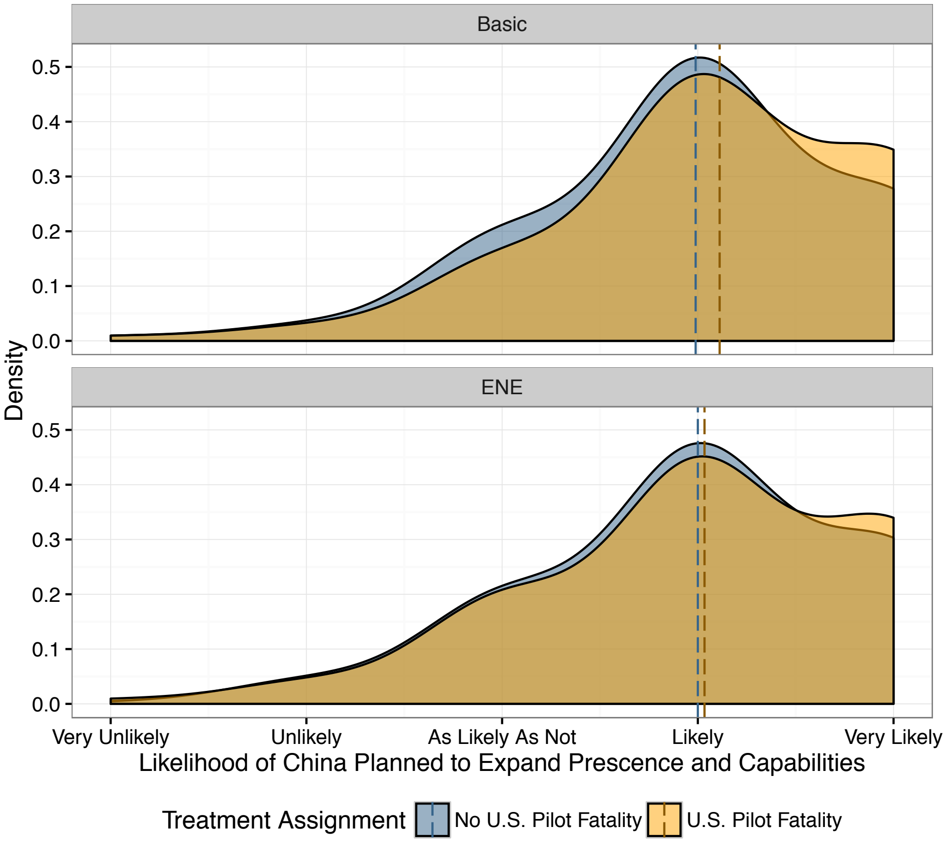
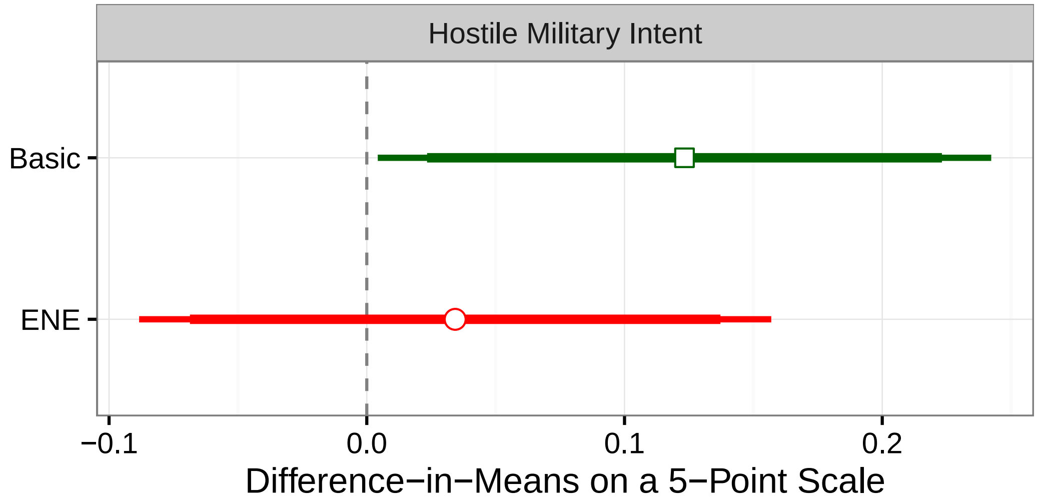


Figure 34: Coercion-Provocation Survey Experiment (2016): Placebo Test Results (Non-standardized)



I Latura's (2015) Survey Experiment

I.1 Overview of Experiment

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

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

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

I.2 Text of the Survey

In this subsection, we present the text of the survey related to the results shown in the paper.

I.2.1 Directions

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

I.2.2 Experimental Vignettes

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

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

ENE Design Text Imagine yourself in the following scenario.

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

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

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

Only a subset of female respondents see the next paragraph:

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

I.2.3 Substantive Outcome Question

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

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

I.2.4 Placebo Test Questions

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

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

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

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

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

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

I.3 Results of Placebo Test Questions

I.3.1 Coding and Analyzing the Placebo Outcomes

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

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

I.3.2 Placebo Test Results

In Figure 35, we plot the distribution of the placebo outcomes by treatment assignment and vignette type. The distributions of placebo outcomes are more similar for the treatment and control groups in the ENE design than in the Basic design. In Figure 36, we use coefficient plots to show our estimates of the non-standardized difference-of-means in the placebo outcomes by vignette type. Overall, we find that the imbalance in placebo outcomes is smaller in the ENE design than in the Basic design. Nevertheless, for two of the placebo outcomes in the ENE design, we still detect statistically significant imbalance in placebo outcomes at $\alpha = 0.05$.

Figure 35: Latura (2015): Distribution of Responses to Placebo Test Questions by Treatment Assignment and Vignette Type

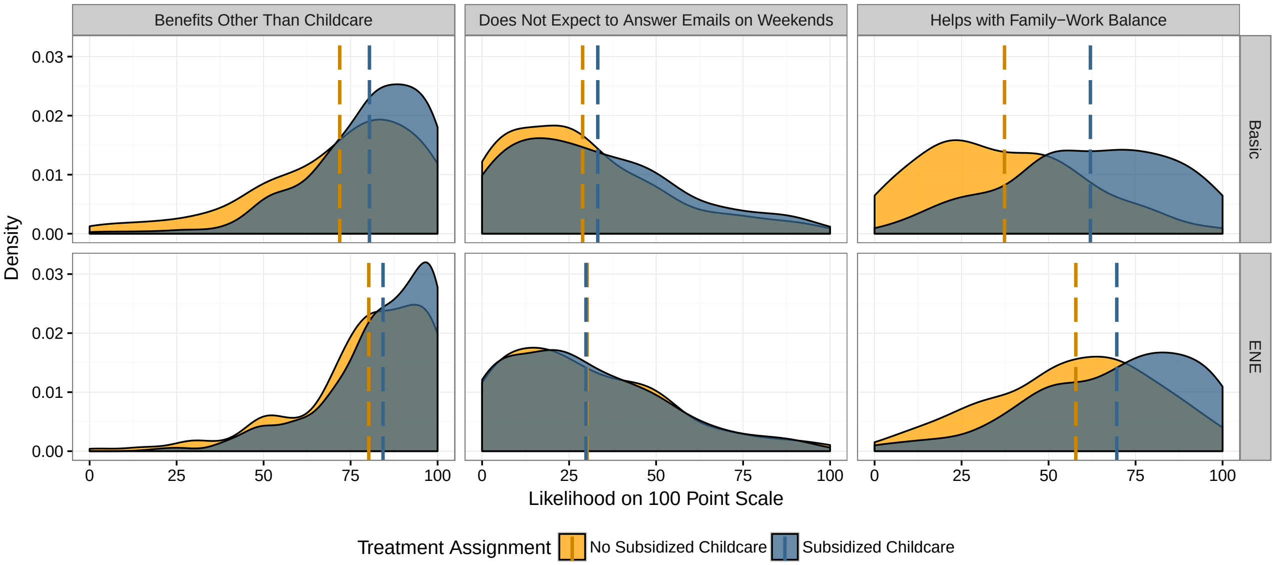
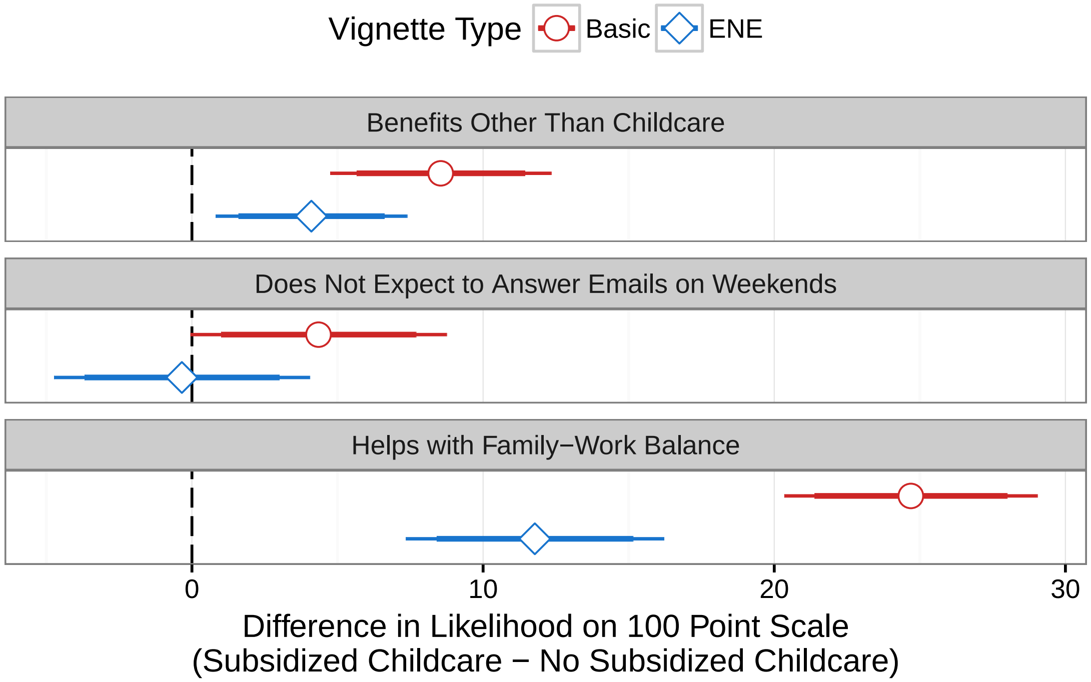


Figure 36: Latura (2015): Placebo Test Questions Results (Non-standardized)



J Replication and Expansion of DeSante (2013)

J.1 Survey Overview

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

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

J.2 Text of the Survey

In this subsection, we present the text of the survey related to the results shown in the paper.

J.2.1 Introduction

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

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

J.2.2 Welfare Application

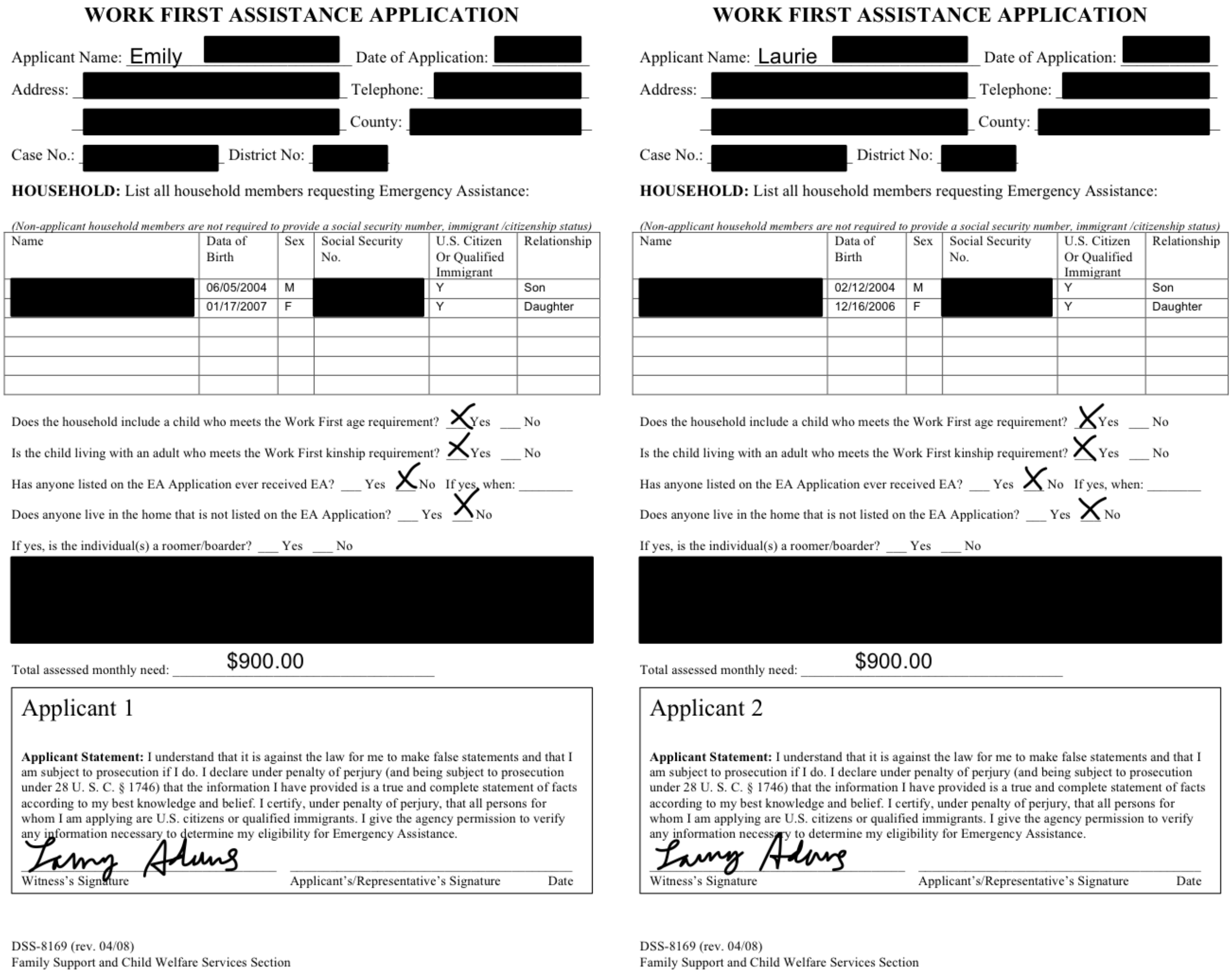
Note: In Table 4, we list the experimental conditions in the survey experiment. We vary the name of the two applicants, whether the worker quality was listed, and if worker quality was listed, the worker quality assessment.

Table 4: Experimental Conditions in Replication and Expansion of DeSante (2013)

Left Applicant Name	Right Applicant Name	Left Applicant Worker Quality	Right Applicant Worker Quality
Emily	Laurie	N/A	N/A
Emily	Laurie	Poor	Excellent
Emily	Laurie	Excellent	Poor
Latoya	Laurie	N/A	N/A
Latoya	Laurie	Poor	Excellent
Latoya	Laurie	Excellent	Poor

Below are examples of applications respondents saw.

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



WORK FIRST ASSISTANCE APPLICATION

Applicant Name: Emily [redacted] Date of Application: [redacted]

Address: [redacted] Telephone: [redacted]

[redacted] County: [redacted]

Case No.: [redacted] District No: [redacted]

HOUSEHOLD: List all household members requesting Emergency Assistance:

(Non-applicant household members are not required to provide a social security number, immigrant /citizenship status)

Name	Data of Birth	Sex	Social Security No.	U.S. Citizen Or Qualified Immigrant	Relationship
[redacted]	06/05/2004	M	[redacted]	Y	Son
	01/17/2007	F		Y	Daughter

Does the household include a child who meets the Work First age requirement? [✓ Yes] ___ No

Is the child living with an adult who meets the Work First kinship requirement? [✓ Yes] ___ No

Has anyone listed on the EA Application ever received EA? ___ Yes [✓ No] If yes, when: ________

Does anyone live in the home that is not listed on the EA Application? ___ Yes [✓ No]

If yes, is the individual(s) a roomer/boarder? ___ Yes ___ No

[redacted]

Total assessed monthly need: $900.00

Applicant 1

Applicant Statement: I understand that it is against the law for me to make false statements and that I am subject to prosecution if I do. I declare under penalty of perjury (and being subject to prosecution under 28 U. S. C. § 1746) that the information I have provided is a true and complete statement of facts according to my best knowledge and belief. I certify, under penalty of perjury, that all persons for whom I am applying are U.S. citizens or qualified immigrants. I give the agency permission to verify any information necessary to determine my eligibility for Emergency Assistance.

Larry Adams

Witness's Signature Applicant's/Representative's Signature Date

DSS-8169 (rev. 04/08)
Family Support and Child Welfare Services Section

WORK FIRST ASSISTANCE APPLICATION

Applicant Name: Laurie [redacted] Date of Application: [redacted]

Address: [redacted] Telephone: [redacted]

[redacted] County: [redacted]

Case No.: [redacted] District No: [redacted]

HOUSEHOLD: List all household members requesting Emergency Assistance:

(Non-applicant household members are not required to provide a social security number, immigrant /citizenship status)

Name	Data of Birth	Sex	Social Security No.	U.S. Citizen Or Qualified Immigrant	Relationship
[redacted]	02/12/2004	M	[redacted]	Y	Son
	12/16/2006	F		Y	Daughter

Does the household include a child who meets the Work First age requirement? [✓ Yes] ___ No

Is the child living with an adult who meets the Work First kinship requirement? [✓ Yes] ___ No

Has anyone listed on the EA Application ever received EA? ___ Yes [✓ No] If yes, when: ________

Does anyone live in the home that is not listed on the EA Application? ___ Yes [✓ No]

If yes, is the individual(s) a roomer/boarder? ___ Yes ___ No

[redacted]

Total assessed monthly need: $900.00

Applicant 2

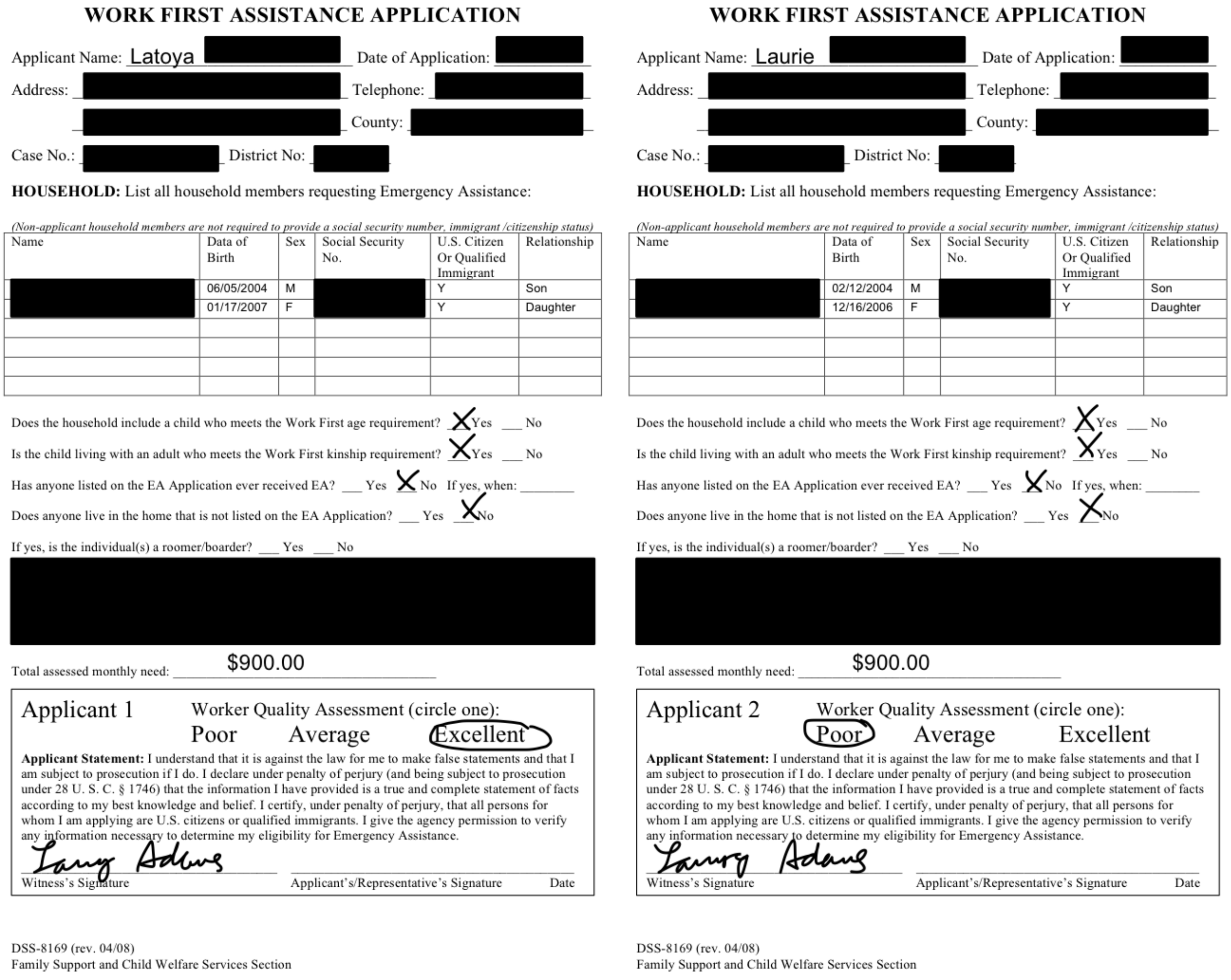
Applicant Statement: I understand that it is against the law for me to make false statements and that I am subject to prosecution if I do. I declare under penalty of perjury (and being subject to prosecution under 28 U. S. C. § 1746) that the information I have provided is a true and complete statement of facts according to my best knowledge and belief. I certify, under penalty of perjury, that all persons for whom I am applying are U.S. citizens or qualified immigrants. I give the agency permission to verify any information necessary to determine my eligibility for Emergency Assistance.

Larry Adams

Witness's Signature Applicant's/Representative's Signature Date

DSS-8169 (rev. 04/08)
Family Support and Child Welfare Services Section

Figure 38: Welfare Applications from Replication and Expansion of DeSante (2013): With Worker Quality Assessment



WORK FIRST ASSISTANCE APPLICATION

Applicant Name: Latoya [redacted] Date of Application: [redacted]

Address: [redacted] Telephone: [redacted]

[redacted] County: [redacted]

Case No.: [redacted] District No: [redacted]

HOUSEHOLD: List all household members requesting Emergency Assistance:

(Non-applicant household members are not required to provide a social security number, immigrant /citizenship status)

Name	Data of Birth	Sex	Social Security No.	U.S. Citizen Or Qualified Immigrant	Relationship
[redacted]	06/05/2004	M	[redacted]	Y	Son
[redacted]	01/17/2007	F	[redacted]	Y	Daughter

Does the household include a child who meets the Work First age requirement? [✓ Yes] ___ No

Is the child living with an adult who meets the Work First kinship requirement? [✓ Yes] ___ No

Has anyone listed on the EA Application ever received EA? ___ Yes [✓ No] If yes, when: ________

Does anyone live in the home that is not listed on the EA Application? ___ Yes [✓ No]

If yes, is the individual(s) a roomer/boarder? ___ Yes ___ No

[redacted]

Total assessed monthly need: $900.00

Applicant 1 Worker Quality Assessment (circle one):
Poor Average [✓ Excellent]

Applicant Statement: I understand that it is against the law for me to make false statements and that I am subject to prosecution if I do. I declare under penalty of perjury (and being subject to prosecution under 28 U. S. C. § 1746) that the information I have provided is a true and complete statement of facts according to my best knowledge and belief. I certify, under penalty of perjury, that all persons for whom I am applying are U.S. citizens or qualified immigrants. I give the agency permission to verify any information necessary to determine my eligibility for Emergency Assistance.

Larry Adams
Witness's Signature Applicant's/Representative's Signature Date

DSS-8169 (rev. 04/08)
Family Support and Child Welfare Services Section

WORK FIRST ASSISTANCE APPLICATION

Applicant Name: Laurie [redacted] Date of Application: [redacted]

Address: [redacted] Telephone: [redacted]

[redacted] County: [redacted]

Case No.: [redacted] District No: [redacted]

HOUSEHOLD: List all household members requesting Emergency Assistance:

(Non-applicant household members are not required to provide a social security number, immigrant /citizenship status)

Name	Data of Birth	Sex	Social Security No.	U.S. Citizen Or Qualified Immigrant	Relationship
[redacted]	02/12/2004	M	[redacted]	Y	Son
[redacted]	12/16/2006	F	[redacted]	Y	Daughter

Does the household include a child who meets the Work First age requirement? [✓ Yes] ___ No

Is the child living with an adult who meets the Work First kinship requirement? [✓ Yes] ___ No

Has anyone listed on the EA Application ever received EA? ___ Yes [✓ No] If yes, when: ________

Does anyone live in the home that is not listed on the EA Application? ___ Yes [✓ No]

If yes, is the individual(s) a roomer/boarder? ___ Yes ___ No

[redacted]

Total assessed monthly need: $900.00

Applicant 2 Worker Quality Assessment (circle one):
[✓ Poor] Average Excellent

Applicant Statement: I understand that it is against the law for me to make false statements and that I am subject to prosecution if I do. I declare under penalty of perjury (and being subject to prosecution under 28 U. S. C. § 1746) that the information I have provided is a true and complete statement of facts according to my best knowledge and belief. I certify, under penalty of perjury, that all persons for whom I am applying are U.S. citizens or qualified immigrants. I give the agency permission to verify any information necessary to determine my eligibility for Emergency Assistance.

Larry Adams
Witness's Signature Applicant's/Representative's Signature Date

DSS-8169 (rev. 04/08)
Family Support and Child Welfare Services Section

J.2.3 Allocation of Money to Applicants

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

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

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

J.2.4 Pop-up Window

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

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

We strongly recommend you click on the link above.

J.2.5 Placebo Test Questions

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

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

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

The placebo test questions are:

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

J.3 Results of Placebo Test Questions

J.3.1 Coding and Analyzing the Placebo Outcomes

Each respondent provided responses to the six placebo test questions. For each non-standardized response $R_{i,j}^N$ to placebo test question j, we code “Very Unlikely” as 1, “Unlikely” as 2, “Odds About Even” as 3, “Likely” as 4, and “Very Likely” as 5. We construct the standardized response $R_{i,j}$ using the method in the Democratic Peace survey experiment (see G.4.1 for details).

The placebo outcome of interest is the difference between respondents' assessment of the left applicant and the right applicant, i.e., the standardized paired difference $Y_{i,j} = R_{i,j}^{Left} - R_{i,j}^{Right}$. As a robustness check, we also analyze the non-standardized paired difference.

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

J.3.2 Placebo Test Results

In Figure 39, we plot the distribution of the paired differences for each placebo variable by vignette type. In Figure 40, we use coefficient plots to show our estimates of non-standardized paired difference for each placebo variable by vignette type. Overall, we find that the Covariate Control design exhibits imbalance in fewer placebo outcomes than the Basic design.

Figure 39: Replication and Expansion of DeSante (2013): Distribution of Responses to Placebo Test Questions by Treatment Assignment

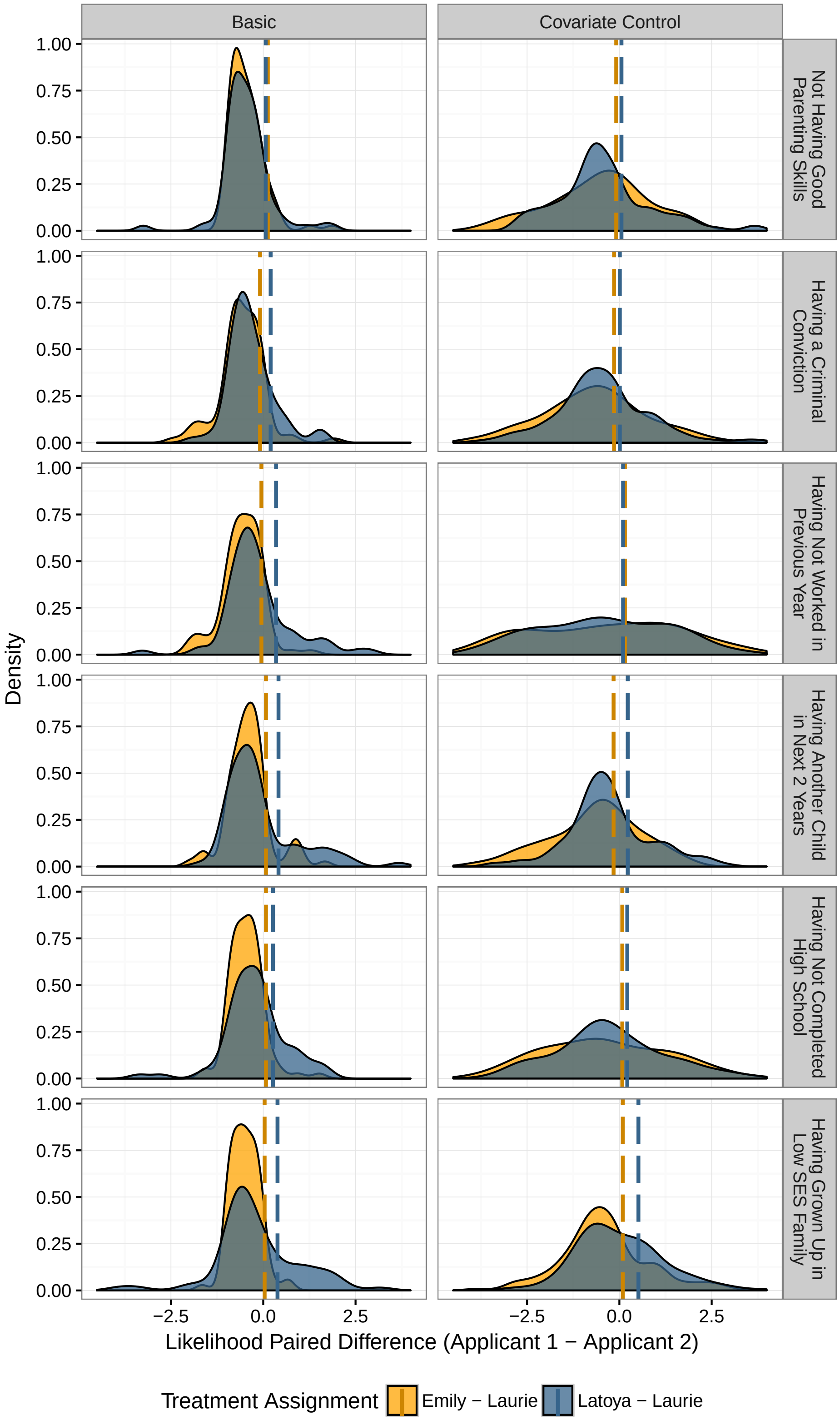
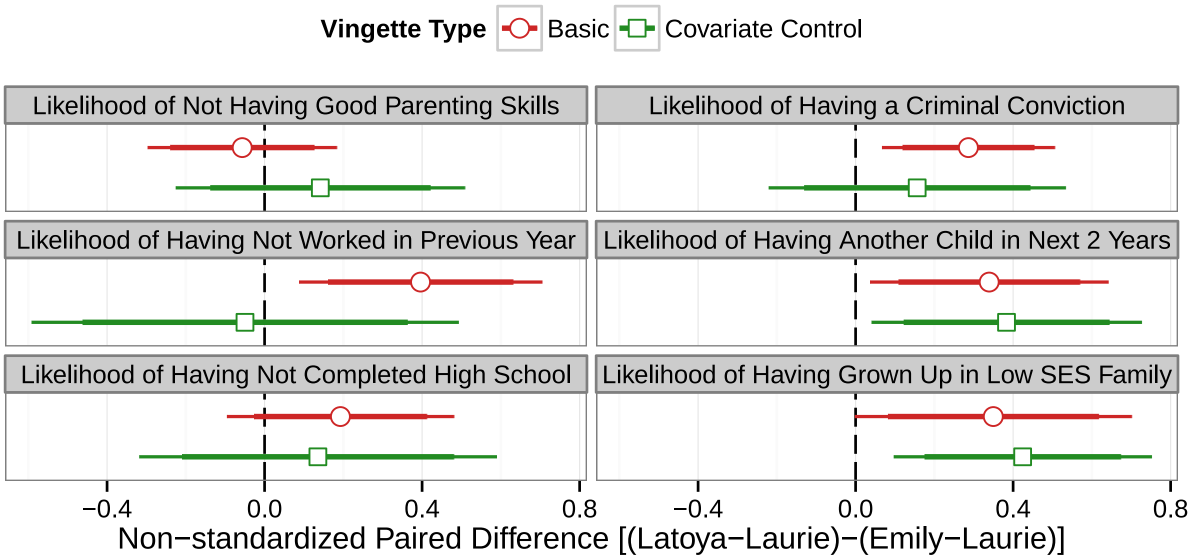


Figure 40: Replication and Expansion of DeSante (2013): Placebo Test Questions Results (Non-standardized)



References

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2017. "Analyzing Causal Mechanisms in Survey Experiments." Unpublished manuscript, March 30. Accessed July 20, 2017. http://www.mattblackwell.org/files/papers/survey-experiments.pdf.

Angrist, Joshua D., and Guido W. Imbens. 1995. "Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity." *Journal of the American Statistical Association* 90 (430): 431–442.

Caughey, Devin, Allan Dafoe, and Jason Seawright. 2017. "Testing Elaborate Theories in Political Science: Nonparametric Combination of Dependent Tests." *Journal of Politics* 79 (2).

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.

Frölich, Markus, and Martin Huber. 2017. "Direct and Indirect Treatment Effects—Causal Chains and Mediation Analysis with Instrumental Variables." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 27:1282–22.

El-Gamal, Mahmoud A., and David M. Grether. 1995. "Are People Bayesian? Uncovering Behavioral Strategies." *Journal of the American Statistical Association* 90 (432): 1137–1145.

Gamson, William A., and Katherine E. Lasch. 1983. "Political Culture of Social Welfare Policy." In *Evaluating the Welfare State,* edited by Shimon E. Spiro and Ephraim Yuchtman-Yaar. New York: Academic Press.

Gilliam, Franklin D., Jr., and Shanto Iyengar. 2000. "Prime Suspects: The Influence of Local Television News on the Viewing Public." *American Journal of Political Science* 44 (3): 560–573.

Henry, P. J., Christine Reyna, and Bernard Weiner. 2004. "Hate Welfare But Help the Poor: How the Attributional Content of Stereotypes Explains the Paradox of Reactions to the Destitute in America." *Journal of Applied Social Psychology* 34 (1): 34–58.

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.

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.

Tomz, Michael, and Jessica L. Weeks. 2013. "Public Opinion and the Democratic Peace." *American Political Science Review* 107 (4): 849–865.

VanderWeele, Tyler J. 2015. *Explanation in Causal Inference: Methods fo Mediation and Interaction.* New York: Oxford.