Confounding in Survey Experiments

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

[1]Department of Political Science, Yale University

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

This draft: July 21, 2015

Prepared for presentation at the annual meeting of
The Society for Political Methodology,
University of Rochester, July 23, 2015

For printing, consider skipping the lengthy appendix beginning on page 36.

Abstract

Survey experiments are susceptible to confounding, in ways similar to observational studies. Scenario-based survey experiments randomize features of a vignette, usually intended to manipulate subjects' beliefs about the scenario. However, the manipulation may change subject's beliefs in unintended ways, confounding causal inferences. We show how to theorize *ex ante* about these biases and how to use placebo tests as diagnostics. We illustrate with several examples, including a study of the effect of democracy on support for force: describing a country as a "democracy" makes respondents more likely to think the country is wealthy, European, majority Christian and white, and interdependent and allied with the US. We evaluate two strategies for reducing the risk of confounding: controlling for other factors in the vignette, and embedding a hypothetical natural experiment in the scenario. We find that controlling reduces the risk of confounding from controlled characteristics, but not other characteristics; the embedded natural experiment reduces the risk from all characteristics.

Contents

1 Introduction

Many questions regarding people's attitudes, preferences, and choices are hard to answer using observational survey data. Are citizens of democracies more willing to use military force against non-democracies than against democracies, as some theories of the democratic peace predict (Tomz and Weeks, 2013)? How much does the race of a potential welfare recipient affect Americans' willingness to give them welfare benefits (Desante, 2013)? To what extent does anti-immigrant sentiment arise from concerns about labor market competition or from concerns about the burden on public services (Hainmueller and Hiscox, 2010)? For several reasons, simply posing these questions to research subjects is likely to yield misleading answers. The question may fail to elicit appropriate consideration of the trade-offs involved. Direct comparisons may make respondents aware of researchers' hypotheses, prompting them to give the answers they think the researchers expect. Or subjects may feel pressure to provide socially desirable answers to questions about sensitive topics, such as race or immigration.

One response to these problems is to seek out natural variation in the causal factor of interest and then evaluate how survey responses correlate with this factor. Scholars of the democratic peace, for example, could compare the public's actual public support for using force in conflicts with democracies with its support for conflicts with non-democracies. Because the causal factor—whether the opponent is a democracy or non-democracy—is not randomly assigned, however, this strategy is susceptible to the *problem of confounding*.[1] In real-world conflicts, regime type is associated with a host of other characteristics that could affect the outcome, such as countries' wealth, political culture, and economic integration with other democracies. This correlation between democracy and other characteristics makes it hard to be confident that observed associations represent the causal effect of democracy, and not the effect of these or other characteristics.

Survey experiments appear to provide a solution to these challenges. A *survey experiment* involves the (random) manipulation of one or more features of the survey instrument, such as the phrasing of question prompts, the ordering of response categories, or the informational content of a hypothetical scenario. Tomz and Weeks (2013), for example, study the popular basis of the democratic peace by presenting respondents with a hypothetical scenario about a conflict with another country that was randomly described either as a democracy or as not a democracy. Because the experimental manipulation in a survey experiment is not observed

The most recent version of this paper, as well as our pre-analysis plans and other related materials, can be found at allandafoe.com/confounding. For helpful comments, we would like to thank Peter Aronow, Cameron Ballard-Rosa, Adam Berinsky, David Broockman, Alex Debs, Chris Farriss, Alan Gerber, Donald Green, Sophia Hatz, Susan Hyde, Josh Kalla, Gary King, Audrey Latura, Jason Lyall, Elizabeth Menninga, Nuno Monteiro, Jonathan Renshon, Bruce Russett, Cyrus Samii, Robert Trager, Mike Tomz, Jessica Weeks, 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 Summer Methods Meeting, and the Survey Experiments in Peace Science Workshop. For support, we acknowledge the MacMillan Institute at Yale University, and the National Science Foundation Graduate Research Fellowship Program.

[1] We define confounding as the existence of a common cause of treatment and the outcome that accounts for some of the treatment-outcome association. This has also been called *common-cause confounding bias* (Winship and Elwert, 2014, 32).

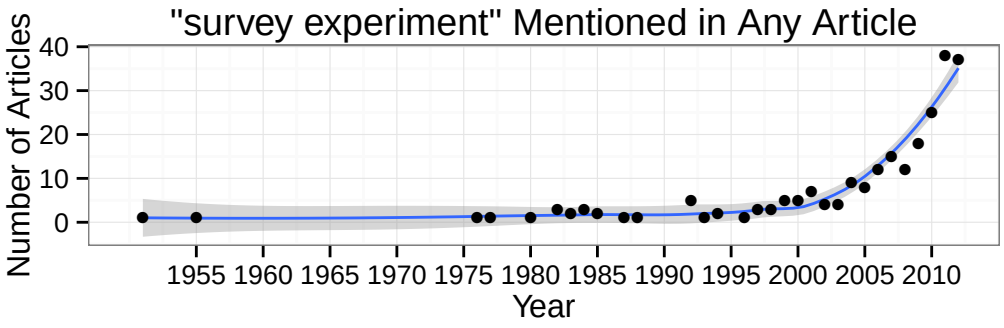


Figure 1: Mentions of "Survey Experiments" in Political Science Journals
The article counts come from searches within political science journals in JSTOR.

by respondents, they are less likely to infer the researcher's question. And since survey experiments involve random assignment, the experimental manipulation is independent of all other background causes of the outcome, thus eliminating (certain kinds of) confounding.

Due to these and other advantages, survey experiments are increasingly recognized as a powerful methodological tool (Brady, 2000; Gilens, 2002; Mutz, 2011, 8–10), and their use in political science is growing rapidly (see Figure 8 and Appendix A). Survey experiments have been used to study a diverse range of phenomena, including racial discrimination,[2] electoral appeals,[3] immigration attitudes,[4] and public support for using military force.[5] They also come in a diverse array of forms, including *scenario-based* survey experiments (which manipulate aspects of a hypothetical scenario), *priming* and *framing* experiments (which manipulate the context or wording of questions), and *list* experiments (which manipulate whether a sensitive item is included in a list of response options). Every kind of survey experiment could be susceptible to the generic problem discussed in this paper arising when the experimental manipulation is not the same as the causal factor of interest (Morton and Williams, 2010, §3.2); this paper will focus on the particular flavor of this problem within scenario-based survey experiments, which by our coding are the most common kind of survey experiment within top political science articles (see Table A).

Although survey experiments are extremely useful tools, they are not a panacea for the major challenges to causal inference. We argue and demonstrate that inferences from survey experiments are often at risk of being confounded, in a manner similar to that which would occur in the analogous observational study. This is because manipulation of one feature of a scenario will generally change subjects' beliefs about other features of the scenario. Subjects fill in details about other aspects of the scenarios in a reasonable way, using their knowledge about real-world associations. For example, in studies of the democratic peace, informing a respondent that a country is a "democracy" will also make the respondent more likely to

[2] Desante (2013); White (2007)

[3] Bullock (2011); Druckman, Peterson and Slothuus (2013); Grimmer, Messing and Westwood (2012); Tomz and Van Houweling (2008); Tomz and Houweling (2009)

[4] Hainmueller and Hiscox (2010); Sniderman, Hagendoorn and Prior (2004)

[5] Gartner (2008); Tomz (2007); Trager and Vavreck (2011)

think the country is wealthy, Christian, European, and interdependent and allied with the US. In studies of racial discrimination, providing information about the race of an individual will also make the respondent more likely to think the individual has other characteristics associated with that race, such as high/low education and socio-economic status. In studies of employment choices, informing a respondent that a firm has a generous child leave policy will also make the respondent more likely to think the firm is progressive, supportive of employees, and family friendly. In studies of anti-immigrant attitudes, informing a respondent that a group of potential immigrants is "low skill" will make the respondent more likely to think the immigrants are from certain ethnic and cultural backgrounds.

This "information leakage" (Tomz and Weeks, 2013, 853, fn. 7) is not necessarily a problem, depending on the research question. The research question can either be about the *effects of a particular feature of the world*, or the *effects of* ***the presentation*** *of a particular feature of the world*. If the question is about the presentation of a feature, then information leakage is not a problem since it is part of the causal effect of interest. Examples of this kind of question include asking about the effect of *describing* a country as a democracy, of *describing* a person as African-American, of *describing* a firm as having a generous child-leave policy. These sorts of questions are of interest for understanding the effects of the framing and presentation of information.

According to our literature review, however, in most survey experiments the research question is not about the effects of the presentation of a feature of the scenario, but the effects of (belief about) the feature itself: what is the effect of a country *being* a democracy, of a person *being* African-American,[6] of a firm *having* a generous child-leave policy. When we ask these counterfactuals we want to manipulate respondent's beliefs about the characteristic of interest, while holding fixed beliefs about other background characteristics. However, achieving this identification condition does not follow from experimental manipulation, but rather, as we will show, requires similar kinds of theorizing and methodological tools as are used for identifying causal effects in observational studies.

For studying the effects of (belief about) a feature of a scenario, the manipulation of the vignette (the actual text of the scenario) should be conceptualized as a potential instrumental variable (Z) for the causal factor of interest (D). The causal factor of interest, also called *treatment*, is the belief of the respondent about a specific feature of the scenario. For Z to be a valid instrument, it must only affect the outcome (Y) through treatment (D): manipulation of Z must not change beliefs about other unspecified features of the scenario (e) that influence the outcome (Y). When this happens, D will be correlated with e, confounding our causal inference about the effect of D. This is the problem of *confounding in survey experiments*.[7]

[6] Difficulties concerning counterfactuals about race are discussed below.

[7] There are several different vocabularies or frameworks that can be used for discussing this problem. We employ two: the framework of instrumental variables and the framework of confounding. The former is necessary to think clearly about the situation. We also employ the latter because it draws attention to the near perfect mapping to the methodological problem of confounding in analogous observational studies; we have also found it helpful for communicating these issues to non-methodologists. Some other terms, vocabularies, and frameworks that scholars have used or could use for thinking about this problem are: information leakage (Tomz and Weeks, 2013); masking (Hainmueller, Hopkins and Yamamoto, 2015); bundled treatments (Gerber and Green, 2012); estimating a specific indirect effect in mediation analysis (where D is the mediator of Z on Y); construct validity (Shadish, Cook and Campbell, 2002, ch.3).

We develop and evaluate our argument through several examples. Our primary example involves a reanalysis of a prominent survey experiment studying the democratic peace (Tomz and Weeks, 2013). We chose this study because it uses best practices in scenario-based survey experiment methodology and it addresses an important topic that is especially hard to study observationally. We also apply these methods to a study of the effect of child-care policies of firms on employment decisions, to a study of racial discrimination (Desante, 2013), and to a study of the effect of coercion on resolve (see Section 5).

We show how to theorize about possible confounding in survey experiments. We argue that respondents will update their beliefs about unspecified features of the scenario in a reasonable manner. We formalize "reasonable" updating as the updating that a rational (hence Bayesian) agent would do, given realistic beliefs about the world. This model of respondent beliefs implies that scenario-based survey experiments will be at a similar risk of confounding as an analogous observational study. For example, just as in the real-world regime type is correlated with—and possibly confounded by—GDP, trade, region of the world, religion, and race (to name just a few), so in scenario-based survey experiments will beliefs about regime type tend to be correlated with beliefs about GDP, trade, region of the world, religion, and race. This model of respondent beliefs allows scholars to deduce *ex-ante* what kinds of characteristics are most likely to confound their inference, and to focus their diagnostics and solutions towards these potential confounds. Finally, this model specifies a condition that would guarantee no confounding: the respondent must believe that variation in the causal factor of interest is as-if random in the context of the scenario.

We offer tools for diagnosing and addressing possible confounding. To diagnose confounding we recommend *placebo tests*. Placebo tests are tests of known (usually zero) effects used to evaluate a design and estimator (Rosenbaum, 2002, ch. 6; Sekhon, 2009, Dunning, 2012, §8.1.1, Dafoe and Tunón, 2014). Specifically, our placebo tests are survey questions that measure whether the experimentally manipulated features of the vignette (Z) affected subjects' beliefs in unintended ways (e). In our examples, we find that the manipulation of the vignette (Z) does affect our placebo variables, in the way one would expect if respondents are updating in a Realistic Bayesian manner. Thus, these survey-experimental designs confront a risk of confounding similar to what would confront an analogous observational opinion survey that used real-world variation in the feature of interest. To the extent that confounding is a threat to causal inference in an observational study, so will confounding be a threat to causal inference in the analogous scenario-based survey experiment.

We evaluate a solution for confounding found in the literature, which we call the *Controlled Details Design*. Controlled Details designs involve specifying potential confounds explicitly in the vignette. For example, in the democratic peace example the vignette could specify the target country's military capabilities and trade with the US (as is done in Tomz and Weeks, 2013). Conjoint analysis is a form of Controlled Details design that typically involves tabular presentation of details and often a large number of controlled details (Hainmueller, Hopkins and Yamamoto, 2014, 2015). We theorize and find that Controlled Details designs operate like control strategies in observational studies: they tend to reduce imbalance on characteristics that are explicitly controlled for or are similar to the controls, but not on other characteristics. We consider some issues and limits with Controlled Details designs, such as changing the causal estimand, respondent exhaustion, implausible combinations of

controls (on the previous points, see also Hainmueller, Hopkins and Yamamoto, 2014, 2015), and the creation and amplification of biases through controls.[8]

Finally, we introduce a novel design: the *Embedded Natural Experiment Design*. Embedded Natural Experiments consist of scenarios that describe a hypothetical as-if random source of variation in the causal factor of interest, as perceived by the respondent.[9] For example, in the study of the effect of subsidized child-care, we inform respondents that the firm has a lottery granting some employees subsidized child-care, which the respondent either (hypothetically) won or did not win. In a study of the effect of coercive harm, US and Chinese planes are described flying dangerously close to each other. In the control condition they nearly collide. In the treatment condition they collide, killing one of the pilots. Given a plausible natural experiment (a story where the cause was plausibly as-if randomly assigned), treatment assignment will not change the beliefs about background characteristics of respondents who update in a Bayesian way. A story about an as-if random process thus offers a way to achieve the identification conditions that otherwise allude researchers using scenario-based survey experiments: to make respondents' beliefs about all background characteristics in the scenario independent of treatment assignment. Our results confirm this conjecture: the embedded natural experiment design is the most successful at reducing imbalance on background characteristics, even as much or more so on those very characteristics that were explicitly controlled in the Controlled Details design. We discuss concerns about weak manipulations, implausible natural experiments, and generalizability.

In summary, survey experiments are extremely useful tools for studying the determinants of people's attitudes, preferences, and choices, especially when researchers are interested in the effects of describing or presenting a scenario in a particular way. When used to infer the effects of (subjects' beliefs about) specific features of the scenario, however, survey experiments face barriers to causal inference similar to those in observational studies. In particular, they face the risk that the apparent effects of (subjects' beliefs about) the factor of interest may be caused by (their beliefs about) other causes of the outcome. We show how the risk of confounding can be anticipated *ex ante*, diagnosed using placebo tests, and minimized using Controlled Details and Embedded Natural Experiment designs. Better understanding of these challenges and tools will improve our ability to use survey experiments to draw credible causal inferences.

2 Why Random Assignment of Vignettes is Not Enough

2.1 Two Kinds of Research Questions

Scenario-based survey experiments may be used to investigate two different kinds of research questions. The first kind concerns the effects of particular features of the world, such as the effect of the regime type of an opponent country on popular support for war. The second

[8] Hainmueller, Hopkins and Yamamoto (2015) introduce the term *masking*, which is related to what we call *confounding* or a manipulation (Z) that is not a valid instrument for treatment (D). They "define [masking] as the extent to which estimated ACMEs change in the presence of other attributes".

[9] Following Sekhon and Titiunik (2012) and Dunning (2012), we define a "natural experiment" as an observational setting in which causes are assigned haphazardly, and ideally in a manner that is as good as random.

kind of research question concerns the effects of presenting information about the world in a particular way. Our focus in this paper is on the first class of questions, those concerned with the effects of particular features of the world. Given the challenges that we later show to be endemic to such studies, however, it is worth briefly considering the merits of asking the second type of research question.

Consider, for example, a researcher interested in how popular support for war is affected by how the media portrays the political regime of the opponent country. A plausible way of investigating this is to conduct a survey experiment in which countries are randomly assigned to be described as "democratic" or "non-democratic." In this experiment, the fact that presenting information in a certain way may influence subjects' beliefs about other characteristics of the country is not necessarily a problem, since such inferences may be part of (i.e., a mechanism of) the effect of interest to the researcher. To use the language of instrumental variables, the causal quantity of interest in this experiment is the "intention-to-treat" (ITT) effect, which can be validly estimated with a simple difference of means.[10]

The relatively weak assumptions required for ITT estimation suggest one potential response to the problem of confounding in survey experiments, which is to redefine the causal quantity of interest from the effect of *being* X to the effect of *being described as* X, where X could be "a democracy" (for a country), "African-American" (for a welfare applicant), or "low-skilled" (for an immigrant). While internally valid, this analytic move is likely to be unsatisfying to applied researchers truly interested in effects of the first kind, which are often of great theoretical and policy importance. Thus before detailing the challenges to estimating such effects, we wish to emphasize that these challenges should not be regarded as a reason for restricting scholarly attention to the effects of information presentation only. Rather, such decisions should take into account the substantive significance of the research question as well as the ease and certainty with which it can be answered.

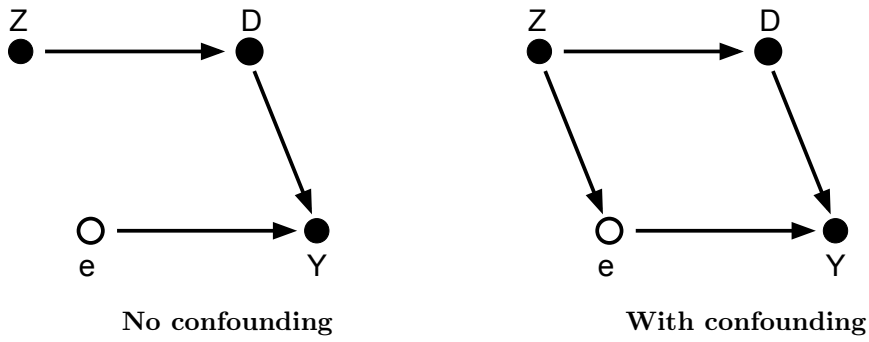
2.2 The Vignette as an Instrumental Variable

We now restrict our focus to addressing the first kind of question, about the effects of (beliefs about) a feature of a scenario on a subject's response. To get at this, the experimentalist manipulates the description of the scenario (denoted as Z) so as to manipulate the subjects' beliefs about the intended feature of the scenario (denoted as D). For example, a study of the democratic peace may manipulate whether a country in a scenario is described as a "democracy" or "not a democracy" (Z), in order to manipulate the subjects' beliefs about the regime type of the country in the scenario. The researcher may then ask the respondent about their support for using force in such a scenario (Y). The researcher would like to infer that the observed effects (the change in Y due to Z) tell us something about the effects of regime type (D) on public support for using force.

This situation is depicted in Figure 2. For ease of exposition, we assume that D is dichotomous. Several conditions are required so that we can draw inferences about the

[10] ITT effects are identified under the assumptions of random assignment and the Stable Unit Treatment Value Assumption (Angrist, Imbens and Rubin, 1996, 446–7). The external validity of these effects depends on further assumptions, such as whether the experimental and real-world versions of the instrument are sufficiently similar (Hernán and VanderWeele, 2011).

Figure 2: Vignettes as Instruments



● represents an observed random variable and ○ represents a latent random variable.
$(A \longrightarrow B)$ means that A affects B. The absence of an arrow implies the absence of causation.

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

effects of D on Y. Specifically, identifying the sign of the complier average causal effect (CACE) of D on Y requires the following assumptions:

$\mathcal{A}_1$ (Independence of Manipulation): Z itself must not be confounded. Fortunately, random assignment of Z guarantees that Z will be independent of all factors that are not affected by Z. Formally, where W denotes pre-survey characteristics such as the gender and age of the respondent or the political environment of the survey, random assignment of Z implies: $Z \perp\!\!\!\perp W$. This implies that Z will not be confounded by pre-survey characteristics.

$\mathcal{A}_2$ (Monotonicity and Non-Zero First Stage): the direction of the effect of Z on D for every respondent must be known, and for some respondents it must be non-zero. $\mathcal{A}_2$ is usually reasonable. For example, in studies of the democratic peace we assume that for every respondent, receiving the vignette describing a country as a democracy ($Z = d$), as opposed to non-democracy ($Z = nd$), will make the respondent more likely to think the country in the scenario is a democracy ($D = d$) rather than non-democracy ($D = nd$). If $\mathcal{A}_2$ is false, for example if some unknown set of respondents draws the opposite inference about D, then the association between Z and Y cannot even tell us about the sign of the average effect of D on Y.

$\mathcal{A}_3$ (Exclusion Restriction): Z must not affect (Y) except through D (denote this as $Z \not\rightarrow e$). Random assignment of Z provides no leverage over $\mathcal{A}_3$. Rather, $\mathcal{A}_3$ requires making a social scientific argument that Z does not affect Y except through its effect on D. For some Z this will be plausible, others it will not be plausible. When $\mathcal{A}_3$ is

false, Z affects e ($Z \rightarrow e$). The association between Z and Y then no longer provides a clear inference about the effect of D on Y. Even if we see a significant association between Z and Y, it could be that D has no effect, and all of the effect of Z is due to e. Or, even worse, it could be that D has the opposite effect as the observed association between Z and Y, but the effect of e swamps the effect of D. When $\mathcal{A}_3$ is false, so that $Z \rightarrow e$, we say that the survey experimental design is confounded. Figure 2 represents this issue using causal graphs, where $\mathcal{A}_3$ is true in the left causal graph and false in the right causal graph.

Given $\mathcal{A}_1$, $\mathcal{A}_3$, and $\mathcal{A}_2$, a significant positive association between Z and Y provides evidence that the CACE of D on Y is positive. In addition, if scholars measure the causal factor of interest (D) then they can estimate the magnitude of the CACE using an IV estimator.[11] We demonstrate how this is done in Section 4.4.

2.3 A Realistic Bayesian Model of Respondent Beliefs

In order to think about $\mathcal{A}_3$ we need to have a model of how respondents think about a scenario, and how they revise their beliefs about the scenario after being given information. Specifically, we need a theory that speaks to $Z \rightarrow e$: the effect of the manipulation of the vignette on the respondents' beliefs about other aspects of the scenario that are not consequences of the causal factor of interest (D) and that affect the respondents' outcome answer (Y). We propose a *Realistic Bayesian Model* of respondent beliefs and contrast it with the implicit *No-Confounding Null Model* that would be required to be true in order for us to not worry about confounding.

2.3.1 No-Confounding Null Model

The manipulation of the vignette manipulates only the intended beliefs of the respondent. This is the model implicitly assumed by any analysis of a scenario-based survey experiment that does not worry about confounding. Formally we assume that $f(e|Z = d) = f(e|Z = nd)$, where $f(e|Z = d)$ is the probability mass function describing the respondents' beliefs about characteristics e in the scenario, for respondents who receive vignette with $Z = d$.

2.3.2 Realistic Bayesian Model

Realistic: Respondents have a model of the distribution of characteristics of a scenario based on the actual distribution of the characteristics of similar scenarios in the real-world. For instance, when told that the population of a country suffers from malaria, realistic respondents think this country has a high probability of being in the tropics, since in the real-world the proportion of countries suffering from malaria is much greater for countries in the tropics than those that are not. Formally, this assumes that $f(D, e, Y) = f_r(D, e, Y)$, where $f(D, e, Y)$ is the multivariate probability mass function[12] describing the respondents' beliefs about characteristics D, e, and Y in

[11] Survey experiment scholars have long advocated for manipulation checks to ensure vignettes have affected subjects' beliefs about the causal factor as intended (Mutz, 2011, 102–104).

[12] For ease of exposition we confine ourselves to discrete variables.

the scenario, and $f_r(D, e, Y)$ is the probability mass function describing the real-world distribution of these characteristics.

Bayesian: Respondents revise their beliefs according to the laws of conditional probability (Bayesian updating). Respondents will use characteristics specified in a vignette to condition their beliefs about unspecified characteristics, as well as the meaning of the other words employed in the vignette. Formally:

$$f(D, e, Y|X = x) = \frac{f(X = x|D, e, Y)f(D, e, Y)}{f(X = x)}$$

So long as we have information about the actual association of characteristics in the real-world this model then yields precise implications about respondents beliefs, unconditionally and conditional on any set of characteristics being true.[13]

A third model that scholars might consider is an Ignorant Bayesian Model in which respondents have non-realistic beliefs, perhaps reflecting the portrayal of the world by media, but still update in a Bayesian manner. This model would also yield precise predictions if we first measured respondents beliefs about the world.

We hypothesize that the Realistic Bayesian Model better accounts for respondents' beliefs than the No-Confounding Null. Specifically, this means that describing a country in a scenario as a "democracy" vs "non-democracy" will lead respondents to believe that this country is more likely to have other characteristics correlated with democracies in the real-world, such as being in Europe, having liberal values and norms, being wealthier, being more economically interdependent, and sharing strategic interests with the U.S. We will now discuss how to diagnose confounding, $f(e|Z = d) \neq f(e|Z = nd)$, by developing measures for e.

3 Diagnosing and Addressing Confounding

3.1 Diagnosing Confounding Through Placebo Tests

To diagnose confounding in survey experiments we propose the use of placebo tests. Specifically, our placebo tests are survey questions that measure whether the experimentally manipulated features of the vignette (Z) affected subjects' beliefs in unintended ways.[14] For

[13] 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. For example, humans often believe the probability of a scenario increases as restrictive details are added, and do not give enough consideration to alternative hypotheses (Bowers and Davis, 2012). However, there does not yet exist a model of human belief updating that, in our view, offers as good a first approximation as the Bayesian model. Any such alternative model can be empirically evaluated against the Bayesian model using the empirical strategy we use in this paper.

[14] It is possible that the placebo questions themselves will actually induce confounding. For example, respondents might not think about the religion of the country until they are asked about it. In general, prior questions in surveys can impact responses in subsequent questions (Benton and Daly, 1991; Gaines,

instance, in the democratic peace survey experiment, we use placebo tests to evaluate whether describing the target country as a dictatorship makes subjects more likely to think the target country is located in the Middle East.

The best placebo variables are *valid* and *powerful* (the following material is from Dafoe and Tunón, 2014). A valid placebo is a variable that should have the same distribution across treatment levels ("balance") if the identifying assumptions are true: $P(b|ia) \approx 1$, where b denotes balance in the placebo across levels of the manipulation Z, and ia denotes that the identifying assumption is true. For testing as-if random treatment assignment, a variable is a valid placebo if it is not affected by treatment. Scholars typically look to pre-treatment variables for valid placebos since these can not be affected by treatment, though post-treatment variables can also be valid so long as we are confident that treatment does not affect them.[15]

A powerful placebo is one that should be dependent with treatment if our identifying assumptions are false (that is, there is confounding): $P(im|\neg ia) \approx 1$, where im denotes imbalance (some divergence from the placebo prediction). The best examples of powerful placebos are those characteristics that we think are most likely to confound the association: they should be on confounding causal pathways so that, under confounding, they are dependent with treatment and affect the outcome. A *dispositive placebo* is a placebo that is valid and powerful. These terms can similarly be applied to sets of placebos. A *dispositive placebo test* is then a test of a set of placebos which is jointly valid and powerful. Dispositive placebo tests generate the strongest evidence (largest likelihood ratio) for or against confounding.

Dispositive placebo tests are an ideal. In practice there are trade-offs between validity and power. If we confine ourselves only to valid placebos we may fail to diagnose confounding due to causes for which there are not valid placebos. Accordingly, we recommend that scholars choose a set of placebos at different points on the frontier of maximum validity and power. In our examination of the democratic peace our most valid placebos ask about characteristics that are unlikely to be affected by regime-type on the time scales of the scenario, such as *region* of the country, *oil reserves*, *religion*, and *race*. Given our Realistic Bayesian model of respondents, a necessary condition for a placebo to be powerful is a real world correlation between the characteristic and regime type. Accordingly, we systematically examined correlates of democracy to determine which characteristics were candidates as powerful placebos (see Table 5); all of our placebos showed real world imbalance. A second condition for a placebo to be powerful is that it is on a confounding causal pathway: it affects the outcome or is dependent with factors (other than treatment) that affect the outcome. Some placebos that meet these criteria as more powerful, but still plausibly valid,

Kuklinski and Quirk, 2007; McFarland, 1981; Schwarz and Hippler, 1995; Siegelman, 1981). For this reason, we recommend that, in general, placebo questions be asked after the outcome question. To investigate the extent to which our placebo questions are, themselves, affecting the outcome we vary whether the placebo questions are asked before or after the outcome question in our 2014 pilot study. For all vignette types, we do not find evidence that the order of the questions affected the outcome (all two-sided p-values > 0.05).

[15] We refer to valid placebos for testing as-if random treatment assignment as *randomization valid* placebos, or simply *valid* placebos. If the identifying assumptions are weaker, such as ignorability, then a placebo will only be valid if it is also a cause of the outcome. A valid placebo for testing ignorability is *ignorability valid*. A placebo that is randomization valid and powerful will also be ignorability valid. Since the ideal placebos are valid and powerful in any case, we limit our discussion to randomization validity, which is a simpler concept.

are *alliance* status of the country, *trade* with the US, whether the country has performed a *joint military exercise* with the US, and *FDI* in the US. See Appendix B.5 for the full text of and additional details about our placebo questions. We considered other variables but rejected them as insufficiently valid or powerful; for example, while *population* was relatively valid, it was not a powerful placebo because regime type has a low correlation with population size in the real-world.

3.2 Addressing Confounding: Controlled Details Designs

Some scholars are implicitly aware of the possibility of confounding in scenario-based survey experiments and adopt what we call a *Controlled Details design*.[16] The work that most explicitly articulates these problems that we are aware of is Tomz and Weeks (2013). These authors refer to the problem of confounding as "information leakage," noting that manipulation of the regime-type of the target country may lead respondents to draw inferences about other characteristics of the target country such as whether it is "also an ally, a major trading partner, or a powerful adversary" (Tomz and Weeks, 2013, 853).

In recognition of this threat to inference, many survey experiments employ *Controlled Details designs*: a vignette that includes additional details to control respondent's beliefs about these potentially confounding characteristics (examples include Bechtel and Scheve, 2013; Desante, 2013; Grieco et al., 2011; Johns and Davies, 2012). For example, Tomz and Weeks specify in their scenarios alliance status, trade with the US, and military capabilities. The prominent tool of *conjoint analysis* (Hainmueller, Hopkins and Yamamoto, 2014) can be regarded as a form of Controlled Details design in which (often many) aspects of a scenario are controlled.[17] The principle of the Controlled Details design is the same whether the controls are held fixed for every respondent or experimentally manipulated, just as the principle in observational studies behind conditioning on a confounder is the same whether you stratify within a single level of a confounder, or average across stratum-specific effects (such as by using regression, matching, or inverse-probability weighting).[18]

Consistent with the Realistic Bayesian Model of respondents, we argue that Controlled Details designs will operate similar to conditioning strategies in observational studies: they will reduce or eliminate confounding on the variables specified, and often reduce confounding

[16] Our early thinking on this topic was discussed by Cyrus Samii in a March 2011 blog.

[17] Work on conjoint analysis (Hainmueller, Hopkins and Yamamoto, 2014) shares themes with this paper in being concerned about improving causal inference in scenario-based survey experiments. However, Hainmueller, Hopkins and Yamamoto (2014) address a different problem. Hainmueller, Hopkins and Yamamoto (2014) confront the problem that some survey experiments manipulate multiple aspects of a vignette, such as a design that varies the ethnicity of a person by altering the immigrant's "face, name, and country of origin" (Hainmueller, Hopkins and Yamamoto, 2014, 2; see also Bullock, 2011, ft 15). By manipulating multiple aspects of a vignette in a collinear manner, it is not possible to identify the specific effects of each of these *words*. Conjoint analysis solves this problem by independently manipulating each relevant feature of a vignette. The problem of confounding that we discuss remains even if one manipulates a *single word* of a vignette, or multiple single words in a factorial design as is done in conjoint analysis. We are concerned with how manipulation of an aspect of a vignette, be it a single or multiple words, will change a set of beliefs in addition to the beliefs that the scholar wishes to manipulate.

[18] These different approaches to conditioning will change the causal estimand since it will weight observations differently depending on the value of the covariates, but are all designed to recover unbiased (local) causal effects by removing confounding.

on other characteristics that are correlated with the controls. However, as with observational control strategies, Controlled Details designs will not address confounding on characteristics not correlated with the controls, and could even induce or amplify confounding.

For our Controlled Details design we use Tomz and Weeks' (2013) design, which explicitly mentions whether the country has a military *alliance* with the US, *trade* with the U.S., and its non-nuclear military *capabilities*. Our *Basic* design is then this same design without these controls. As summarized in Figure 3, we find that controlling for these variables reduces imbalance on them. In addition, imbalance on the similar variables of *FDI* in the US and likelihood of a *joint military exercise* also become close to zero. However, all other potential confounds that we examined (*GDP* per capita, *religion*, *race*, *oil reserves*) remained significantly imbalanced and by the same magnitude. Consistent with our predictions, controlling reduces imbalance on the variables controlled for, and on correlated variables, but not on other variables.

3.3 Addressing Confounding: Embedded Natural Experiments

We also introduce a new strategy for overcoming confounding in scenario-based survey experiments: basing the hypothetical scenario on a natural experiment, a source of as-if random variation in the causal factor of interest. Just as natural experiments in the real-world allow observational studies to identify plausibly as-if random variation in the causal factor of interest, consistent with the Realistic Bayesian Model we conjecture that scenarios based on plausible natural experiments will eliminate confounding. The reason for this is as follows.

In a plausible natural experiment, by definition, variation in the causal factor is perceived to be as-if random. This implies that treatment is independent of the potential values of all other variables: $D \perp\!\!\!\perp X(D = d) \quad \forall d$, where $X(D = d)$ denotes the value that X would have taken if D had been set to d. Since this implies ignorable treatment assignment, in the mind of a Realistic Bayesian respondent an as-if randomly assigned causal factor cannot be confounded. Telling a Realistic Bayesian respondent about the specific value of treatment, conditional on the natural experiment, provides no information to the respondent about anything that is not a consequence of treatment.

In our study of the democratic peace, our embedded natural experiment (ENE) involves two narratives. The first concerns a fragile democracy being held together by its popular president; the haphazard outcome of an assassination attempt then determines whether the country stays democratic or becomes ruled by a military regime. The second is about a fragile dictatorship; likewise, the outcome of the assassination attempt determines whether pro-democracy forces topple the dictatorship or the country remains an autocracy.[19] The exact text that respondents are assigned to read is presented in Table 1.[20]

[19]The inspiration for our vignette design comes from Jones & Olken's (2009) observational study using the outcome of assassination attempts as a natural experiment to study democratization.

[20]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

As we conjecture, the *Embedded Natural Experiment design* exhibits the least confounding. It was largely balanced on all placebo variables, significantly more balanced than the *Controlled Details design*. Even considering the characteristics explicitly or implicitly controlled for in the *Controlled Details design*, the *Embedded Natural Experiment design* was superior or as good. We conclude that Embedded Natural Experiments work similar to natural experiments in observational studies: when credible ones exist, they are extremely useful for causal identification since they eliminate confounding on all factors. Just as observational natural experiments are hard to find, however, so are Embedded Natural Experiments often hard to construct. Finally, as with observational natural experiments, Embedded Natural Experiments identify a particular local causal effect which may not necessarily be of interest or generalize; however, this local nature of the causal estimand is equally true for Basic and Controlled Details designs, except that with them it is often less clear what is the distribution of the (respondent's beliefs about) background characteristics.

4 An Application to the Democratic Peace

4.1 Survey Experimental Study of the Democratic Peace

Scenario-based survey experiments have been increasingly used to test important theories in international relations (e.g. Grieco et al., 2011; Hainmueller and Hiscox, 2010; Tomz, 2007), and in particular the democratic peace. Mintz and Geva's (1993) and Rousseau's (2005) studies show that Americans express greater support for going to war against dictatorships than democracies. More recent studies have controlled for other aspects of the opponent country. Johns and Davies (2012) test whether subjects would respond differently to democracies versus autocracies, as well as to the majority religion of the opponent country (Christian versus Muslim). In several large-N survey experiments, Tomz and Weeks find that subjects are more likely to support military strikes against non-democracies than democracies, even after controlling for the target country's military capabilities, trade, and alliances (Tomz and Weeks, 2013). Further, Tomz and Weeks (2013) found that in vignettes involving democracies (versus non-democracies), respondents had similar expectations of the costs of conflict and the probability of failure, but decreased perceptions of threat and increased perceptions of the immorality of a US attack. This provides insight into the possible mechanisms of the democratic peace.

Our contribution to this literature is to evaluate the extent to which these scenario-based survey experiments provide evidence of the effect of the causal factor of interest—the regime-

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.

type of the country in the scenario—as opposed to the possible effects of other characteristics of the country. As we will show, we as yet cannot rule out that the possibility that the U.S. public's aversion to using force against a country described as democratic is due to beliefs about other features of the country. Such possible confounders include the target country's liberal culture, religion, race, history of conflict with the West, willingness to be a responsible global citizen, the orientation of its economy, or other factors correlated—in the real-world and in the minds of respondents—with its regime-type. Among other research strategies, future survey experiments that are sensitive to these challenges will be better equipped to further our understanding of this important phenomenon.

4.2 Survey: Scenario and Questions

Our survey was fielded July 1-3 2015 using the Qualtrics survey platform on 3000 American respondents recruiting using Amazon's Mechanical Turk. This subsection briefly summarizes the survey designs; see Appendix B for a more complete description of the survey; see Appendix D for the full summary of our analysis of the data.

Our survey closely follows (Tomz and Weeks, 2013). Table 1 summarizes our three vignette types. In all cases they read the basic scenario (**Scenario1** and **Scenario2**). The **Basic** design just manipulates regime type. The **Controlled Details Design** also provides information about the country's *military capability*, *trade*, and *alliance*. The **Embedded Natural Experiment Design** consists of the ENE narrative, plus the basic scenario.

After reading the vignette the respondents received questions related to the placebos, the outcome, the mechanisms, and the treatment, as well as demographic questions. The order of these questions were in part randomized, and depended on the survey wave. Appendix B provides more detail.

4.3 Results: Imbalance Exists and Is Similar to Confounding in Observational Studies

Figure 3 summarizes the main results for the placebo tests: *region*, *GDP*, *religion*, *race*, *oil reserves*, *alliance*, *trade*, *joint military exercise* and *FDI*. We also included the placebo test outcomes for *military spending*, but discuss it separately because it was not a dispositive placebo.[21] More detailed results are presented in Appendix D.

The data is highly consistent with our hypotheses. The **Basic** design exhibits evidence of confounding (see red circles). Every placebo is imbalanced (significantly different from zero at $\alpha = 0.05$) for the **Basic** design. The imbalance in the **Basic** design is in the direction predicted by the Realistic Bayesian Model. Namely, countries described as a democracy are more likely to have 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 or white, to not have large oil reserves, to have an alliance with the U.S. or have conducted a joint military exercise with the U.S., or to trade with or invest in the U.S.

[21] In our real-world data, we did not find a significant difference in military spending between democracies and non-democracies. Accordingly, if respondents behave as Realistic Bayesians, we should be less likely to detect imbalance for it; it is not a powerful placebo test. Nevertheless, we included this placebo test question because Tomz and Weeks included details about nonnuclear military capabilities in their vignettes.

Table 1: Text of the Three Vignette Types

Scenario1: "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."
Scenario2: "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."
(1) Basic: **Scenario1** + "[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.]" + **Scenario2**
(2) Controlled Details: **Scenario1** + **Basic** + 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. + **Scenario2**
(3) Embedded Natural Experiment: "We are going to describe a hypothetical country, called Country A. Please read this passage about Country A carefully. 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.**] + **Scenario1** + **Scenario2** ——————————————————— 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.**] + **Scenario1** + **Scenario2**

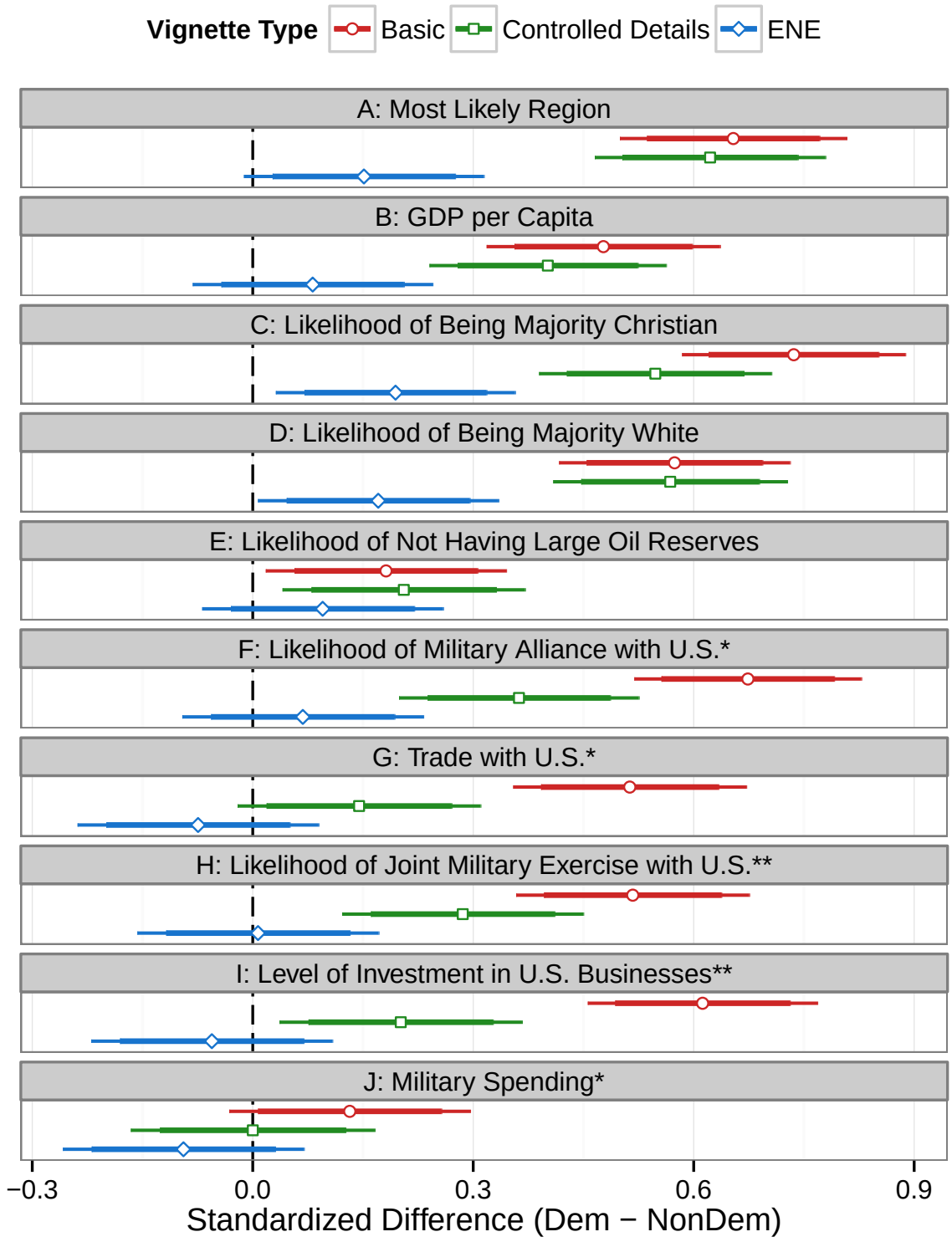


Figure 3: Results of Placebo Tests by Vignette Type

This figure summarizes the evidence for and against confounding for each of the three vignette types. The **Basic** design appears confounded: the difference is significantly greater than 0 for each placebo variable at $\alpha = 0.05$. The **Controlled Details** design exhibits great imbalance on those variables that were not controlled and even some imbalance on those that were. The **Embedded Natural Experiments** design only exhibits imbalance on three placebo outcomes at $\alpha = 0.05$. The x-axis is the estimated difference between the level of the placebo under the democracy condition ($Z = d$) and the non-democracy condition ($Z = nd$), each placebo standardized so that the effects can be seen on the same scale. The thick line is the 95% confidence interval, the thin line the 99% confidence interval. $*$ denotes characteristics of the target country we explicitly controlled for in the Controlled Details. $**$ denotes characteristics of the target country we implicitly controlled for in the Controlled Details

The **Controlled Details** design exhibits significant imbalance on placebos that were not controlled, specifically *region*, *GDP*, *religion*, *race*, and *oil reserves*. The **Controlled Details** design exhibits smaller, though still statistically significant, imbalances on placebos that were explicitly controlled, *alliance* and *trade*, and on variables that were implicitly controlled, *joint military exercise* and *FDI*.[22]

The **Embedded Natural Experiments** design exhibits the least amount of imbalance, being balanced (not significantly different from zero at $\alpha = 0.05$) on all placebos except *region*, *religion*, and *race*. Even for these three placebo outcomes, the imbalance in the ENE design is much smaller than the imbalance in the other two designs. The ENE manipulation also changed the reported beliefs about the country's regime type and support for using force by similar or greater magnitudes than the other designs (see Figure 4), so it is not that the ENE design just failed to change the respondents' beliefs about anything.

4.4 Estimating a (Local) Average Treatment Effect

So far we have confined ourselves to the modest goal of estimating the sign of the effect of D on Y. However, scholars often want to estimate the magnitude of the effect of D on Y. To do so, scholars need to employ an IV estimator which will lean on several additional assumptions. First, one must be able to correctly measure D. Second, one must know the correct functional form of the effect of D on Y.

Given those assumptions, it is possible to estimate a local average treatment effect (LATE) of D (Morgan and Winship, 2007, §7; Imbens and Angrist, 1994; Angrist and Pischke, 2008, §4; Sovey and Green, 2010). If D is dichotomous, then one can estimate the complier average treatment effect, which is the average treatment effect for respondents who changed their beliefs about D because of Z. The complier average treatment effect might be closer to our desired causal estimand, since it gives weight to those respondents who paid enough attention to the vignette to process the change in regime-type. On the other hand, suppose our desired causal estimand is the effect of regime-type on public opinion in a real-world crisis. For this estimand, the complier average treatment effect could overestimate the effect of regime-type, since respondents who do not pay attention to the details of a vignette may also be less likely to pay attention to the details of a real-world crisis. In this case the ITT estimate may be closer to the actual average treatment effect.

Further, if D has multiple levels then things become more complicated. An IV estimator will weight unit level causal effects in two ways. As before, it will give more weight to respondents whose beliefs about D are more sensitive to Z. But in addition, it will give more weight to the kinds of changes in D induced by Z. In our case we measure D in several ways. One of them involves imputing a D on an almost-continuous scale of Polity IV units (from -10 to 10, see Appendix 4). It is likely that the effect of making a country more democratic is not linear in the Polity scale. For example, the effect of increasing from Polity=6 to Polity=10 could be much greater than the effect of increasing from Polity=0 to Polity=4. If our survey experiment tends to induce changes around the middle of the Polity

[22]The placebo *military spending* was the least imbalanced under the Basic design (consistent with the Realistic Bayesian Model which predicts that to be the least powerful placebo test), though it was imbalanced at $p_{one-sided} < 0.05$. It was perfectly balanced in the Controlled Details design.

scale, but our intended causal estimand is focused on the upper end of the scale (as we think it usually is), then our IV estimator would underestimate the target average treatment effect.

The IV framework clarifies why we should be careful about interpreting the ITT estimate (the estimated effect of Z on Y) as an average treatment effect. The bivariate IV estimator is $\hat{\delta}_{IV} = \frac{Cov(Z,Y)}{Cov(Z,D)}$. The numerator is the ITT estimate. The denominator is the "first stage", the estimated effect of Z on D. If $Cov(Z,D) = 1$, so that every respondent interprets the terms "non-democracy" and "democracy" as is intended by the researchers, then the ITT estimate is the same as the complier average treatment effect (because everyone is a complier). On the other hand, to the extent that the first-stage is weaker than intended (that is than 1), the ITT estimate will be less than the complier average treatment effect.

In our case, this is what we observed. Respondents' beliefs about the regime-type of the country did change in the intended direction in all vignettes: under the democracy vs the non-democracy conditions, respondents assigned a higher likelihood to the country being fully democratic or democratic and a lower likelihood of being non-democratic or fully non-democratic (see Figure 30). However, the baseline levels and the magnitude of the changes were different than what are implied by a literal interpretation of the manipulated text.

Considered in isolation, a literal reading of the phrase "a country that is a democracy and shows every sign that it will remain a democracy" implies that the country is at least "democratic" (Polity score 6-9), if not "fully democratic" (Polity=10; using the categories from one of our treatment measures). But for each regime type, respondents think the target country is less likely to be "democratic" or "fully democratic" than the other categories ("somewhat democratic/somewhat non-democratic", "non-democratic", "fully non-democratic"). Using one method of conversion, under the democracy condition the average respondent's belief is that the country has a Polity score of 3.3, 3.6, 2.3 (under Basic, CD, ENE), so clearly not everyone is fully complying with the intended treatment.

The respondents' beliefs about the country in the non-democracy condition were highly autocratic, possibly more so than is warranted by the literal phrasing of the non-democracy condition. The scenario read that "the country is not a democracy and shows no sign of becoming a democracy." Respondents assigned the highest probability to "Fully Non-democratic" (Polity score -10 to -6), for which our examples included China, Saudi Arabia, Vietnam, North Korea and Iran.

The reason that respondents' perceived the regime type to be more authoritarian than is implied by the literal text seems clear: respondents did not read the sentence about regime-type independent of the other features of the scenario. The fact that this country was developing nuclear weapons, "had refused all requests to stop its nuclear weapons program" and is otherwise portrayed as a threat led respondents to condition their interpretation of the country's regime type. While Russia and Iraq are not the countries one thinks of when reading about "democracy", they appear to be the kinds of countries one thinks about when reading about "democracies" building nuclear weapons in a threatening manner.In general this result speaks to the broader message of this paper that it is rarely possible to simply manipulate a specific feature of a scenario-based survey experiment without also manipulating the respondents' interpretation and understanding of other features of the scenario.

Respondents also perceived the difference in the level of democracy to be much smaller than a literal interpretation of the regime-type portion of the vignettes would suggest. If we interpret the "democracy" phrasing to refer to countries centered in the middle of our "democracy" category (Polity = 8), and the "non-democracy" phrasing to refer to countries centered in the middle of the Polity scale (Polity = 0), then the change in level of democracy should be about 8 points on the Polity scale. We could also adopt a more autocratic interpretation of "non-democracy", centering the interpretation at about a Polity= −3. The effect of our vignettes on perceived level of democracy, then, should be about 8 to 11 points on the Polity scale. We found that the level of democracy increased on average by 5.5, 5.4, and 4.3 Polity points (for Basic, CD, ENE; see Figure 31).

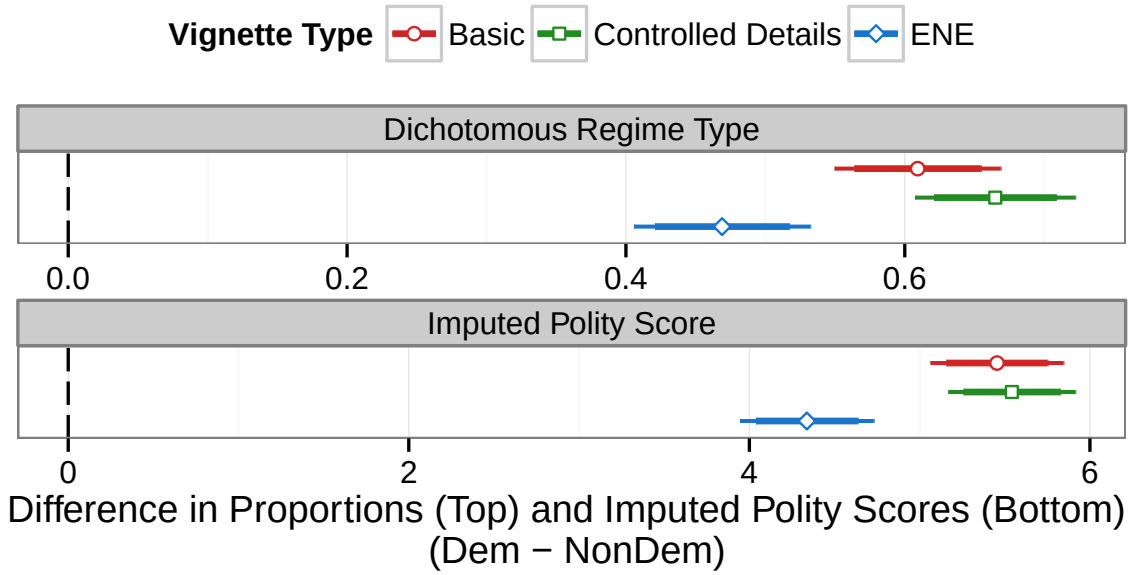
The magnitude of these "first-stage" effects matter for drawing correct substantive interpretations of the results from scenario-based survey experiments. Studies of the democratic peace that interpret the ITT estimates (the effect of Z on Y) as estimates of an average causal effect (of D on Y) are likely producing underestimates (if D is dichotomous, they will definitely be underestimates unless compliance is perfect, so $Z = D$). Suppose researchers have in mind the counterfactual of a target country being a democracy (Polity around 8) vs being a non-democracy (Polity around 0). The ITT estimates are thus about between 60 percent of this intended contrast (Figure 5). For estimating the magnitudes of effects one needs to go beyond ITT estimates to IV estimates, rescaling by the strength of the instrument. Scenario-based survey experiments of the democratic peace reporting only ITT estimates are thus often likely to underestimate the intended quantity of interest because the other details in the scenario will tend to shift the respondent's interpretation of "democracy" and "non-democracy" towards each other, and in general not every respondent will adequately process the manipulation.

4.5 Limits of Controlled Details Designs

While the Controlled Details design that we used did not completely overcome confounding, it did seem to work for those details that were specified. Could we not, then, just specify more characteristics? This logic of inference is typical in observational studies where scholars defend a causal estimate by showing that it is robust to inclusion of a battery of control variables. A researcher could provide a very detailed scenario or a conjoint comparison involving many a long list of characteristics. Another strategy is to specify a real referent in the scenario, but then hypothetically vary one aspect of that referent. For example, a study of the democratic peace could ask a question about Iran, and then manipulate whether it is described as recently democratizing or not.[23]

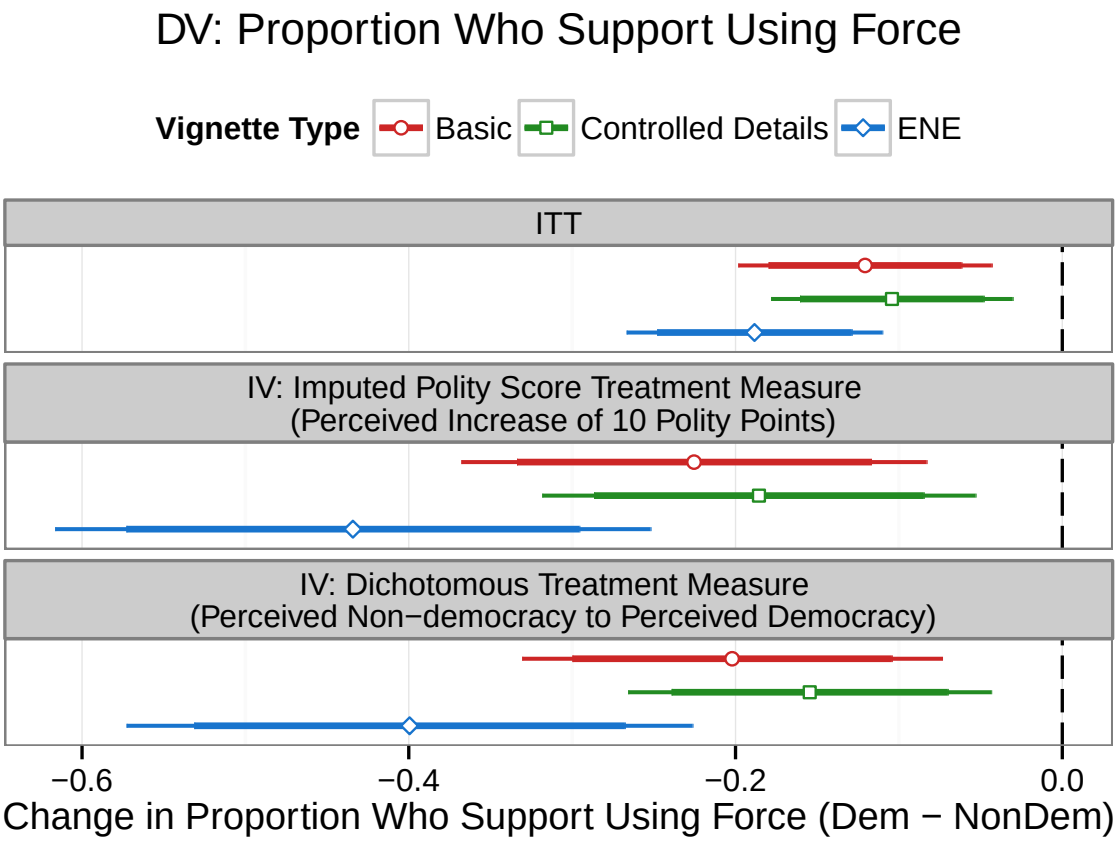
[23]Manipulating the name of the real-world referent alone leads to the same confounding problems found in abstract vignettes. When reading about a scenario involving a specific country, respondents may infer the country's characteristics apart from its regime type. For instance, subjects understand that France and Iran not only have different regime types but also different cultures, histories, and militaries. Therefore, we cannot simply change the name of the aggressor country in the democratic peace survey experiment. Using racialized first names in survey experiments poses a similar challenge since these names also convey information about education level and socioeconomic class; see the next section for further discussion.

Figure 4: Treatment Measure



For the Dichotomous Treatment Measure, we code such 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.
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.

Figure 5: ITT and IV Estimates



The dependent variable is a dichotomous measure for support for using force. Responses "strongly favor" and "favor" are coded 1 and all other responses are coded 0.
The ITT estimate is the average effect of treatment assignment (being assigned to read the target country is a democracy) on their support for using force. The IV estimate is the average effect of respondents perceiving the target country to be a democracy (measured through a dichotomous or Polity score measure) — induced only by treatment assignment — on support for using force.

It may be the case that specifying additional characteristics could shrink confounding to an arbitrarily small mount. Future research needs to explore this possibility. However, there are several reasons why this strategy may not work.

Providing extensive details could attenuate causal effects. If provided in a vignette, the extensive details could exhaust the respondent, leading them to read less closely and satisfice (Hainmueller, Hopkins and Yamamoto, 2015; Krosnick, 1999). Buried in a lengthy vignette, the respondent may also no longer perceive the treatment, though a researcher could always visually emphasize the treatment. There may be deeper problems, though, in asking respondents to hold "all else equal."

Given a large enough number of characteristics, most combinations of characteristics will describe rare or non-existent units. For example, there is simply no empirical referent for a country that has a freely elected head of government, Sharia law for criminal proceedings, and is part of NATO.Cautious researchers could prune away empirically implausible vignettes, as (Hainmueller, Hopkins and Yamamoto, 2014, 20) do. In so doing, however, we confront another problem that besets observational studies: the problem of *rare counterfactuals*. Absent a strong theoretical model from which we can extrapolate, researchers can only estimate the effects of factors for which there is ignorable variation in the real-world. This problem is apparent in how respondents interpreted our "democracy" vignettes as much less democratic than we intended. Democracies developing nuclear weapons in a threatening manner (to the US) are rare or non-existent. We refer to this constraint prohibiting vignettes that are implausible as the *plausibility constraint*.

We believe most scholars are aware of this potential problem, as otherwise why would scholars confine themselves to relatively realistic scenarios? Without some plausibility constraint, we could ask respondents about whatever hypothetical counterfactual we have in mind. What are the actual consequences of posing implausible scenarios?

One issue is of external validity. For scenario-based survey experiments to be useful, the responses to scenarios should approximate how those same people would respond to the real-life analogue of the scenario. If people are better at predicting their opinions under plausible hypotheticals than under implausible hypotheticals, then designs relying on implausible hypotheticals could be biased (and/or noisy). This bias could simply attenuate effects, but it could do worse. For example, suppose a researcher asks a respondent about their willingness to use force against a country that is "**Islamic**, fully democratic, has a free press, gender equality, and an advanced knowledge economy." While religion may matter in practice to a respondent, social desirability and other biases could lead the respondent to suppress that effect. Further, religion may have an effect through other channels, such as how the media talks about the country and the social connections between citizens of the country. While the effects of these various pathways can be approximated for plausible scenarios because the respondent can think about actual countries that fit the counterfactual, they cannot do this for rare counterfactuals.

Another issue is that controlling for characteristics can actually create or amplify confounding bias. For example, suppose that support for force in the US public is greater for target countries that are Islamic. A researcher, unaware of that confound, writes the Basic vignette, except sets in the Middle East to "control for regional confounds" (fixing *Middle East*$= 1$). Then the bias from *Islamic* will become more severe, since the magnitude of the correlation between democracy and *Islamic* is greater in the Middle East ($\rho = -0.67$) then

in the entire world ($\rho = -0.47$). There are many other kinds of situations where controlling for one background characteristic could generate new biases, or amplify the biases from other confounds (Pearl, 2010). And of course controlling for post-D characteristics are even more susceptible to inducing bias. To adjudicate whether controlling for a specific characteristic is a good idea, researchers should use the same criteria as are appropriate for observational studies.[24]

In summary, trying to control for many characteristics could exhaust the respondent, lead to rare counterfactuals for which the respondent is not equipped to give valid answers, and could actually induce or amplify confounding biases if care is not taken in the selection of control details. Researchers using Controlled Details designs should select their controls using similar principles as those appropriate for observational studies. Specifically, controls should be selected that will allow scholars to rule out the most plausible alternative explanations for the result.[25] These are typically, though not necessarily, (1) factors that are thought to affect the outcome and (2) pre-treatment (factors prior in scenario time to the implied change in the causal factor of interest).

4.6 Limits of Embedded Natural Experiments

Embedded Natural Experiments also face limits in their application. While our theory and empirics suggest that, when well constructed, they can overcome all kinds of confounding, they are (1) often hard to construct and (2) they change the causal estimand in ways that might not be desired.

In creating these surveys we brainstormed many possible hypothetical natural experiments. We rejected almost all of them. Some were not plausibly as-if random. Some were too subtle or complicated. Some were distractingly colorful. But most simply had too small of an effect on regime-type (recall that IV bias is larger for weak instruments).

As noted above, respondents did not perceive the "democracy" country as especially democratic (average Polity score for Basic of 3.3, Controlled Details of 3.6, and ENE of 2.3). When reading "democracy", the implied Polity score was closer to that of a semi-democracy or semi-autocracy (like South Sudan or Algeria) than a democracy or full-democracy (such as France or Japan).

We were simply unable to think of a plausible strong natural experiment for which the "democracy" level would be a country like Belgium. Using observational data, strong assumptions are needed to estimate the effect of making Belgium a dictatorship. Similarly, it may not be possible to get at these rare counterfactuals using embedded plausible natural experiments, since they all seem so implausible. Thus, the *plausibility constraint* may bind as much on Embedded Natural Experiments as on Controlled Details designs. Future research should evaluate more the consequences of implausible scenarios.

An additional concern about Embedded Natural Experiment designs is that they only allow us to estimate the local causal effect for the kinds of countries that fit the ENE scenario,

[24] Namely by blocking all "backdoor paths" without opening up new ones, and not controlling for consequences of treatment; Pearl, 2000.

[25] For example, the Controlled Details design in Tomz and Weeks (2013) was thoughtful because military alliance and international trade are prominent rival theories that seek to explain the democratic peace apart from political institutions.

whereas Controlled Details designs seem to allow us to estimate more general causal effects. We believe this issue is often misunderstood. While it is correct that the ENE design only estimates a local causal effect, it is also the case that Controlled Details design only estimate a local causal effect. To see this, consider the set of countries a respondent has in mind after reading the Controlled Details design: countries for which the controlled characteristics are at their fixed values. If controls are selected by the researcher, or interpreted by the respondent, so as to make scenarios plausible, then we will be confined to the kinds of ignorable variation in treatment that exists in the real-world. Causal estimates will be for the kinds of individuals with this ignorable variation; these individuals are often not similar to the broader population. For example, while our Basic design poses the crisis scenario in an abstract way and does not refer to specific countries, most respondents indicated that they were only thinking about a handful of countries (e.g., North Korea, Iran, Iraq, Pakistan) when reading the vignettes. Further, the set of countries that respondents had in mind looked similar across our three designs (see Table 24). Neither the Basic, Controlled Details, or ENE designs allow us to estimate the effect of regime type for countries like Belgium.[26] With any design we should consider what is our effective sample, and theorize whether it is appropriate to generalize our estimated causal effects to other kinds of units. An advantage of ENE designs is that they make the context clear, whereas for more abstract designs much of the context is left to the respondent's imagination.

5 Extensions to Other Studies

We have extended these methods to three other studies: a study of racial discrimination (Desante, 2013), a study about career decisions (Latura, 2015), and a study of the effect of coercive acts on resolve (Dafoe and Hatz, 2014). We will briefly summarize some of the results here.

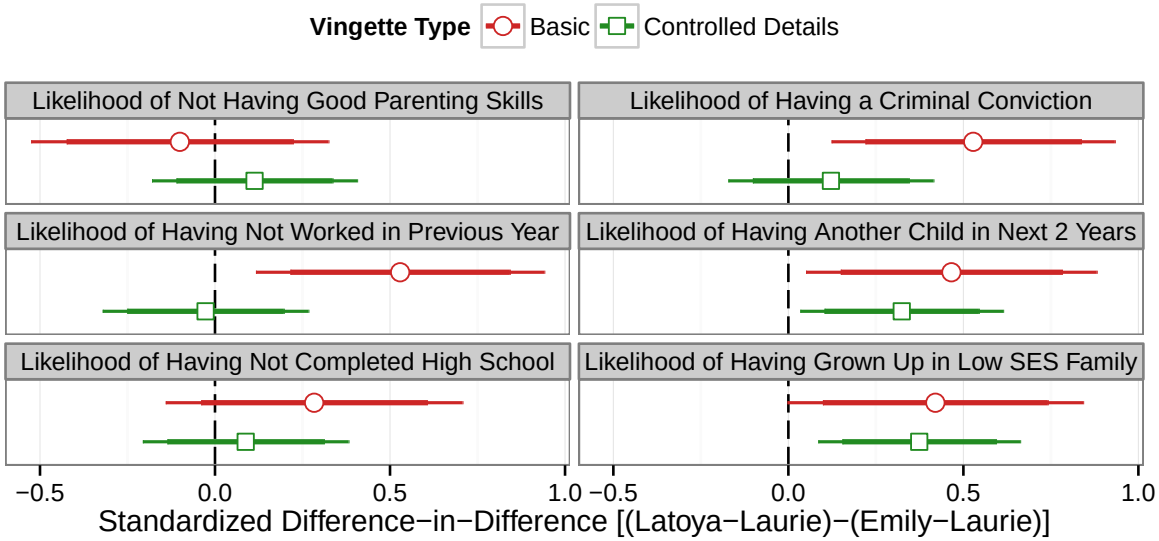
5.1 Why is Latoya Discriminated Against?

Desante (2013) studies whether Americans are more willing to support welfare for people who are white than black, and why this is the case. The manipulation is the name of the welfare applicant (e.g. Laurie vs Latoya). The number and age of the applicant's children are held constant. The experiment also manipulates a "Worker Quality Assessment" as being either "Poor" or "Excellent". In so doing, this design hopes to rule out "principled conservative" reasons for discrimination, leaving only "racial animus" as the basis for discrimination. For placebo questions, we are looking for characteristics that are related to "principled conservative" reasons for discriminating. We use questions used by the North Carolina Work First agency to evaluate welfare applicants. These measure whether the applicant completed high school, worked the previous year, has a criminal record, is from a low socio-economic status background, has good parenting skills, and is likely to have another child. In May

[26] This issue is similar to a point made by Aronow and Samii (2014) that regression estimates local causal effects, with individuals weighted by the conditional variance of treatment. The effective sample is often much smaller than the nominal sample, and can have very different characteristics.

2014, we conducted a survey experiment using respondents recruited from Amazon's Mechanical Turk. We first examined a **Basic** design ($n = 156$) where we didn't control for worker quality. We found (significant) evidence of imbalance on all of the placebos except two (whether they completed high school, and whether they have good parenting skills). In the **Controlled Details** design we used DeSante's control for worker quality assessment ($n = 312$). We then only found (significant) evidence of confounding on low socio-economic status background and on having another child (See Figure 6). This shows that DeSante's control strategy reduces imbalance on some characteristics that a "principled conservative" might discriminate on (prior work experience, criminal conviction), but not on all such characteristics (low SES, intention to have children). Thus, while the results in (Desante, 2013) do provide insight into the reasons for racial discrimination, the results are not able to rule out "principled conservative" reasons for this discrimination.

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



This study raises some thorny issues about experimental manipulations when the causal factor of interest is not well defined, such as is the case with "race".[27] What would an Embedded Natural Experiment design look like if we want to manipulate respondents' perception of someone's race? It's hard to think of a process that as-if randomly assigns race, in large part because race is not a clearly defined phenomenon. It is certainly more than skin-pigment, but if it is thought to include things like education and work-ethic then it becomes inseparable from the other bases for discrimination that are consistent with being "principled conservative". One can reframe the study to be about the effects of "exposure to a racial cue" (Sen and Wasow, 2016), in this case the effects of a welfare applicant having the name Latoya vs Laurie. This response is similar to the general response we mentioned at the beginning of the paper about asking the second kind of question about the effects of

[27]For a detailed discussion of how to conceptualize and study the effects of elements of race, see (Sen and Wasow, 2016).

being described as X. But the problem is also the same: we are no longer directly addressing the question of primary interest, which for Desante (2013) was why Latoya is discriminated against: principled conservative reasons or racial hatred.

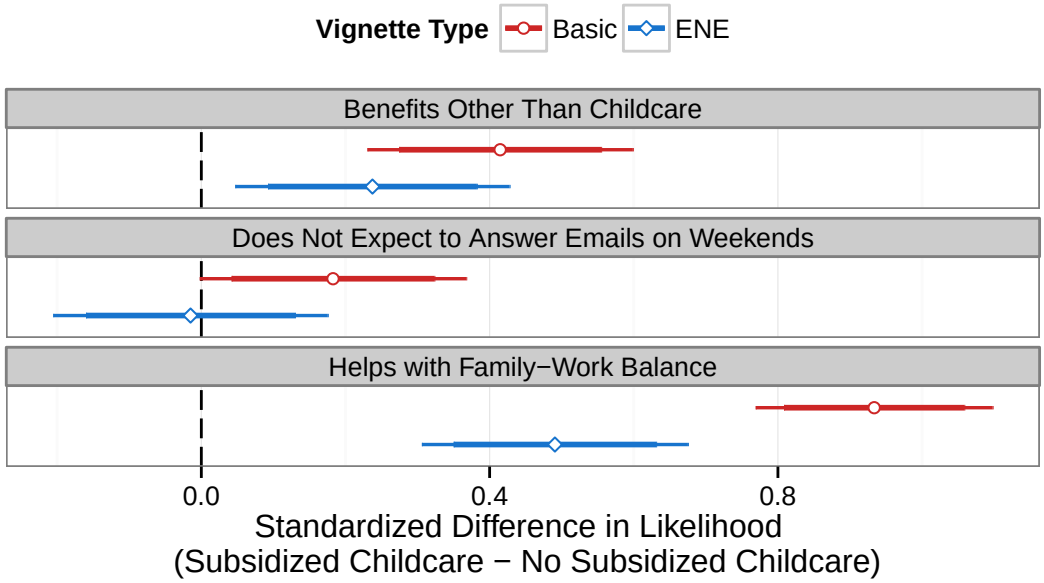
In our thinking about this study we came up with several potential natural experiments for skin-pigment. For example, a person applying for welfare could be described as having a rare mutation making them slightly darker/lighter than their identical twin; we would show pictures of both. But we realized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

5.2 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. We performed an experiment embedded in Latura's survey experiment, which was conducted in April 2015 on Amazon.com's Mechanical Turk. Our experiment 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** ($n = 771$), 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 **ENE Design** ($n = 1003$), 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.

What could a reasonable respondent infer about background characteristics of a company that provides subsidized childcare? They might think that the company in general provides more employee benefits, and that the company is more attentive to family needs. Accordingly, we asked (three) placebo questions to get at these characteristics: (1) Does the company offer other employee benefits than childcare; (2) does the company expect employees to answer work-related email on the weekends; (3) does the company help employees to balance family-work issues. Figure 7 presents the placebo test question results. We see, first, that there is imbalance on all placebos in the basic design, in the direction predicted. Second, that imbalance is reduced in the ENE designs. Third, that the ENE did not eliminate imbalance for two of the characteristics. This suggests either that respondents are not fully Bayesian—incorrectly drawing an inference about the outcome of a lottery—or that they did not fully believe that the lottery was random. Future work should seek to better understand why a "perfect" ENE like this did not completely succeed (perfect because it was based on an allegedly truly random process). Perhaps a narrative that makes the randomness of the process more believable will achieve better balance.

Figure 7: Placebo Test Results from Latura's (2015) Survey Experiment on Professional Decision-Making



5.3 Effects of Coercive Harm

Dafoe and Hatz (2014) study the effect of coercive harm on the resolve of the target population. They create an embedded natural experiment involving US and Chinese planes buzzing each other. In the control condition they nearly collide. In the treatment condition they collide, killing one of the pilots. If this collision is perceived to be as-if random then respondent beliefs about the scenario should be unconfounded. They contrast this ENE with an intentional attack by one state (vs no attack), again shooting down a plane and killing the pilot. This comparison is more likely to be confounded with the intentions and capabilities of the states. Consistent with our claims here, they find that there is less imbalance in perceived intentions and capabilities in the ENE vignette (as-if random collision) than the Basic vignette (attack vs no attack), though their results are preliminary.

6 Recommendations

Survey experiments are extremely valuable tools for social science. But as with any method for causal inference, scientists should be aware of the possible pitfalls of their method. In particular, survey experiments can often be confounded in ways similar to the analogous observational studies. Best practice for survey experiments is thus similar to best practice for observational studies.

1. Theorize confounding. Think about the kinds of characteristics that cause both treatment and the outcome, as well as proxies for these confounding causal pathways.

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

3. Find 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 for confounds. If you can't employ an embedded natural experiment, employ Controlled Details designs to reduce the worst kinds of confounding.

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

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

7. Qualify your inferences. Acknowledge the possibility of confounding biases. Recognize that your estimated causal effects are local to the kinds of scenarios you presented and the respondents' inferences about the context of the scenario.

References

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.

Angrist, Joshua D. and Jorn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, New Jersey: Princeton University Press.

Aronow, Peter M and Cyrus Samii. 2014. Does Regression Produce Representative Estimates of Causal Effects? In *EPSA 2013 Annual General Conference Paper*. Vol. 585.

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

Benton, J Edwin and John L Daly. 1991. "A question order effect in a local government survey." *Public Opinion Quarterly* pp. 640–642.

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

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

Bullock, John G. 2011. "Elite Influence on Public Opinion in an Informed Electorate." *The American Political Science Review* 105(03):496–515.

Dafoe, Allan and Guadalupe Tunón. 2014. "Placebo Tests for Causal Inference." International Studies Association Annual Convention 2014.

Dafoe, Allan and Sophia Hatz. 2014. "Coercion Provocation and Reputation Concerns." Working paper.

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

Druckman, James N, Erik Peterson and Rune Slothuus. 2013. "How Elite Partisan Polarization Affects Public Opinion Formation." *American Political Science Review* 107(01):57–79.

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

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

Gartner, Scott Sigmund. 2008. "The Multiple Effects of Casualties on Public Support for War: An Experimental Approach." *American Political Science Review* 102(1):95–106.

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

Gilens, Martin. 2002. An Anatomy of Survey Based Experiments. In *Navigating Public Opinion: Polls, Policy, and the Future of American Democracy*, ed. Jeff Manza, Fay Lomax Cook and Benjamin J. Page. New York: Oxford University Press pp. 232–250.

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

Grimmer, Justin, Solomon Messing and Sean J Westwood. 2012. "How Words and Money Cultivate a Personal Vote: The Effect of Legislator Credit Claiming on Constituent Credit Allocation." *American Political Science Review* 106(04):703–719.

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

Hainmueller, Jens, Daniel J. Hopkins and Teppei Yamamoto. 2015. "Optimizing Survey Designs in Conjoint Analysis." Working paper.

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.

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

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.

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

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.

Krosnick, Jon A. 1999. "Survey Research." *Annual review of psychology* 50(1):537–567. **URL:** *http://www.annualreviews.org/doi/abs/10.1146/annurev.psych.50.1.537*

Latura, Audrey. 2015. Material and Normative Factors in Women's Professional Advancement: Experimental Evidence from a Childcare Policy Intervention. American Politics Research Workshop, Harvard University.

McFarland, Sam G. 1981. "Effects of question order on survey responses." *Public Opinion Quarterly* 45(2):208–215.

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

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

Morton, Rebecca B and Kenneth C Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab*. Cambridge University Press.

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

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge University Press.

Pearl, Judea. 2010. On a Class of Bias-Amplifying Variables that Endanger Effect Estimates. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, ed. P. Grunwald and P. Spirtes. Corvallis, OR.: AUAI pp. 417–424.

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

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

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

Schwarz, Norbert and Hans-J Hippler. 1995. "Subsequent questions may influence answers to preceding questions in mail surveys." *Public Opinion Quarterly* 59(1):93–97.

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

Sekhon, Jasjeet S. and Rocío Titiunik. 2012. "When Natural Experiments Are Neither Natural Nor Experiments." *American Political Science Review* 106(1):35–57.

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* .

Shadish, William R., Thomas D. Cook and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Hughton Mifflin.

Siegelman, Lee. 1981. "Question-order effects on presidential popularity." *Public Opinion Quarterly* 45(2):199–207.

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

Sovey, Allison J and Donald P Green. 2010. "Instrumental Variables Estimation in Political Science: A Readers' Guide." *American Journal of Political Science* 55(1):188–200.

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

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

Tomz, Michael and Robert P Van Houweling. 2008. "Candidate Positioning and Voter Choice." *American Political Science Review* 102(03):303–318.

Tomz, Michael and Robert P. Van Houweling. 2009. "The Electoral Implication of Candidate Ambiguity." *American Political Science Review* 103(1):83–98.

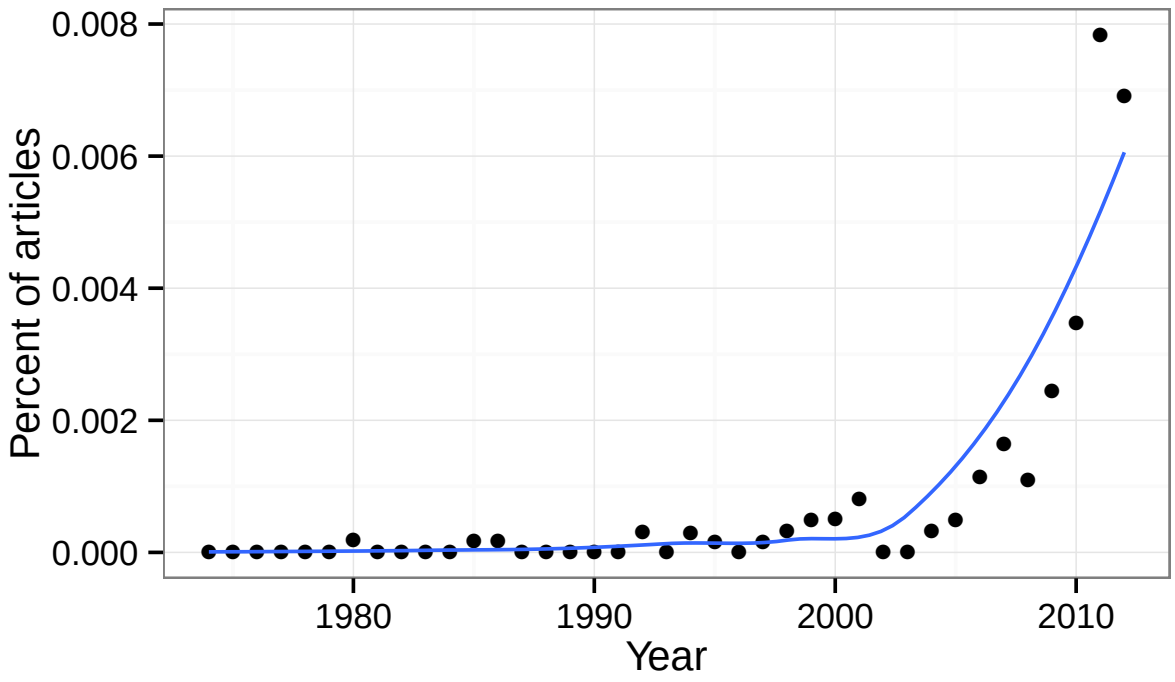
Trager, Robert F. and Lynn Vavreck. 2011. "The Political Costs of Crisis Bargaining: Presidental Rhetoric and the Role of Party." *American Journal of Political Science* 55(3):526–545.

White, Ismail K. 2007. "When Race Matters and When it Doesn't: Racial Group Differences in Response to Racial Cues." *American Political Science Review* 101(02):339–354.

Winship, Christopher and Felix Elwert. 2014. "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable." *Annual Review of Sociology* 40:31–53.

A Literature Review

Figure 8: Increase of Survey Experiments in Political Science



Using JSTOR Data for Research, we searched for the percentage of articles that mention "survey experiment" or "survey experiments" in academic journals within political science for the years between (and including) 1973 and 2013.

Table 2: Survey Experiments Published in Top Journals: Part 1

Title	Authors	Year	Journal	Type
Identity Salience, Identity Acceptance, and Racial Policy Attitudes: American National Identity as a Uniting Force	John E. Transue	2007	AJPS	Framing
Beyond Negativity: The Effects of Incivility on the Electorate	Deborah Jordan Brooks and John G. Geer	2007	AJPS	Framing
Issue Definition, Information Processing, and the Politics of Global Warming	B. Dan Wood and Arnold Vedlitz	2007	AJPS	Framing
How Predictive Appeals Affect Policy Opinions	Jennifer Jerit	2009	AJPS	Framing
Source Cues, Partisan Identities, and Political Value Expression	Paul Goren, Christopher M. Federico and Miki Caul Kittilson	2009	AJPS	Framing
Electoral Incentives and Partisan Conflict in Congress: Evidence from Survey Experiments	Laurel Harbridge and Neil Malhotra	2011	AJPS	Framing
Emotional Substrates of White Racial Attitudes	Antoine J. Banks and Nicholas A. Valentino	2012	AJPS	Framing
Cognitive Biases and the Strength of Political Arguments	Kevin Arceneaux	2012	AJPS	Framing
Polarizing Cues	Stephen P. Nicholson	2012	AJPS	Framing
Stereotype Threat and Race of Interviewer Effects in a Survey on Political Knowledge	Darren W. Davis and Brian D. Silver	2003	AJPS	Other
Designing and Analyzing Randomized Experiments: Application to a Japanese Election Survey Experiment	Yusaku Horiuchi, Kosuke Imai andNaoko Taniguchi	2007	AJPS	Other
Opinion Taking within Friendship Networks	Suzanne L. Parker, Glenn R. Parker and James A. McCann	2008	AJPS	Other
Gender Stereotypes and Vote Choice	Kira Sanbonmatsu	2002	AJPS	Vignette
When Do Welfare Attitudes Become Racialized? The Paradoxical Effects of Education	Christopher M. Federico	2004	AJPS	Vignette
Certainty or Accessibility: Attitude Strength in Candidate Evaluations	David A. M. Peterson	2004	AJPS	Vignette
Racial Resentment and White Opposition to Race-Conscious Programs: Principles or Prejudice?	Stanley Feldman and Leonie Huddy	2005	AJPS	Vignette
The Indirect Effects of Discredited Sterotypes in Judgments of Jewish Leaders	Adam J. Berinsky and Tali Mendelberg	2005	AJPS	Vignette
The "Race Card" Revisited: Assessing Racial Priming in Policy Contests	Gregory A. Huber and John S. Lapinski	2006	AJPS	Vignette
Attributing Blame: The Public's Response to Hurricane Katrina	Neil Malhotra and Alexander G. Kuo	2008	AJPS	Vignette
What Triggers Public Opposition to Immigration? Anxiety, Group Cues, and Immigration Threat	Ted Brader, Nicholas A. Valentino and Elizabeth Suhay	2008	AJPS	Vignette
The Political Costs of Crisis Bargaining: Presidential Rhetoric and the Role of Party	Robert F. Trager and Lynn Vavreck	2011	AJPS	Vignette
Economic Explanations for Opposition to Immigration: Distinguishing between Prevalence and Conditional Impact	Neil Malhotra, Yotam Margalit and Cecilia Hyunjung Mo	2013	AJPS	Vignette
Taking Sides in Other People's Elections: The Polarizing Effect of Foreign Intervention	Daniel Corstange and Nikolay Marinov	2012	AJPS	Vignette

Table 3: Survey Experiments Published in Top Journals: Part 2

Title	Authors	Year	Journal	Type
Social Welfare as Small-Scale Help: Evolutionary Psychology and the Deservingness Heuristic	Michael Bang Petersen	2012	AJPS	Vignette
Working Twice as Hard to Get Half as Far: Race, Work Ethic, and America's Deserving Poor	Christopher D. Desante	2013	AJPS	Vignette
When Race Matters and When It Doesn't: Racial Group Differences in Response to Racial Cues	Ismail K. White	2007	APSR	Framing
Framing Public Opinion in Competitive Democracies	Dennis Chong and James N. Druckman	2008	APSR	Framing
Elite Influence on Public Opinion in an Informed Electorate	John G. Bullock	2011	APSR	Framing
How Elite Partisan Polarization Affects Public Opinion Formation	James N. Druckman, Erik Peterson, and Rune Slothuus	2013	APSR	Framing
Dynamic Public Opinion: Communication Effects Over Time	Dennis Chong and James N. Druckman	2010	APSR	Other
How Words and Money Cultivate a Personal Vote: The Effect of Legislator Credit Claiming on Constituent Credit Allocation	Justin Grimmer, Solomon Messing, and Sean J. Westwood	2012	APSR	Other
Predisposing Factors and Situational Triggers: Exclusionary Reactions to Immigrant Minorities	Paul M. Sniderman, Louk Hagendoorn, and Markus Prior	2004	APSR	Vignette
Challenges to the Impartiality of State Supreme Courts: Legitimacy Theory and "New-Style" Judicial Campaigns	James L. Gibson	2008	APSR	Vignette
Candidate Positioning and Voter Choice	Michael Tomz and Robert P. Van Houweling	2008	APSR	Vignette
The Multiple Effects of Casualities on Public Support for War: An Experimental Approach	Scott Sigmund Gartner	2008	APSR	Vignette
The Electoral Implications of Candidate Ambiguity	Michael Tomz and Robert P. Van Houweling	2009	APSR	Vignette
Attitudes toward Highly Skilled and Low-skilled Immigration: Evidence from a Survey Experiment	Jens Hainmueller and Michael J. Hiscox	2010	APSR	Vignette
Public Opinion and the Democratic Peace	Michael Tomz and Jessica Weeks	2013	APSR	Vignette

Table 4: Survey Experiments Published in Top Journals: Part 3

Title	Authors	Year	Journal	Type
Through a Glass and Darkly: Attitudes Towards International Trade and the Curious Effects of Issue Framing	Michael J. Hiscox	2006	IO	Framing
Sensitivity to Issue Framing on Trade Policy Preferences: Evidence from a Survey Experiment	Martin Ardanaz, M. Victoria Murillo, and Pablo M. Pinto	2013	IO	Framing
Explaining Mass Support for Agricultural Protectionism: Evidence from a Survey Experiment During the Global Recession	Megumi Naoi and Ikuo Kume	2011	IO	Other
False Commitments: Local Mispresentation and the International Norms Against Female Genital Mutilation and Early Marraige	Karisa Cloward	2014	IO	Other
Domestic Audience Costs in International Relations: An Experimental Approach	Michael Tomz	2007	IO	Vignette
International Law and Public Attitudes Toward Torture: An Experimental Study	Geoffrey P.R. Wallace	2013	IO	Vignette
Promises or Policies? An Experimental Analysis of International Agreements and Audience Reactions	Stephen Chaudoin	2014	IO	Vignette
Decision Maker Preferences for International Legal Cooperation	Emilie M. Hafner-Burton, Brad L. LeVeck, David G. Victor and James H. Fowler	2014	IO	Vignette

B "Democratic Peace" Survey Experiment Details

B.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 orders of the section are 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.

B.2 Three Vignette Types

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

B.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.

B.2.2 Controlled Details

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 [**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.

B.2.3 Embedded Natural Experiment

Embedded Natural Experiment d (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 n (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.

B.3 Support for Force and Mediation Questions Order

We randomized the order of these following questions:

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

B.4 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.

B.4.1 Justifications for Placebo Test Questions

We selected placebo test variables by identifying real-world variables that show large and significant imbalances across regime types (see Table 5).[28] For our analysis, we used data from the Quality of Government (GOG) Basic dataset[29], the Correlates of War (COW) formal alliance dataset[30], the COW trade dataset[31], the COW National Material Capabilities dataset[32], the CIA World Factbook Ethnic Group dataset[33], Vito D'Orazio's Joint Military Exercise dataset[34], and U.S. Department of Commerce Bureau of Economic Analysis's Foreign Direct Investment dataset[35].

First, we showed that geographic regions should be included as a placebo question because democracies and non-democracies are distributed differently across regions. Figure 9 displays the percent of countries that are democracies in the ten regions of the world. We defined democracy using the variable `chga_demo` from QOG, which is a binary coding of democracy/non-democracy from the Cheibub et al. 2010 dataset.[36]

In the analysis of our survey experiment, we focused on four regions that exhibit the most imbalance between regime types: Western Europe, North America, Sub-Saharan Africa, and North Africa and the Middle East. The first two have the largest percentage of countries that are democracies and the last two have the smallest percentage of countries that are democracies.[37]

To select the rest of the placebo test variables, we analyzed 114 characteristics of countries in 1998 from all the datasets mentioned in the introduction. We tried to identify variables

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

[29] Dahlberg, Stefan, Sören Holmberg, Bo Rothstein, Felix Hartmann & Richard Svensson. 2015. The Quality of Government Basic Dataset, version Jan15. University of Gothenburg: The Quality of Government Institute, http://www.qog.pol.gu.se.

[30] Gibler, Douglas M. 2009. International military alliances, 1648-2008. CQ Press.

[31] Barbieri, Katherine and Omar Keshk. 2012. Correlates of War Project Trade Data Set Codebook, Version 3.0. Online: http://correlatesofwar.org.

[32] Singer, J. David. "Reconstructing the Correlates of War Dataset on Material Capabilities of States, 1816-1985." International Interactions 14: 115-32. Correlates of War Project National Material Capabilities Codebook, Version 4.0. http://www.correlatesofwar.org/data-sets/national-material-capabilities

[33] Ethnic Groups Dataset, CIA World Factbook, 2000. https://www.cia.gov/Library/publications/the-world-factbook/fields/2075.html

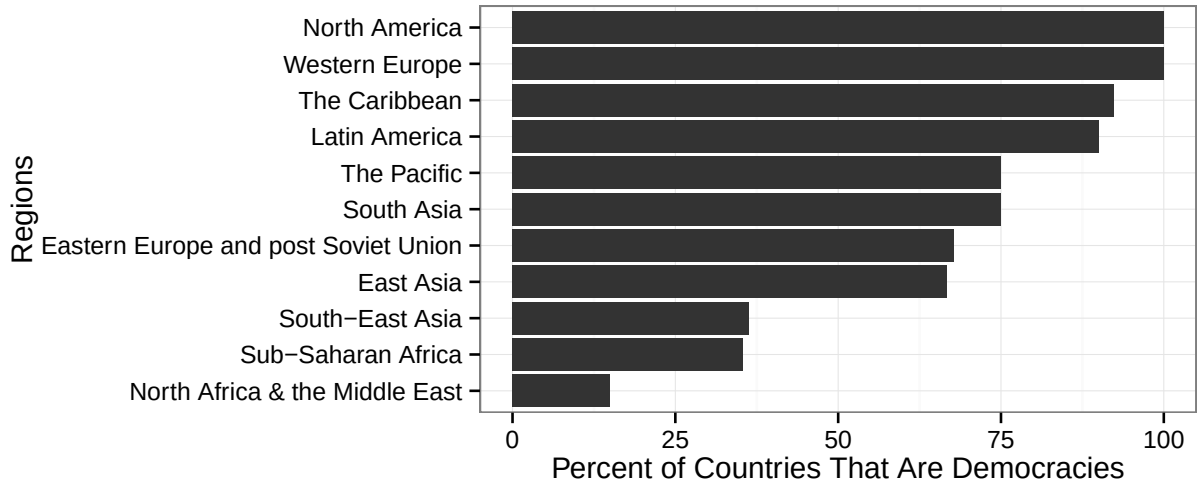
[34] http://vitodorazio.weebly.com/data.html

[35] Foreign Direct Investment in the U.S.: Balance of Payments and Direct Investment Position Data. 2015. U.S. Department of Commerce Bureau of Economic Analysis.

[36] Cheibub, JosÃ© Antonio, Jennifer Gandhi, and James Raymond Vreeland. "Democracy and dictatorship revisited." *Public Choice* 143.1-2 (2010): 67-101.

[37] Although we include East Asia and Central Asia among our answer choices in the placebo test question because those were popular answers in our pilot studies.

Figure 9: Democracies in Regions of the World



that are the most imbalanced across regime types.[38] We selected data from 1998 so that our potential placebo variables are lagged behind the democracy variable by 10 years (the most recent year of the democracy variable `chga_demo`, which we use, is 2008).[39] Furthermore, we selected these variables because they describe characteristics that are not directly related to politics, regime type, or electoral procedure, and are thus more conceptually distinct. For each of these potential placebo variables P_k for $k \in \{1, 2, ..., 114\}$, we standardized them to create $S_{i,k}$ such that for country i:

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

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

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

We can interpret $\widehat{\gamma}_k$ as the estimated difference in means for standardized variable $S_{i,k}$ between democracies and non-democracies. Table 5 presents the coefficient estimates and robust standard errors for the 25 variables that exhibit the greatest imbalance (in absolute value) across regime types.[40] From this list, we identified four potential placebo variables to

[38]The CIA World Factbook Ethnic Group dataset contains too many ethnic groups. Instead, we code the variable `majority_white` using the dataset. For each country, `majority_white` is coded 1 if the country's population is greater than 50 percent white (Causasian) and 0 otherwise. Note the data is from 2000 and not 1998; however, we think whether a country was majority white is unlikely to have changed between 1998 and 2000.

[39]Likewise, in our placebo test questions, we asked subjects to guess what the country in the scenario was like a decade ago so that their answers to the placebo questions are not affected by their beliefs about any recent change in the country's regime type, such as could be induced by the manipulation of the vignette.

[40]We also report the percentage of countries that are missing from each of the variables in the datasets.

use, in addition to the ones related to military capability, alliance, and trade (i.e., variables controlled for in the Tomz and Weeks's vignettes).

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

Second, we created a placebo variable measuring how likely it is that the country in the scenario was majority Christian. As Table 5 shows, democracies had a low percentage of Muslims and a high percentage of Catholics in 1980. Since religion is slowly changing, we regarded it as an especially valid placebo variable (it is unlikely to be affected by regime-type on a short time scale).

Third, we created a placebo variable measuring GDP per capita. Many of the highly imbalanced variables in Table 5 are related to levels of economic development. These variables include employment in agriculture as a percentage of total employment, employment in services as a percentage of total employment, gross enrollment ratio in pre-primary schools, health expenditure as percent of GDP, and mortality rate of children under five. In selecting a placebo question we had several considerations to balance: we wanted to only ask one question to avoid burdening the respondent with multiple redundant questions; we wanted to choose a question that captures much of the common variance to these characteristics; we wanted to ask about a factor that is most likely to influence the outcome (support for using force); we wanted to ask a question that is easy to understand. These considerations lead us to ask about GDP per capita. GDP per capita, itself, is 0.4 standard deviations greater for democracies than non-democracies in 1998 ($p < 0.001$).

Finally, we asked about the racial makeup of the country's population. As Table 5 shows, democracies are more likely to be majority white compared with non-democracies ($p < 0.001$).

We also examined variables that are related to military capability, alliance, and trade, three variables that were included as details in the Tomz and Weeks's survey experiment design. Potential placebo variables include those that were explicitly controlled for by the Tomz and Weeks's vignettes (i.e., non-nuclear military capability, military treaties, and volume of import/export) and those that are highly correlated with alliance and trade (i.e., iron and steel production, energy consumption, population, joint military exercises, and foreign direct investment). We estimated γ_k for these variables using the same model as described in the previous section. In Table 6, we report our coefficient estimates and robust standard errors.[41] We found that none of the variables that describe military capability is statistically significant at $\alpha = 0.05$. On the other hand, variables related to trade and military alliance are all statistically different between regime types at $\alpha = 0.05$.

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

[41] Again, we report the percentage of countries that are missing in each variable.

Table 5: Top 25 Variables Most Imbalanced Across Regime Types

Variables (Standardized)	Coef	SE	% Missing
Fuel exports (% of merchandise exports)	−0.959	0.239	37
Muslims as percentage of population in 1980	−0.953	0.152	11
Employment in agriculture (% of total employment)	−0.941	0.335	56
Population ages 65 and above (% of total)	0.922	0.121	11
Heritage Foundation Economic Freedom Index: Property Rights	0.912	0.147	21
Heritage Foundation Economic Freedom Index	0.877	0.158	21
Employment in services (% of total employment)	0.859	0.295	56
Number of Military Treaties	0.822	0.109	0
Number of Treaties: Defense	0.810	0.109	0
Gross enrollment ratio, pre-primary schools, total.	0.807	0.182	43
Number of Treaties: Entente	0.784	0.110	0
Population ages 0-14 (% of total)	−0.774	0.135	11
Number of Treaties: Non-aggression	0.758	0.111	0
Catholics as percentage of population in 1980	0.740	0.129	11
Social Globalization Index	0.732	0.142	11
Hertiage Foundation Economic Freedom Index: Trade Freedom	0.723	0.154	21
Alternative and nuclear energy (% of total energy use)	0.698	0.149	35
Energy imports, net (% of energy use)	0.686	0.199	35
Country's population was majority white	0.675	0.120	0
Services, etc., value added (% of GDP)	0.675	0.147	16
Health expenditure, total (% of GDP)	0.666	0.135	10
Employment in industry (% of total employment)	0.655	0.336	56
Mortality rate, under-5 (per 1,000 live births)	−0.654	0.147	8
Average Value of Ethnolinguistic Fractionalization	−0.650	0.199	47
Armed forces personnel (% of total labor force)	−0.624	0.168	17

Table 6: Variables Related to Military Alliance and Trade with US

Variables (Standardized)	Coef	SE	% Missing
Iron and Steel Production (Thousands of Tons)	0.137	0.159	8
Military Expenditures (Thousands of $)	0.105	0.155	8
Military Personnel (Thousands)	−0.172	0.173	8
Energy Consumption (Thousands of Coal-Ton Equivalents)	0.108	0.155	8
Total Population (Thousands)	−0.057	0.166	8
Urban Population (Thousands)	−0.089	0.181	8
Composite Index of National Capability Score	−0.010	0.172	8
Number of Treaties: Defense	0.810	0.109	0
Number of Treaties: Non-aggression	0.758	0.111	0
Number of Treaties: Entente	0.784	0.110	0
Number of Military Treaties (All Types)	0.822	0.109	0
Volume of Imports	0.259	0.129	8
Volume of Exports	0.323	0.122	8
Total Volume of Trade (Imports + Exports)	0.291	0.125	8
Number of Joint Military Exercises	0.398	0.119	0
FDI: Position on a Historical-Cost Basis	0.325	0.138	46
FDI: Net Financial Transactions	0.403	0.139	44
FDI: Net Income	0.386	0.128	41

B.5 Placebo Test Questions

B.5.1 Notes on Placebo Test Questions

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

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

Note that the instructions asked country in the past. This is because we wanted to minimize the risk that subjects will 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 are 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.

B.5.2 Text of Placebo Test Questions

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

[Textbox]

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

[Textbox]

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

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

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

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

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

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

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

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

G How likely do you 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?[42]

- 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)

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

- 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.?[43]

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

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

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

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

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

B.6 Treatment Measures

We used two questions to measure how much the democracy condition affects subjects' beliefs about the target country. We called these questions *treatment measures* because they measure 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.[44]

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

[44] Note: Each respondent will input his or her answers using one of the three following matrices randomly assigned to him or her.

	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)	○	○	○	○	○

	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.

B.7 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.

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

Open-ended Mediation Question

Why did you select that answer choice in the previous question?[45]

[Textbox]

B.8 Mediation Questions

The mediation questions was used to investigate the reasons why subjects support or oppose use of force against the target country. We asked the same questions Tomz and Weeks asked in their survey.

[45] This question will not be asked in the Separated-Placebos Design.

B.8.1 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.

B.8.2 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.

B.8.3 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.

B.9 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 do not think the previous questions affect how subjects answer them.

B.9.1 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

B.9.2 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

B.9.3 Age

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

B.9.4 Sex

What is your sex?

- Female
- Male
- Other

B.9.5 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]

C "Democratic Peace" Survey Respondents

C.1 Overview

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

Table 7: Number of Respondents by Experimental Condition

Treatment Assignment	Vignette Type	N
Non-democracy	Basic	513
Democracy	Basic	517
Non-democracy	Controlled Details	512
Democracy	Controlled Details	513
Non-democracy	ENE	516
Democracy	ENE	509

Table 8: 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
Controlled Details	$F(5,942) = 0.69$	0.632
ENE	$F(5,964) = 1.12$	0.349

Figure 10: Demographic Variables by Experimental Condition

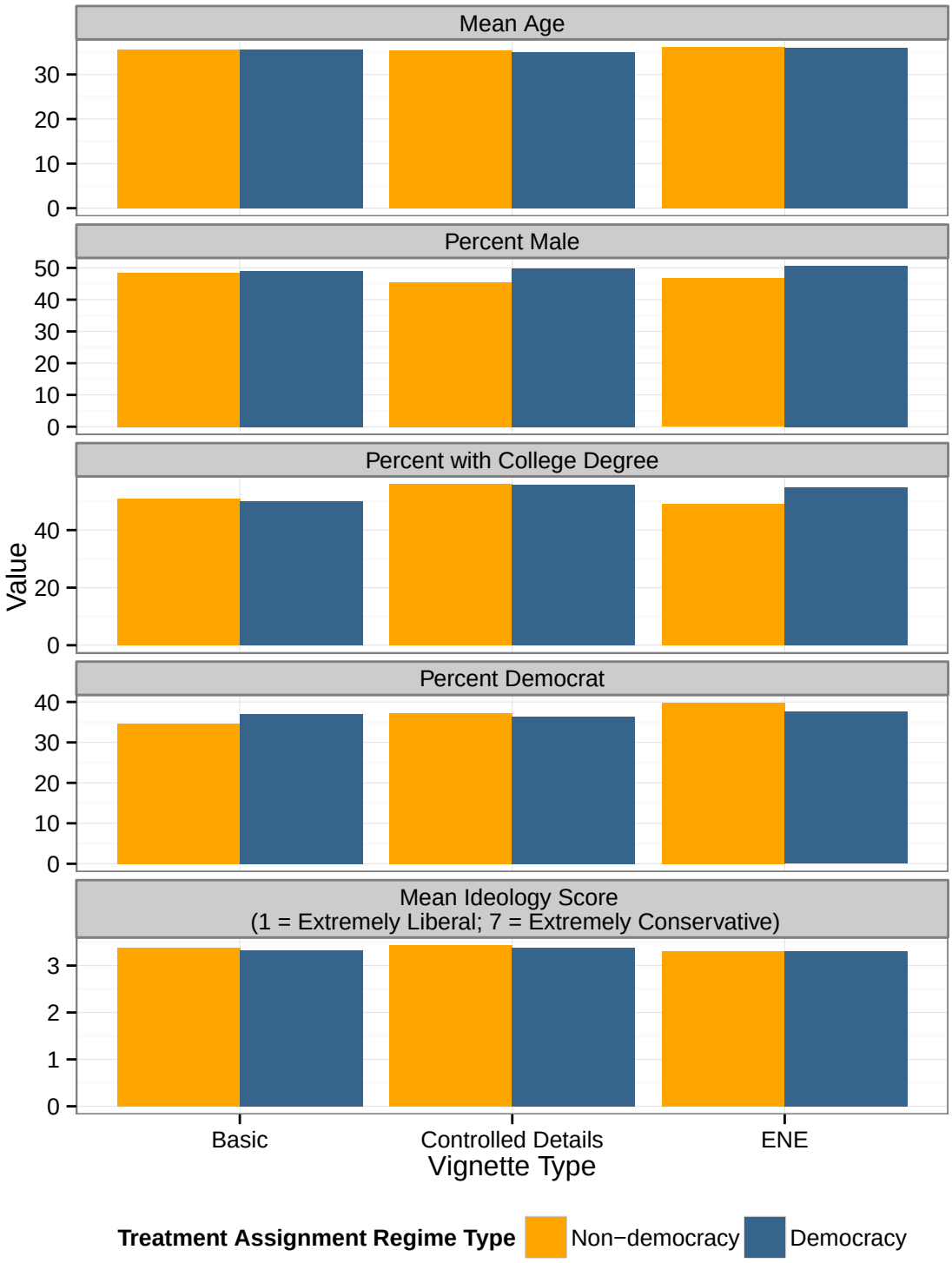
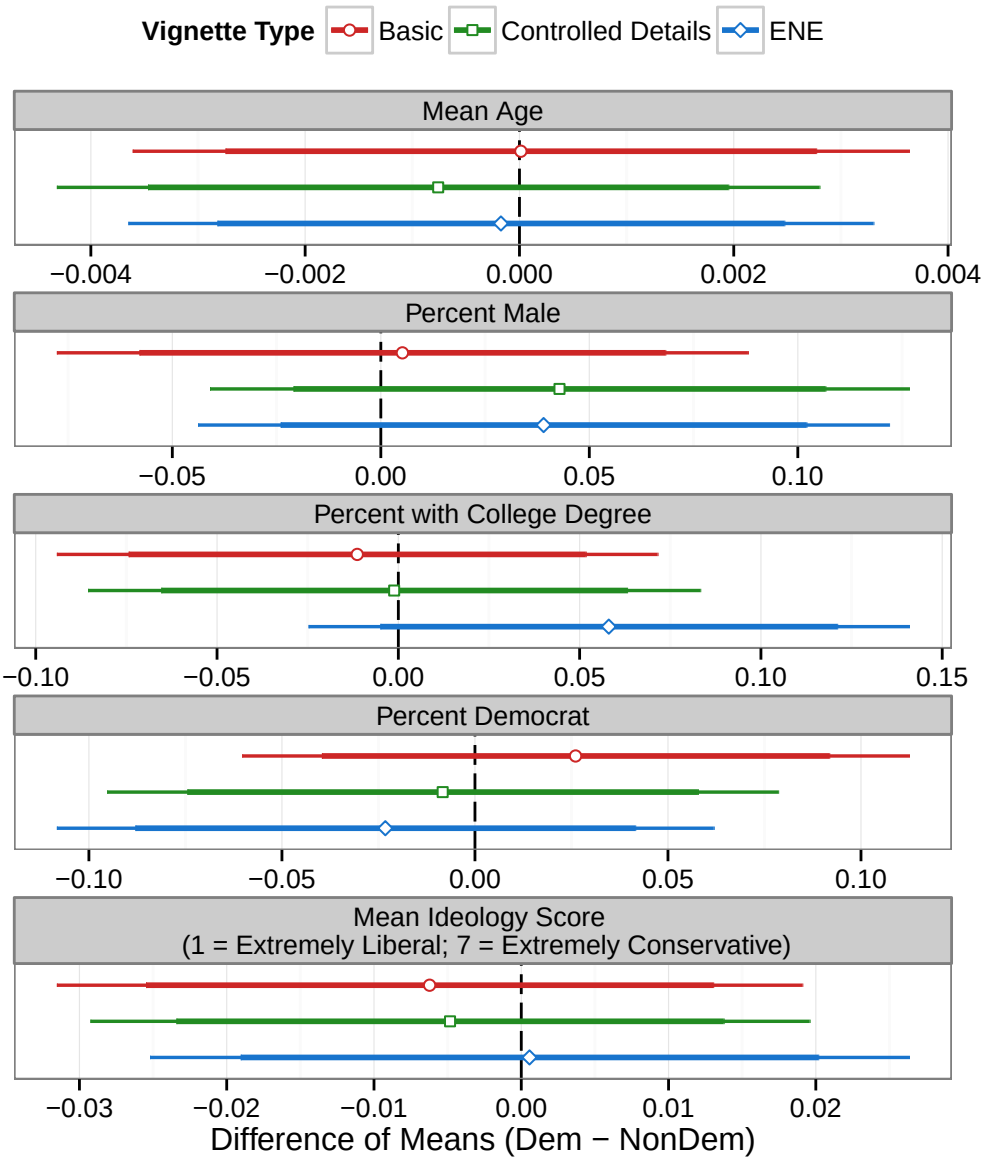


Figure 11: Balance Tests on Demographic Variables



C.2 Balance Tests

Figure 12: Demographic Variables by Experimental Condition

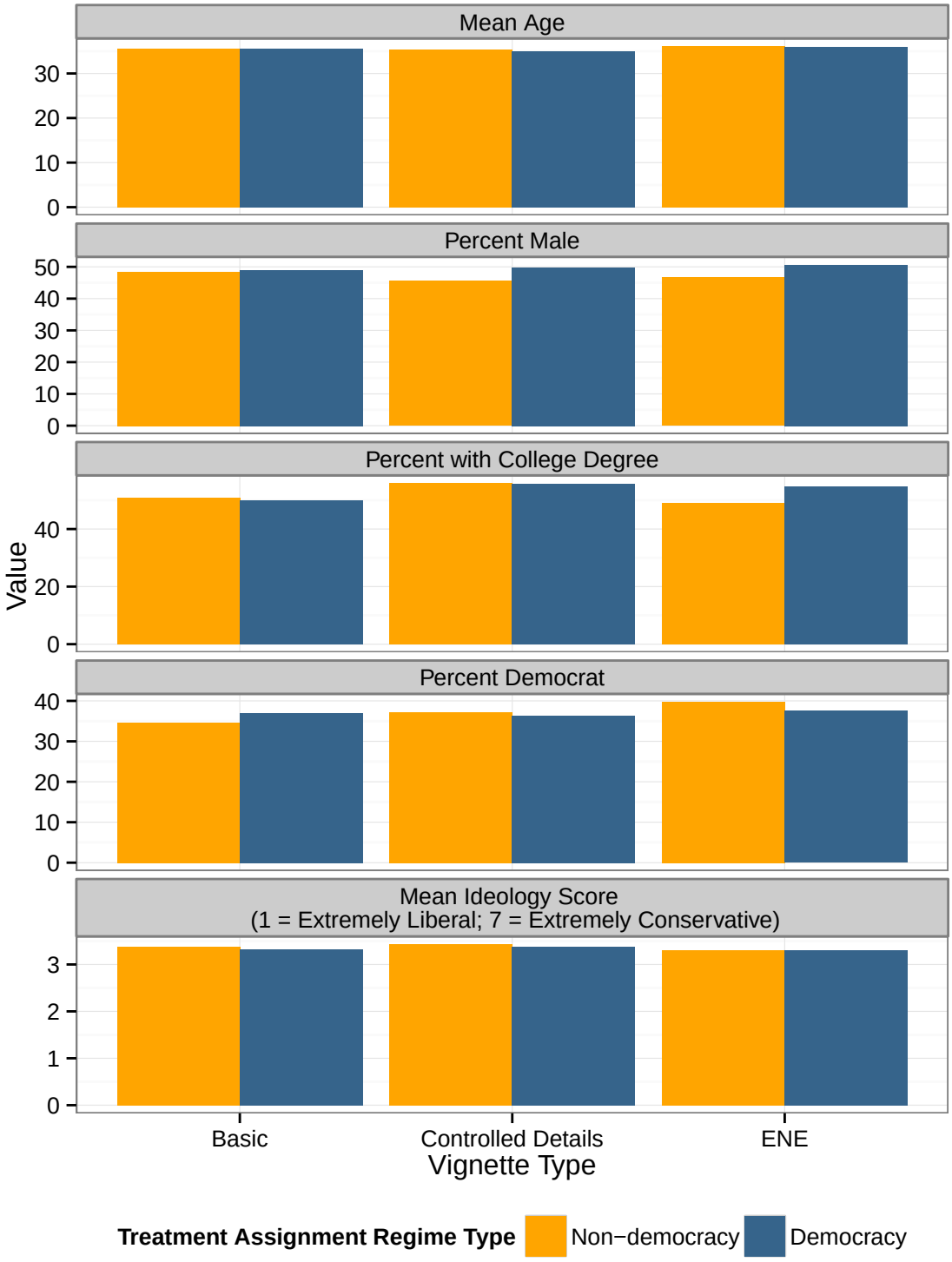


Figure 13: Balance Tests on Demographic Variables

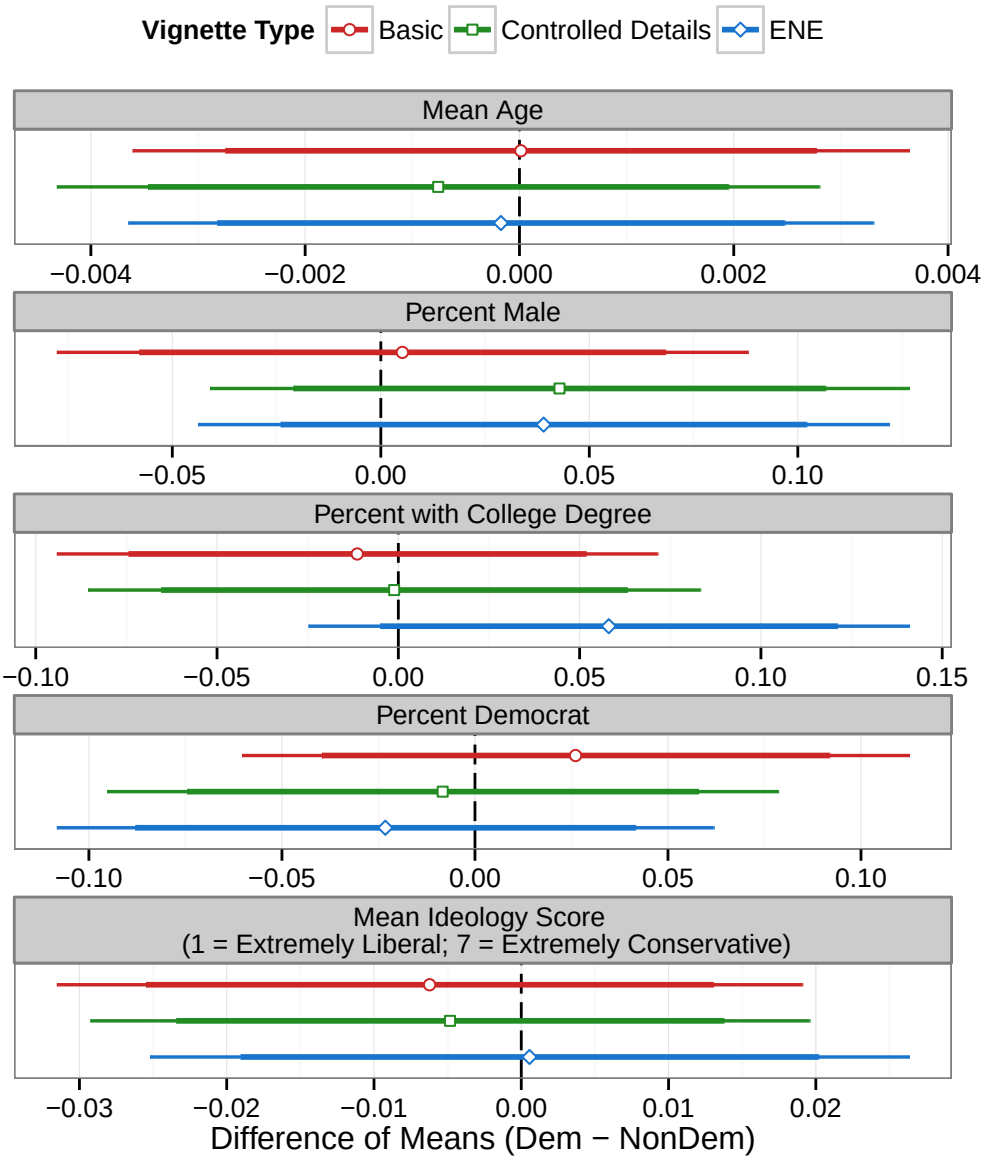


Table 9: Balance Test: Results from Joint F-test Using All Five Demographics Variables to Predict Treatment Assignment

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

D Full Summary of "Democratic Peace" Survey Results

D.1 Coding Placebo Test Results

A: Regions of the World

We reduce the regions to a single dimension $Y_{i,A}^N$ such that $Y_{i,A}^N = 1$ if the subject i mentions North America or Western Europe, $Y_{i,A}^N = 0$ if he mentions Central Asia or East Asia, and $Y_{i,A}^N = -1$ if he mentions the Middle East and North Africa or Sub-Saharan Africa.

B: GDP per Capita

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

C: Religion

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

D: Oil Reserves

We define Y_D^N as one minus the subjects' response to the oil reserves placebo test question: $Y_{i,D}^N$ equals one minus the mean of the probability interval subject i selects.[46]

E: Race

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

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

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

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 he 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 he selects "A Very Large Amount." '

D.2 Placebo Test Results

Figure 14: A: Most Likely Regions

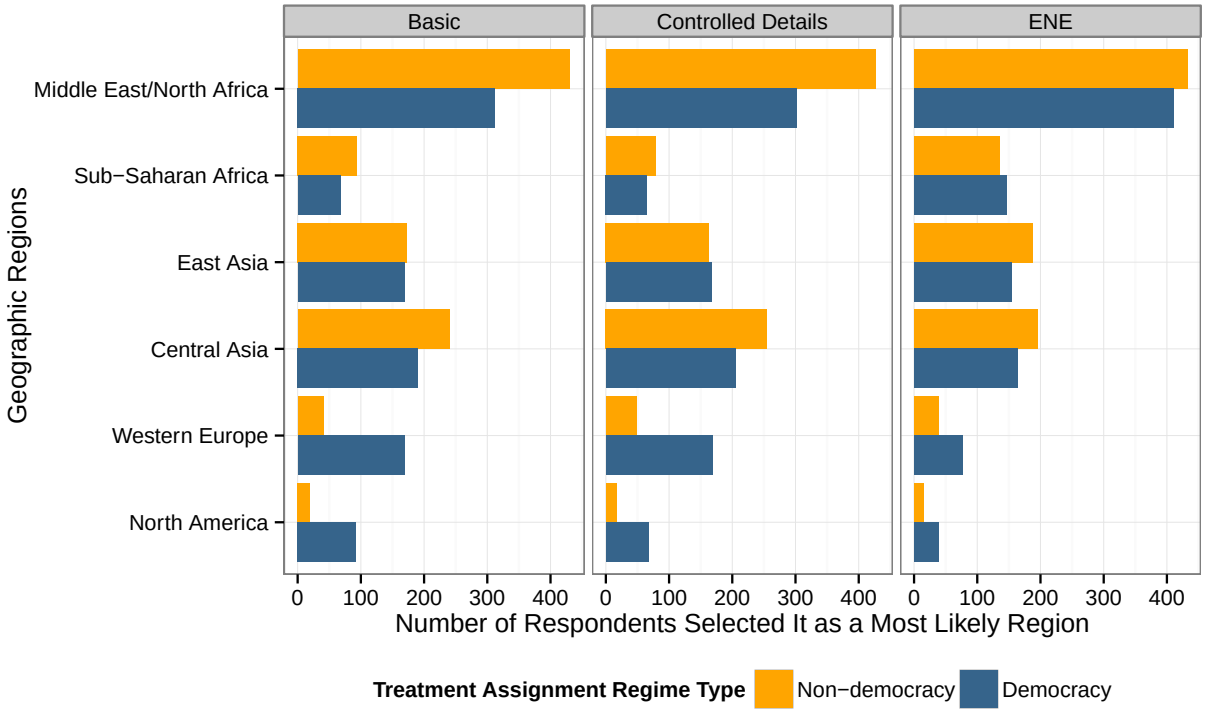


Figure 15: B: GDP per Capita

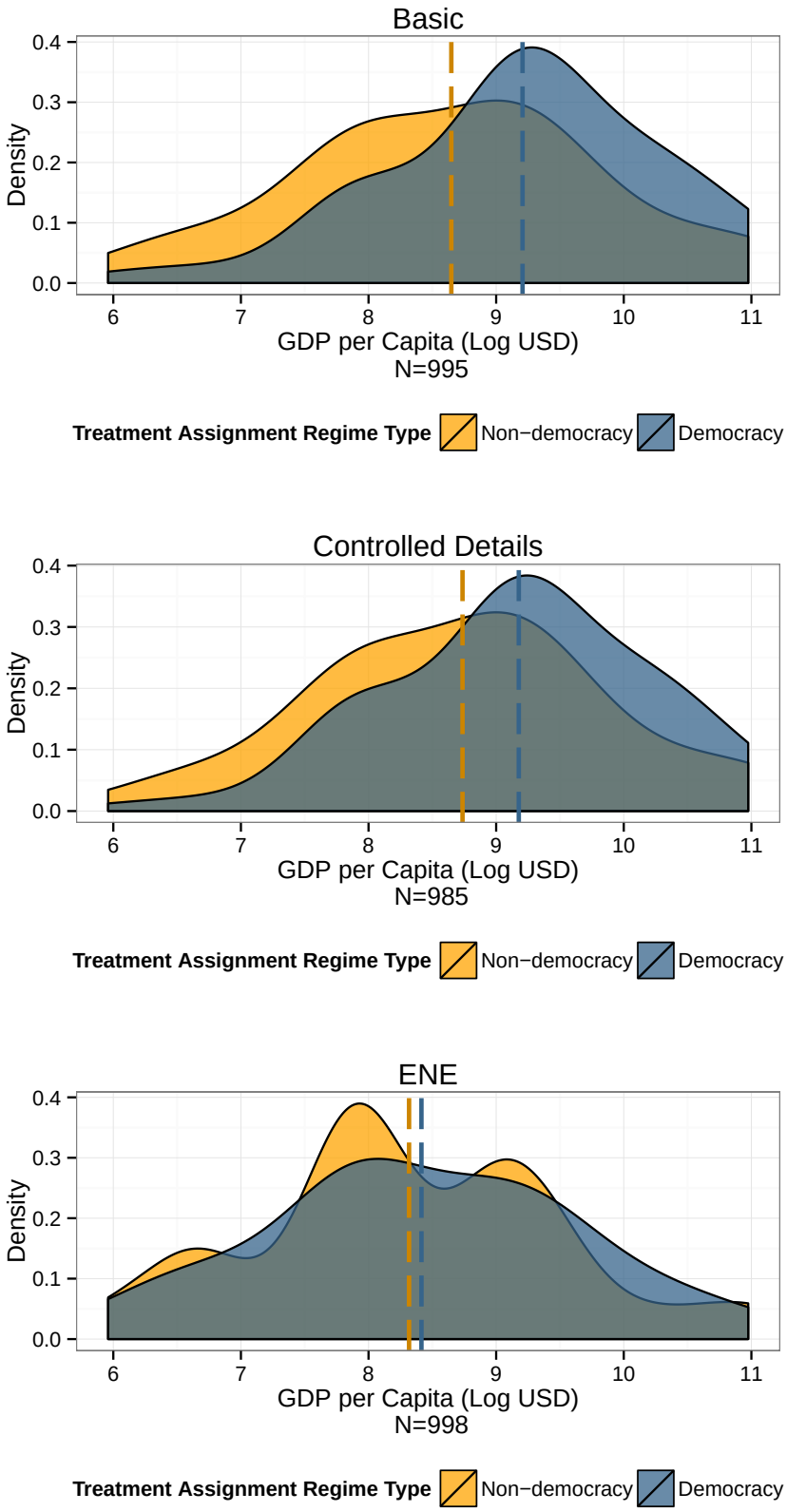


Figure 16: C: Likelihood of Being Majority Christian

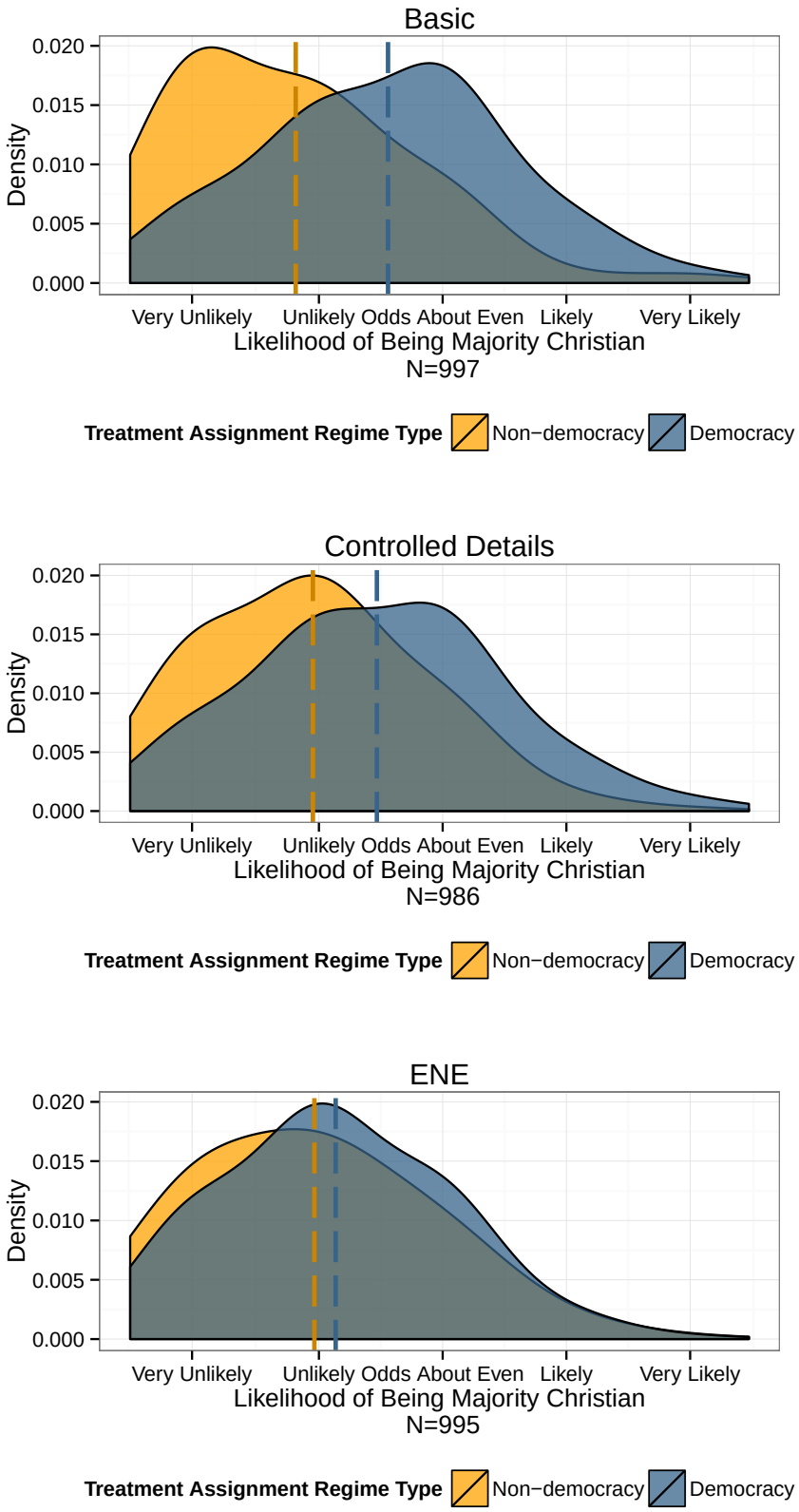


Figure 17: D: Likelihood of Not Having Large Oil Reserves

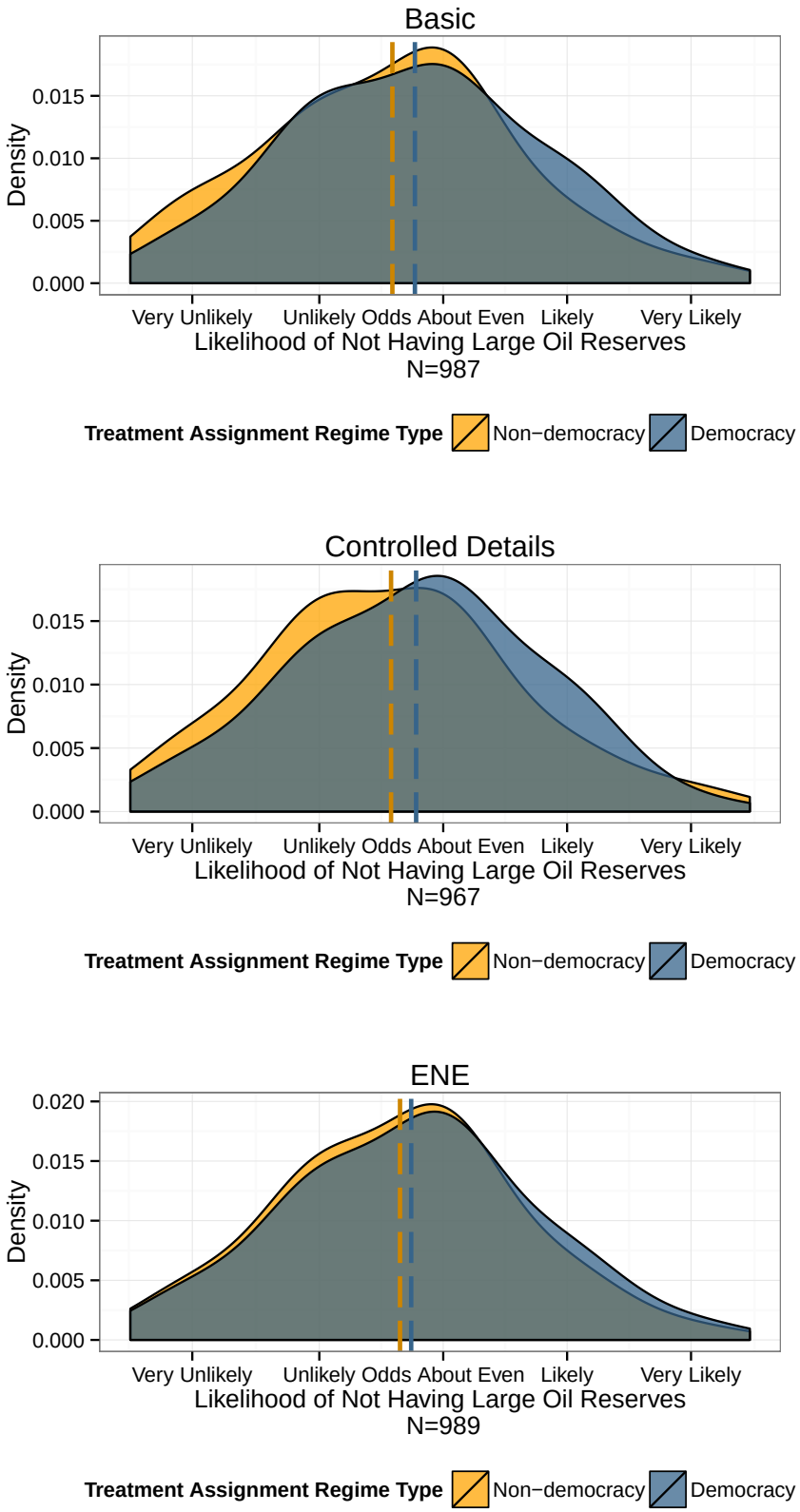


Figure 18: E: Likelihood of Being Majority White

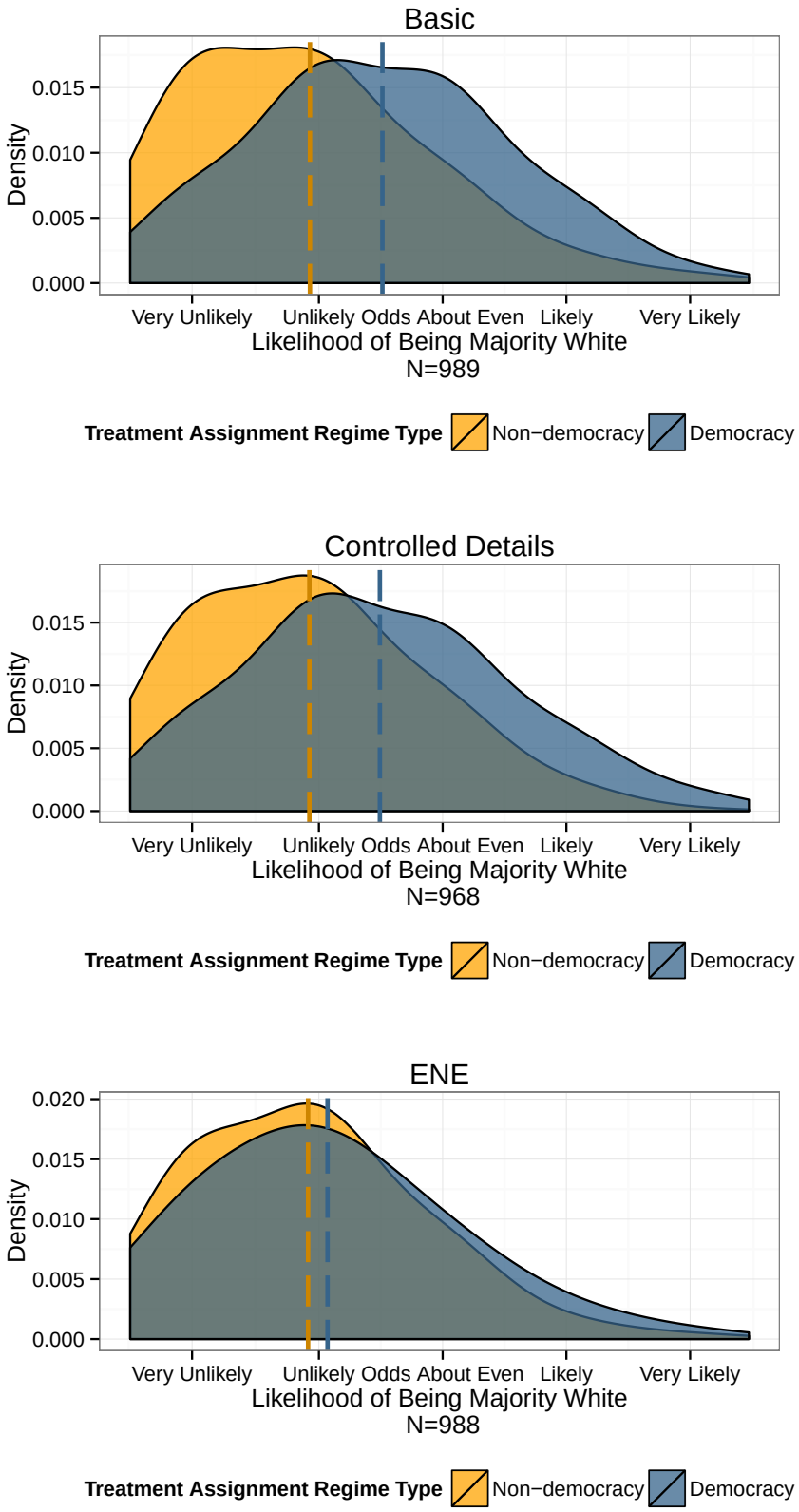


Figure 19: F: Likelihood of Military Alliance with the U.S. since World War II

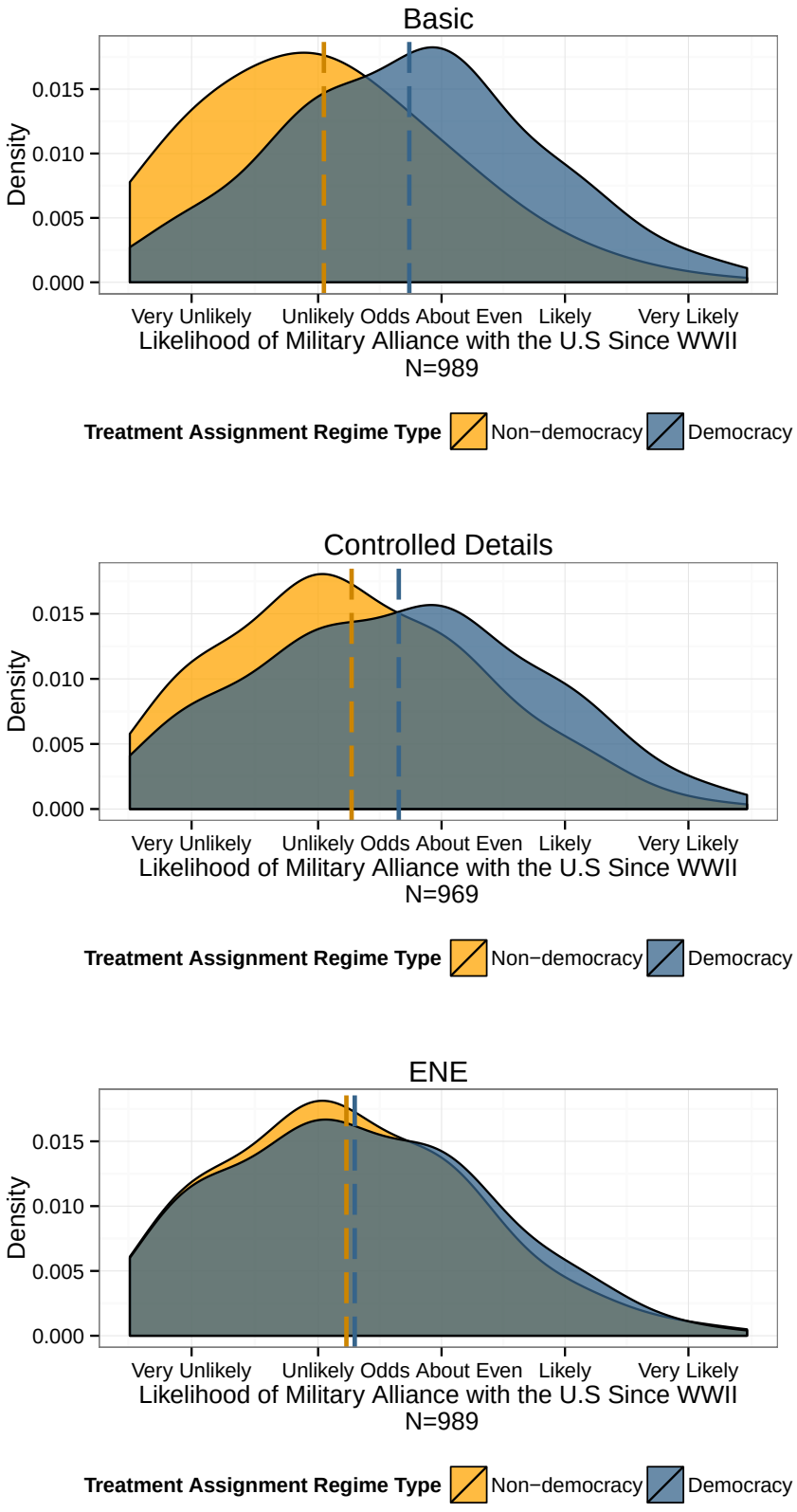


Figure 20: G: Level of Trade with the U.S.

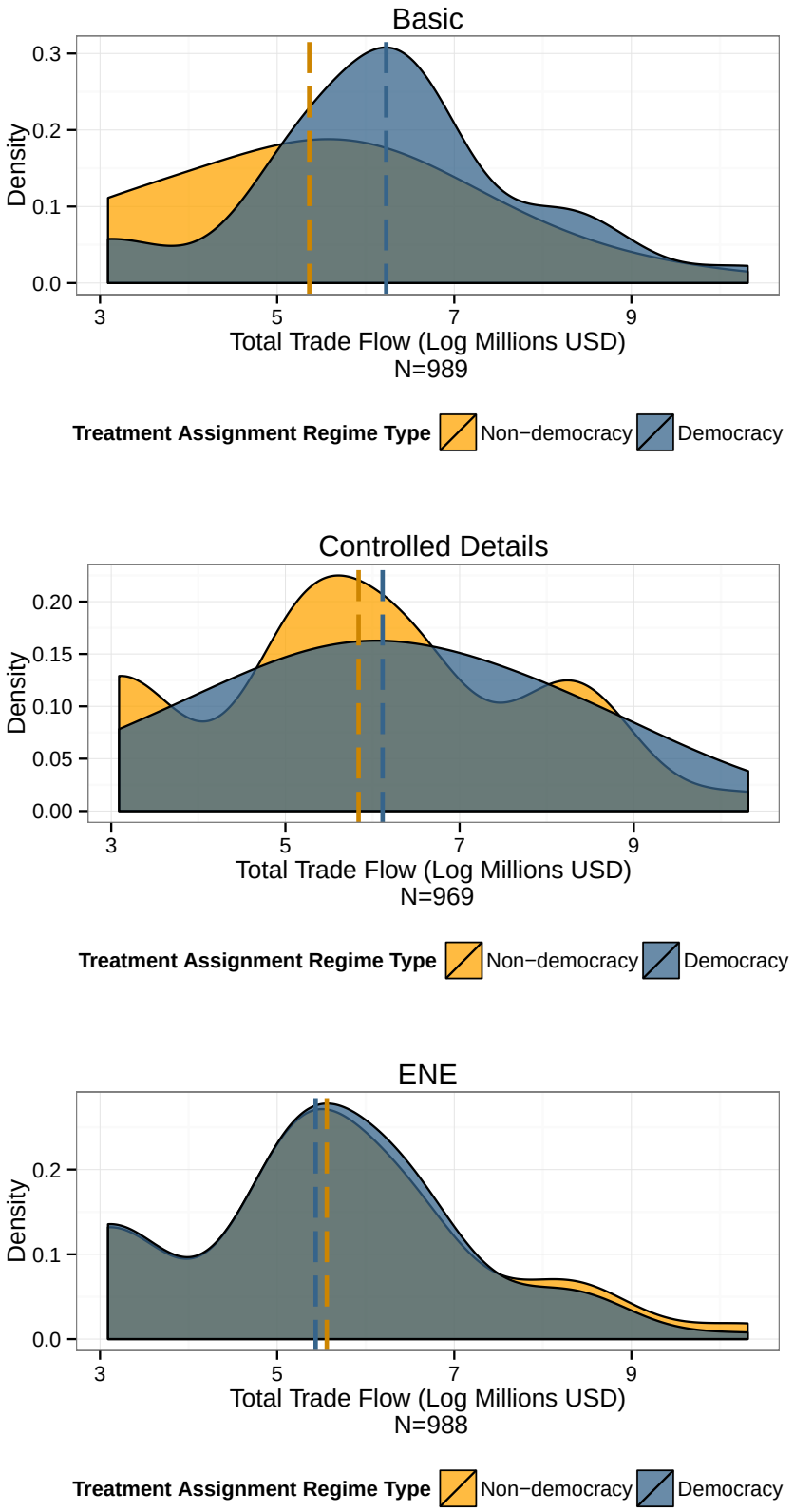


Figure 21: H: Likelihood of Joint Military Exercise with the U.S.

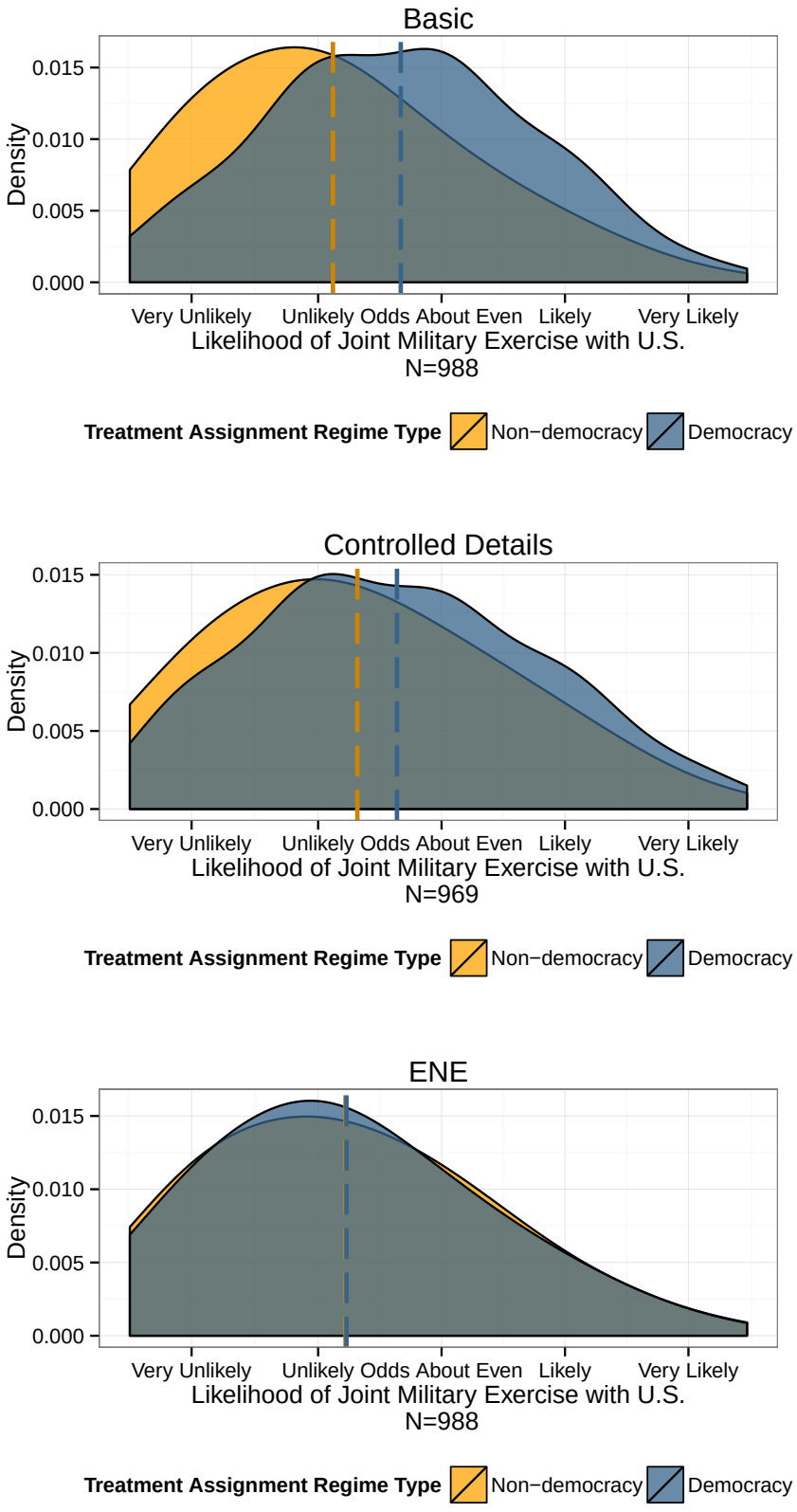
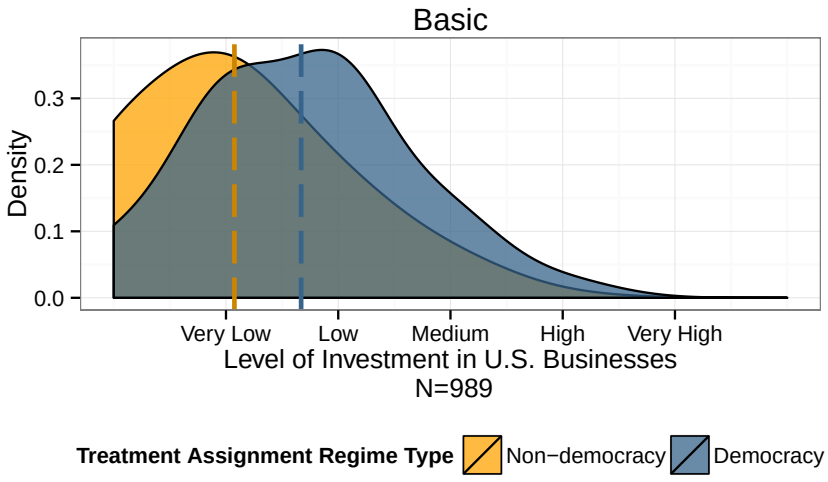
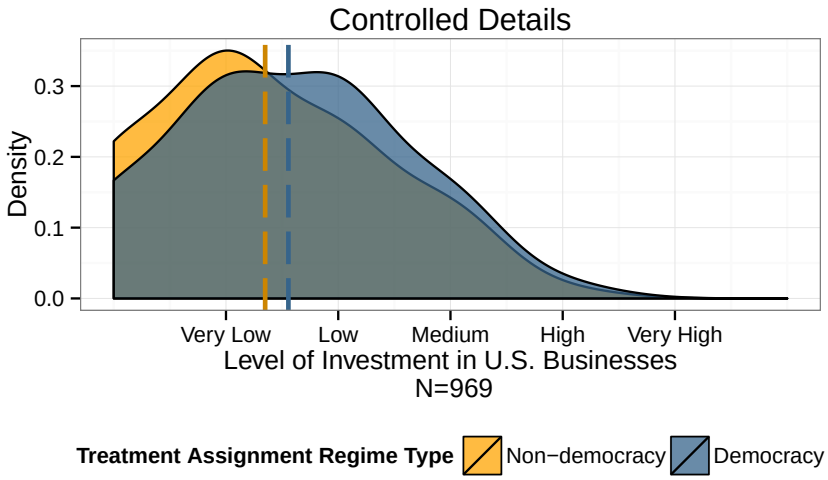


Figure 22: I: Level of Investment in U.S. Businesses





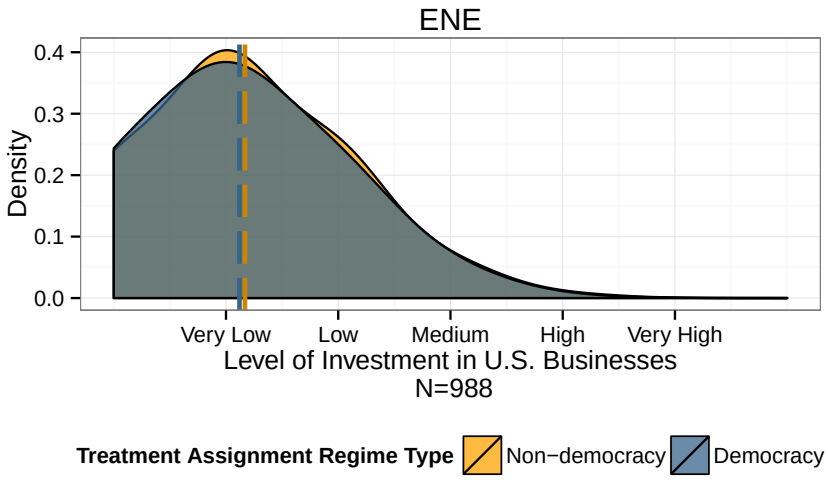


Figure 23: J: Military Spending

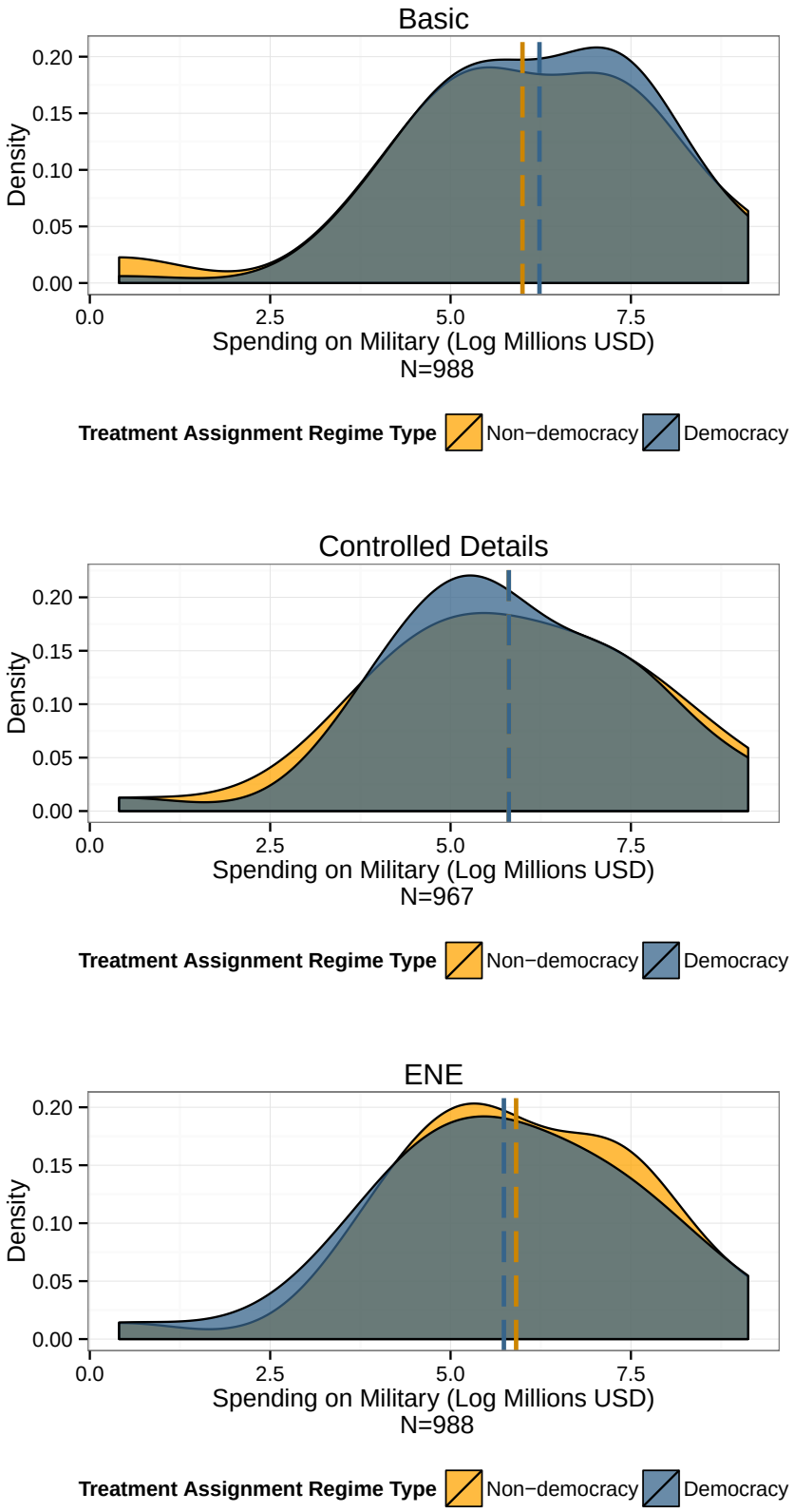


Figure 24: K: Most Likely Countries

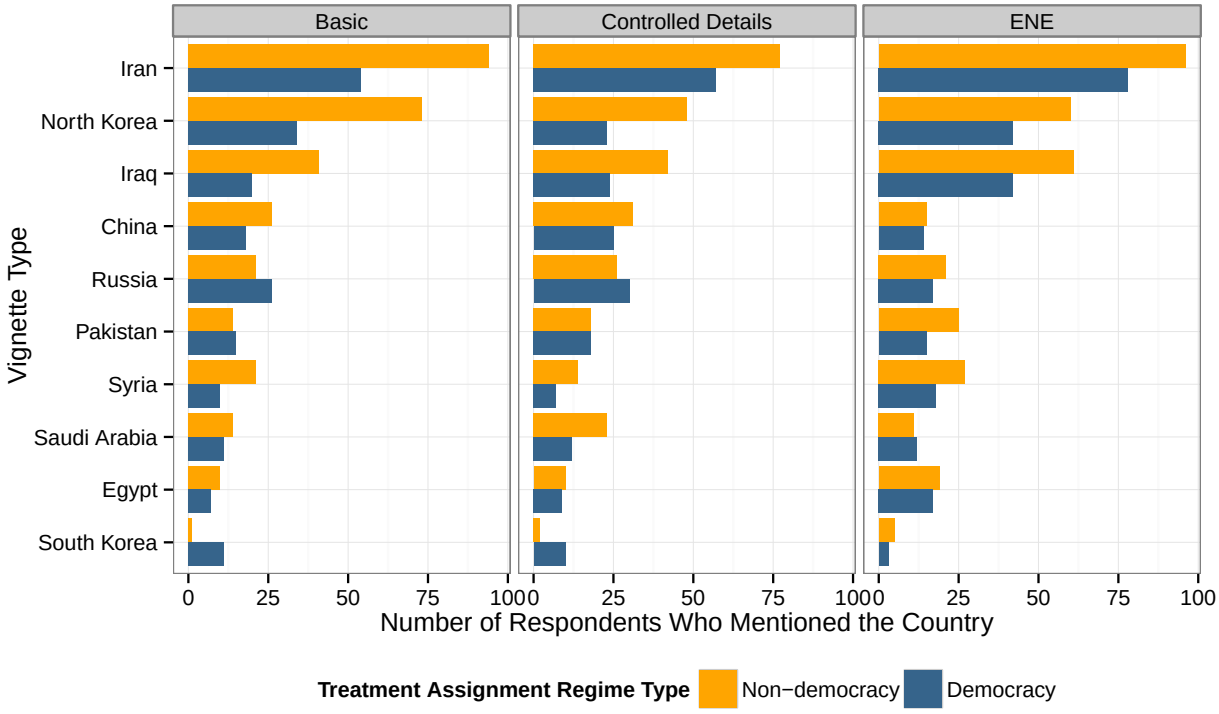
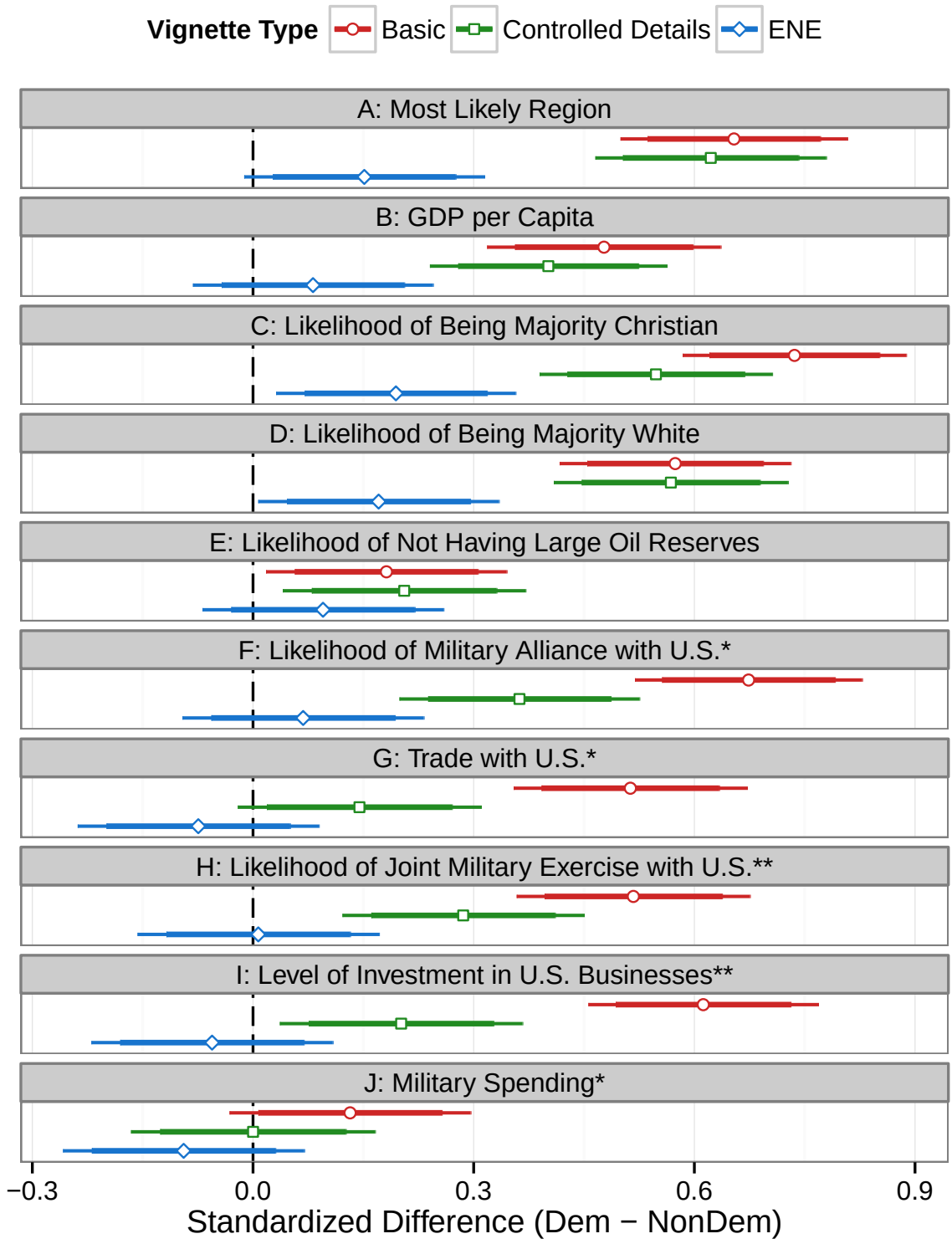
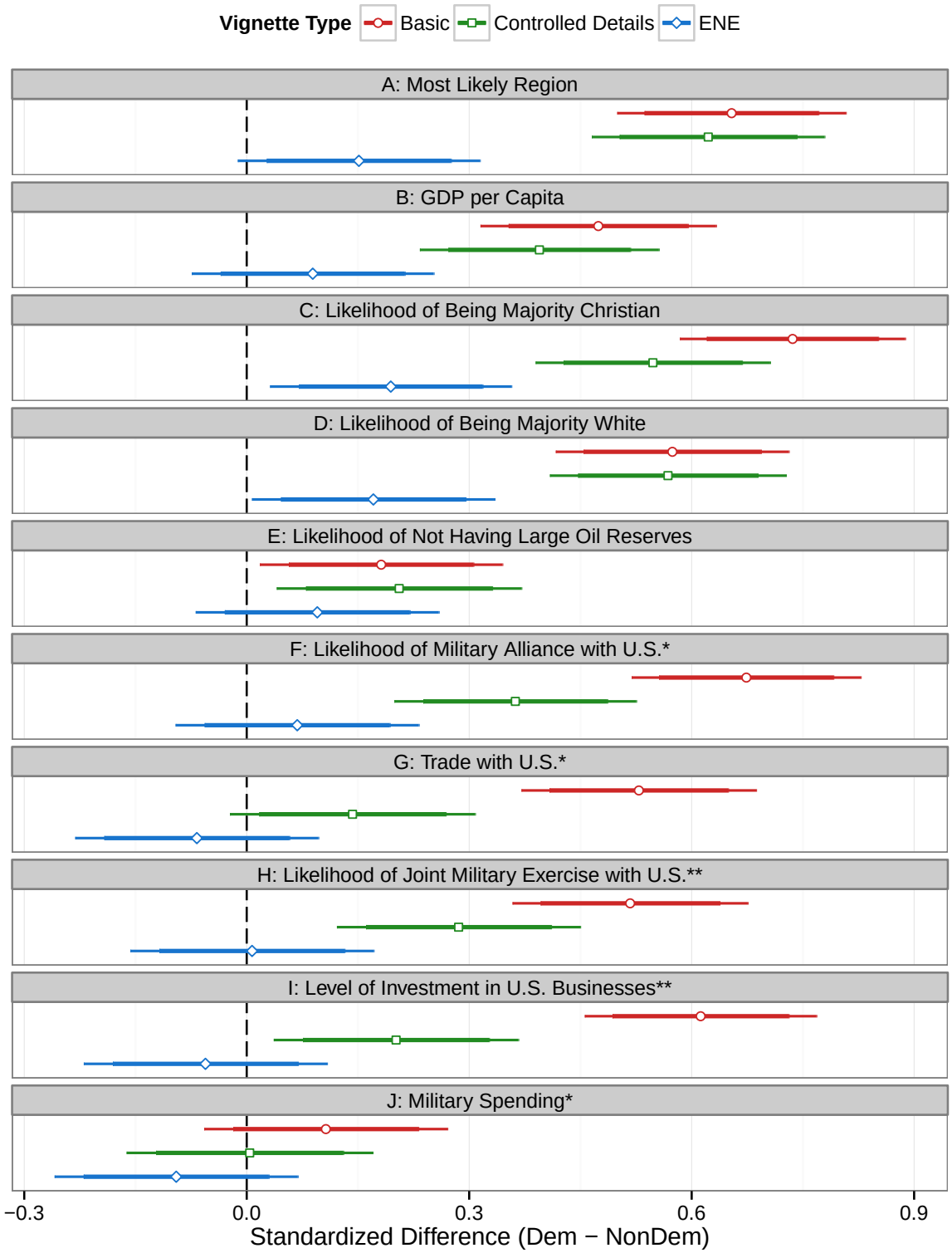


Figure 25: Placebo Test Questions Results (Standardized) including Military Spending



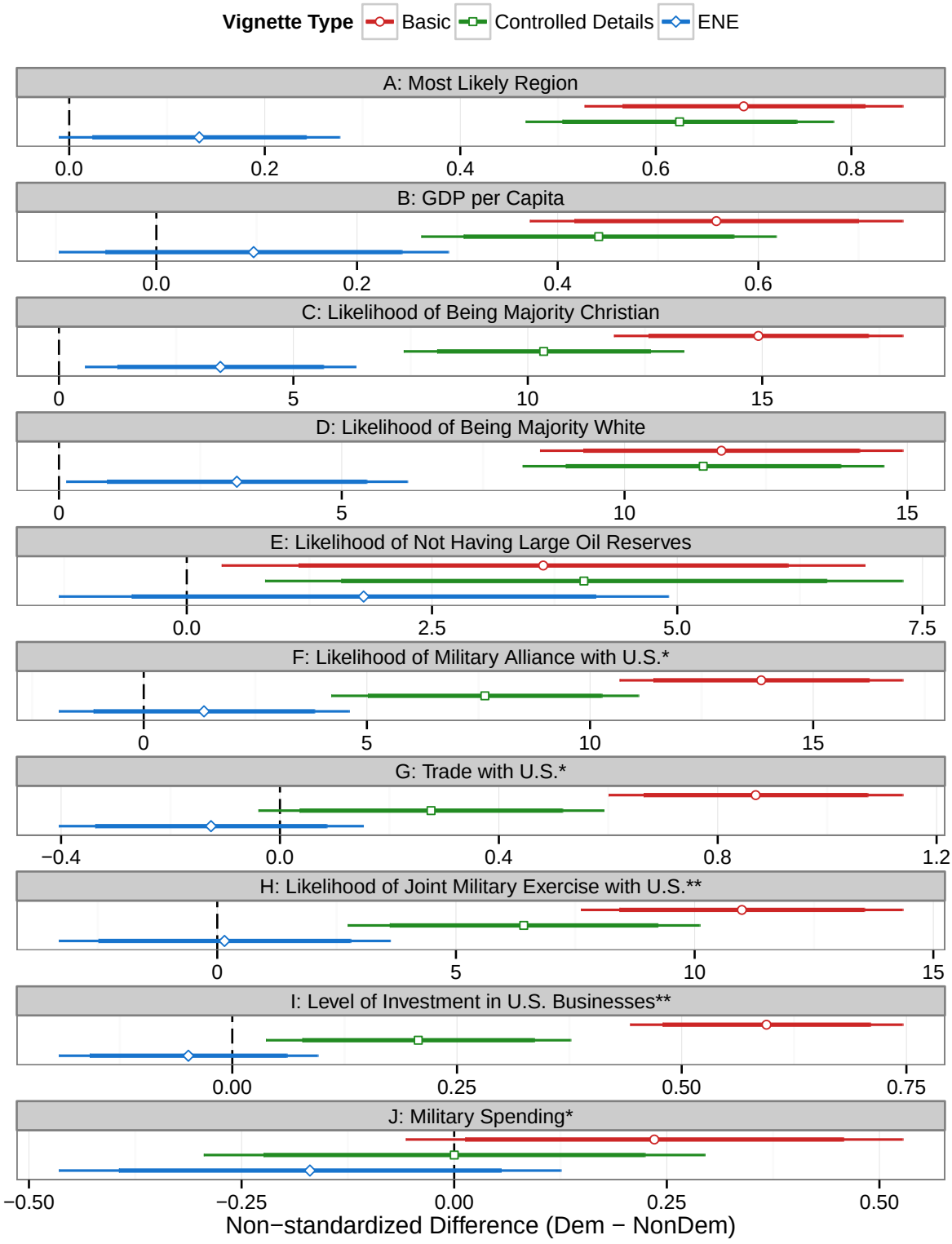
For Placebo Outcomes B, G, and J, we took the natural log of the non-standardized USD outcome before standardizing within vignette type.

Figure 26: 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 27: Placebo Test Questions Results (Non-standardized) including Military Spending



For Placebo Tests B, G, and J, the outcomes are in their original USD values and not in the ordinal scale.

D.3 Coding Treatment Measure Results

Regime 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 assigns to regime type category j. Recall that in the survey, each subject i selects 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[47] 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.
   $R1_i = \sum_j (O_j R1_{i,j})$

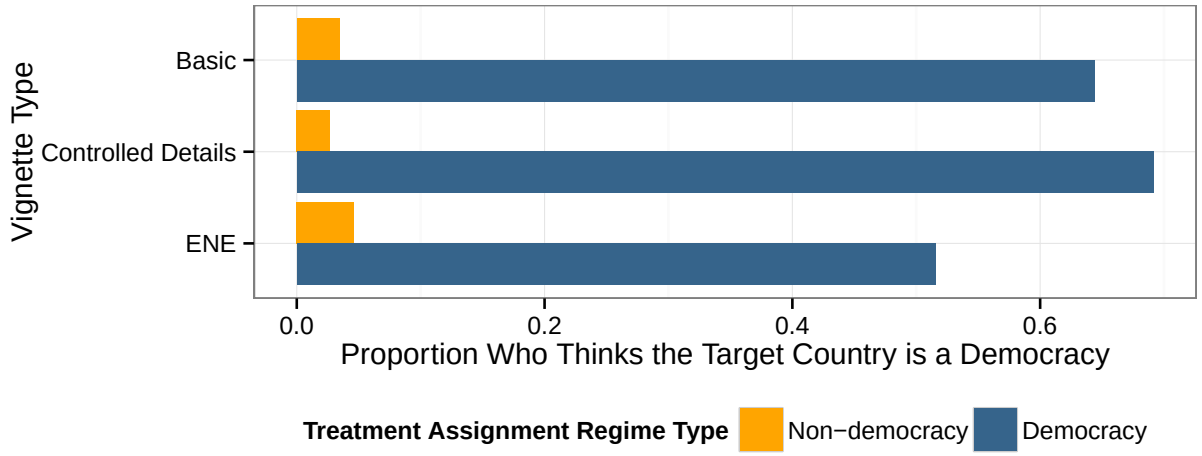
Regime Type Question 2: Characteristics of Democracies

We define $R2_i$ as the number of democratic characteristics respondent i selects, which serves as a proxy for how democratic respondent i thinks the target country is.

[47] 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.

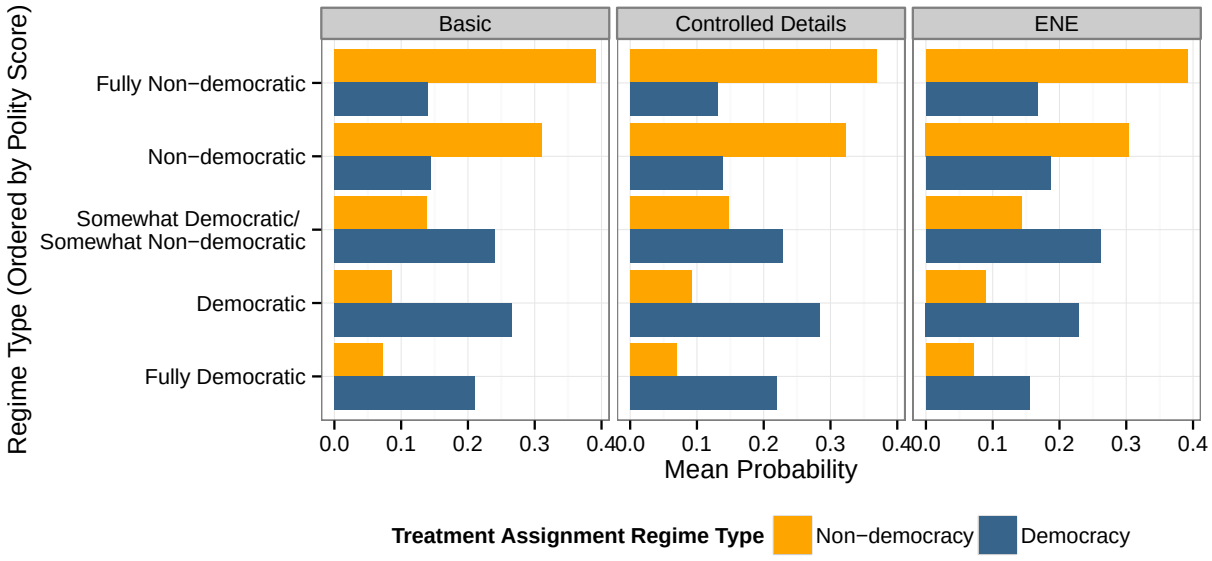
D.4 Treatment Measure Results

Figure 28: Treatment Measure: Dichotomous Measure of Perceived Regime Type



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 29: 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 he/she assigned to each regime type so that his/her probabilities sum up to 100 percent.

Figure 30: Treatment Measure: Probability of Each Regime Type

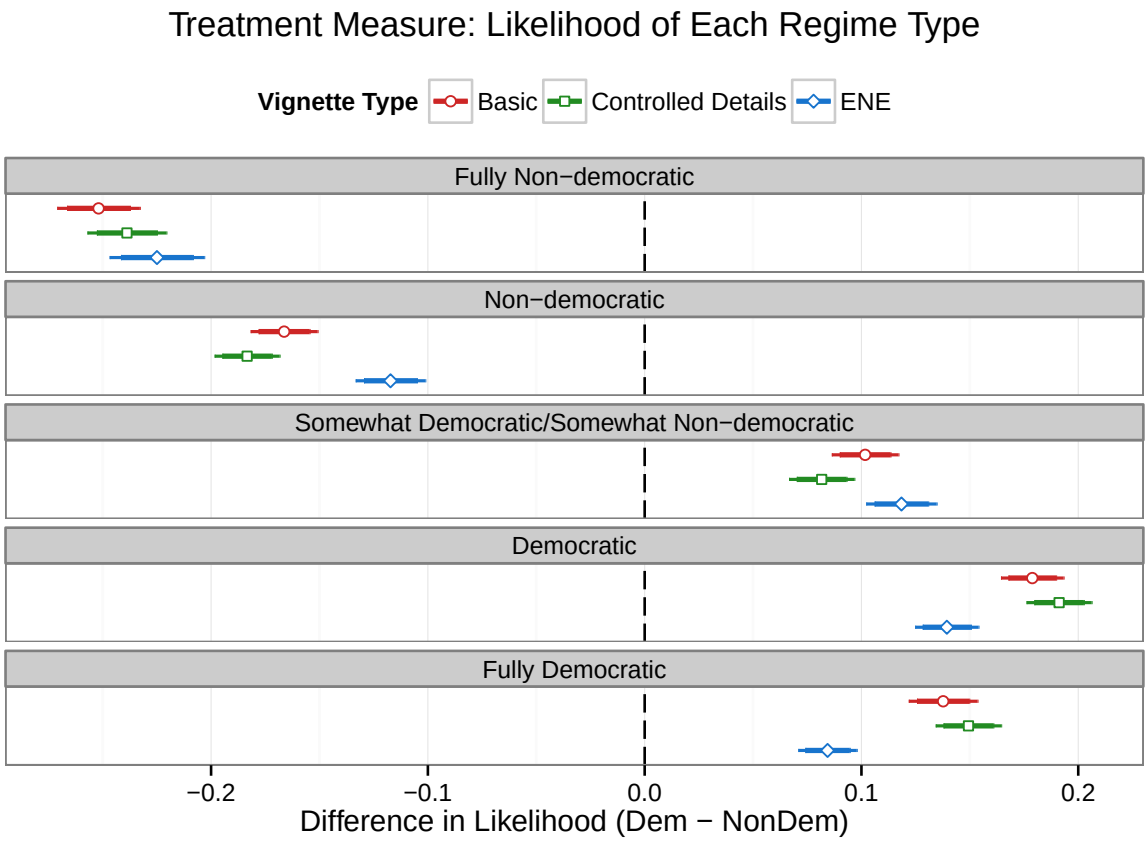


Figure 31: Treatment Measure: Imputed Polity Score

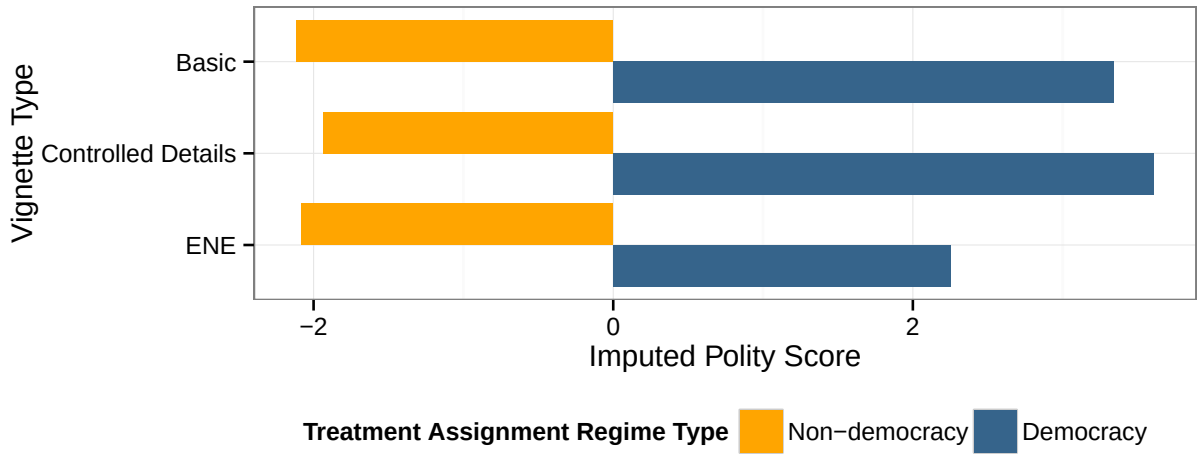


Figure 32: Treatment Measure: Characteristics of Democracies

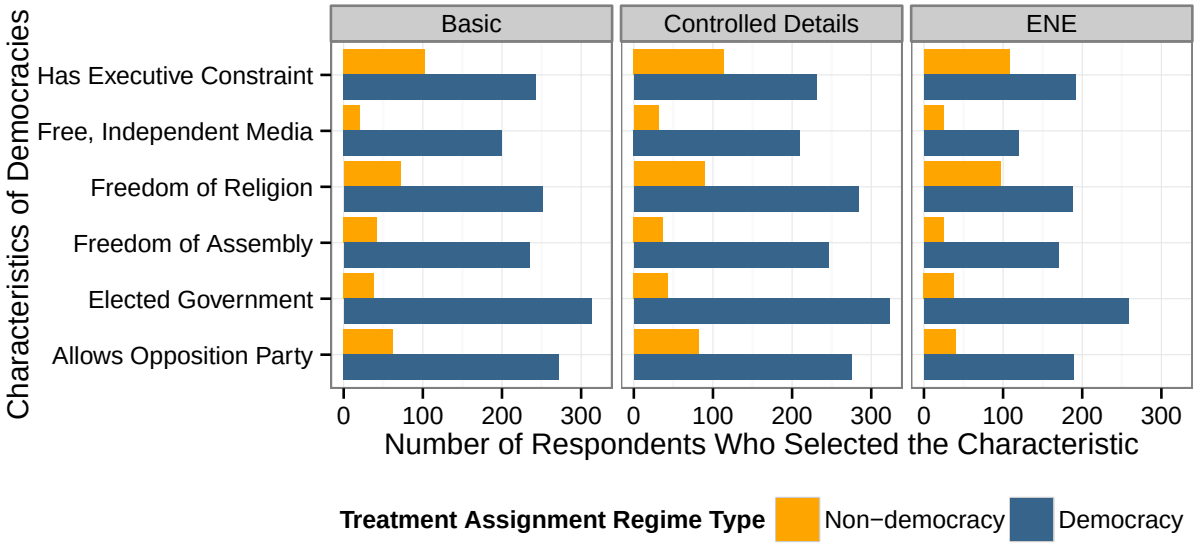


Figure 33: Treatment Measure: Characteristics of Democracies Coefficient Plot

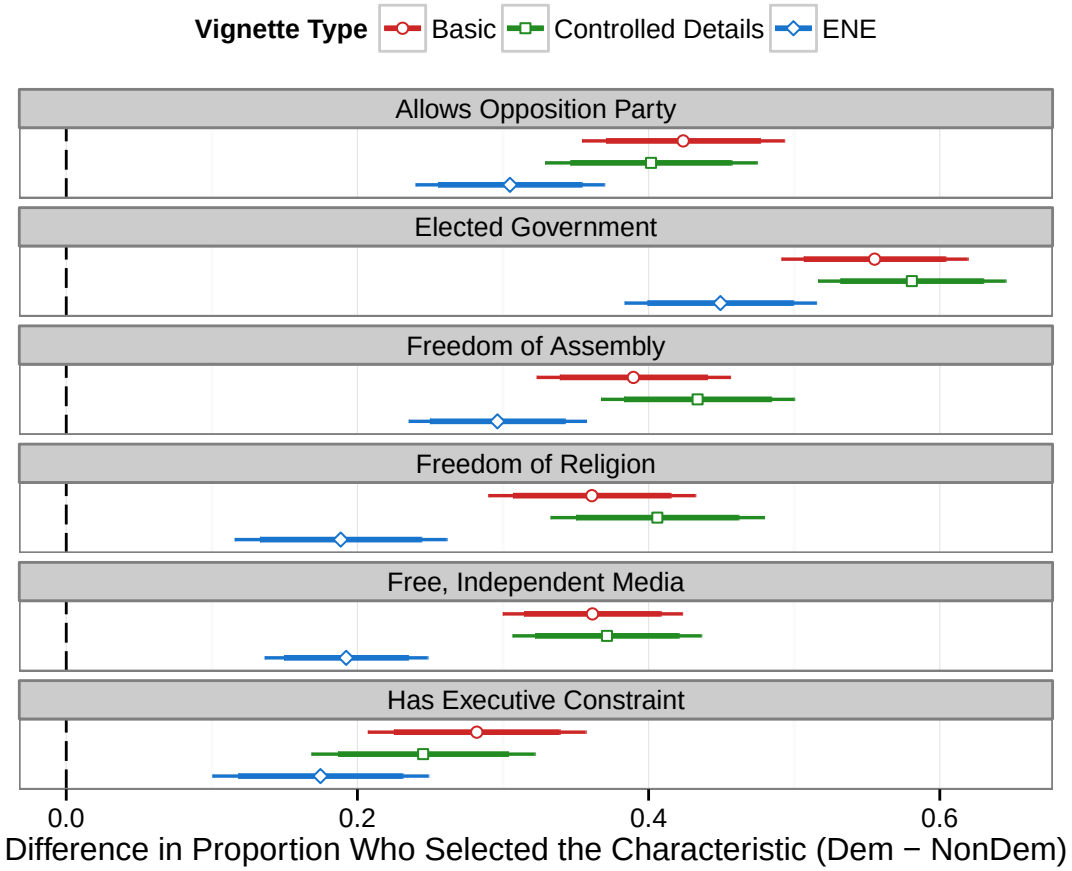
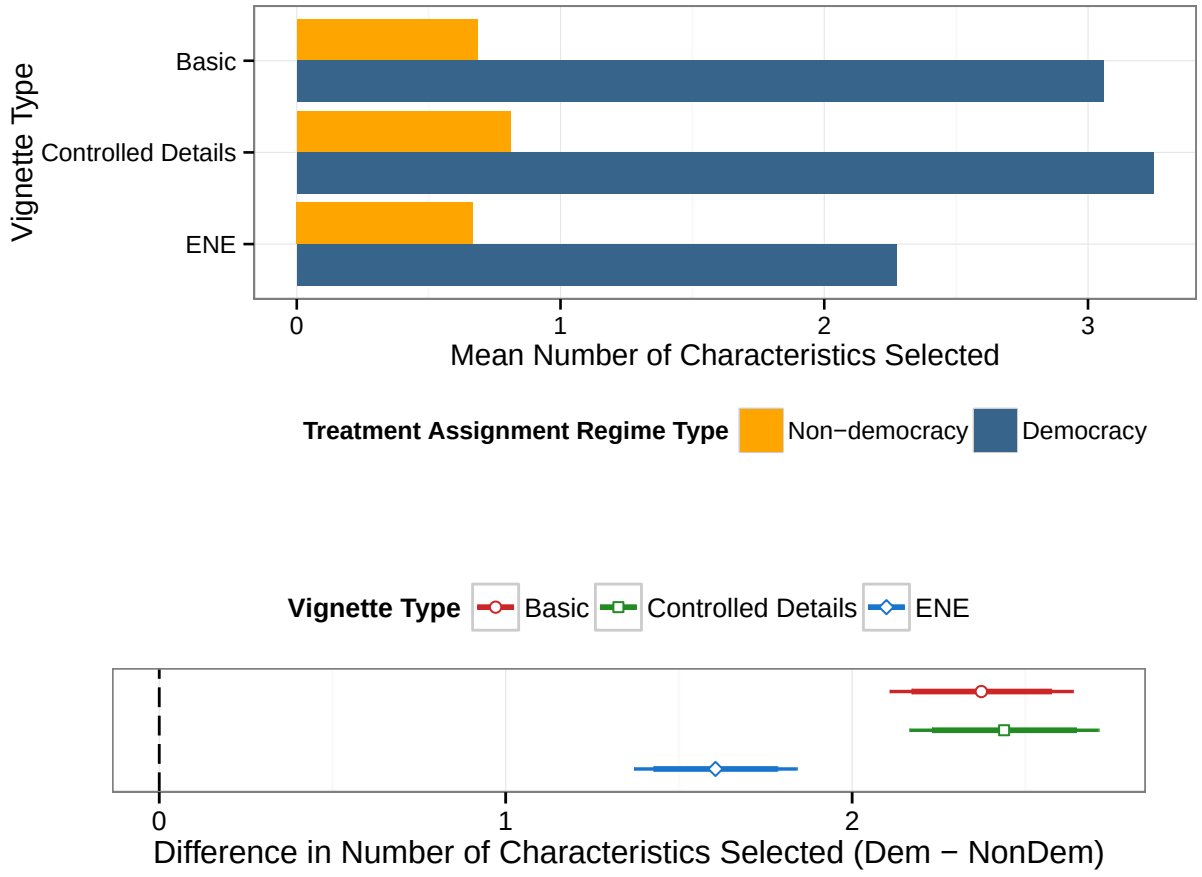
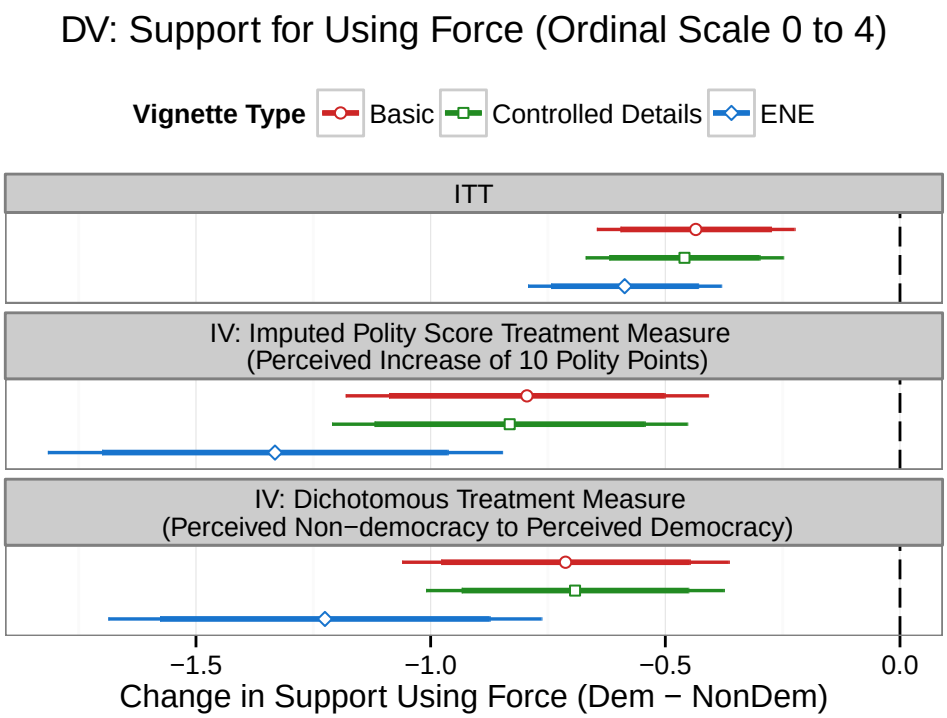


Figure 34: Treatment Measure: Characteristics of Democracies Count by Vignette Type



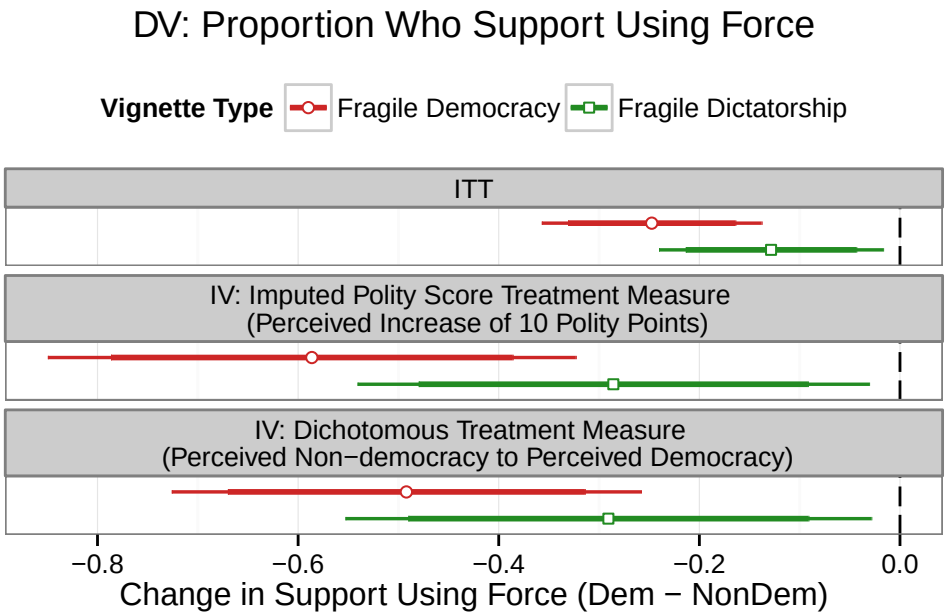
D.5 ITT and IV Estimates

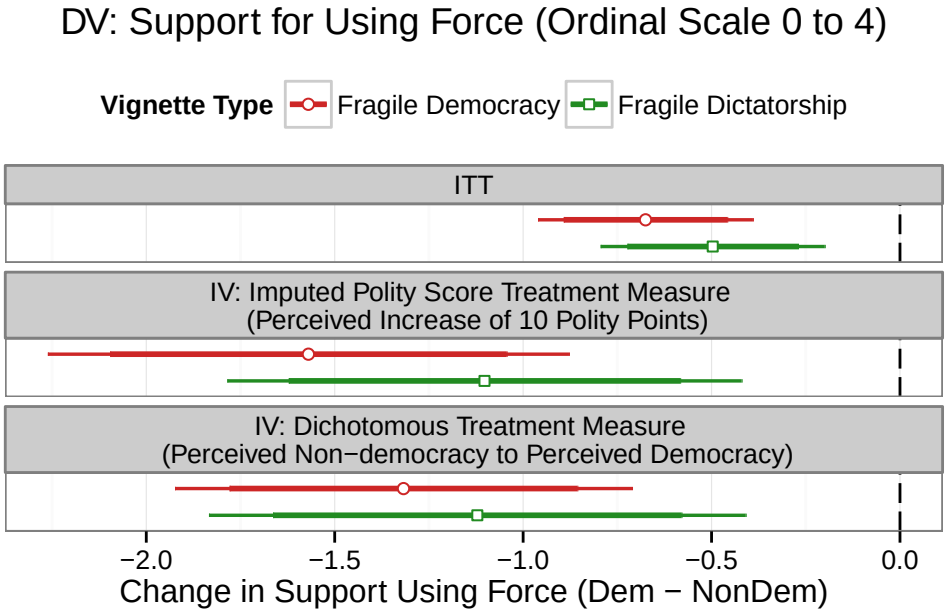
Figure 35: ITT and IV Estimates: Ordinal Measure of Support for Using Force



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 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.
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.

Figure 36: ITT and IV Estimates: Two Versions of the ENE Design





We perform the same type of analysis seen in Figures 5 and 35 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.

D.6 Abstract Encouragement Design

Respondents had 1/2 probability of being randomly assigned to read instructions that encourage them to consider the vignette scenario in the abstract. They were told "For scientific validity the situation is general, and is not about a specific country in the news today." We

examine whether those assigned to the Abstract Encouragement Design produced less imbalances in their placebo test outcomes.

Figure 37: Effect of the Abstract Encouragement Design (Standardized)

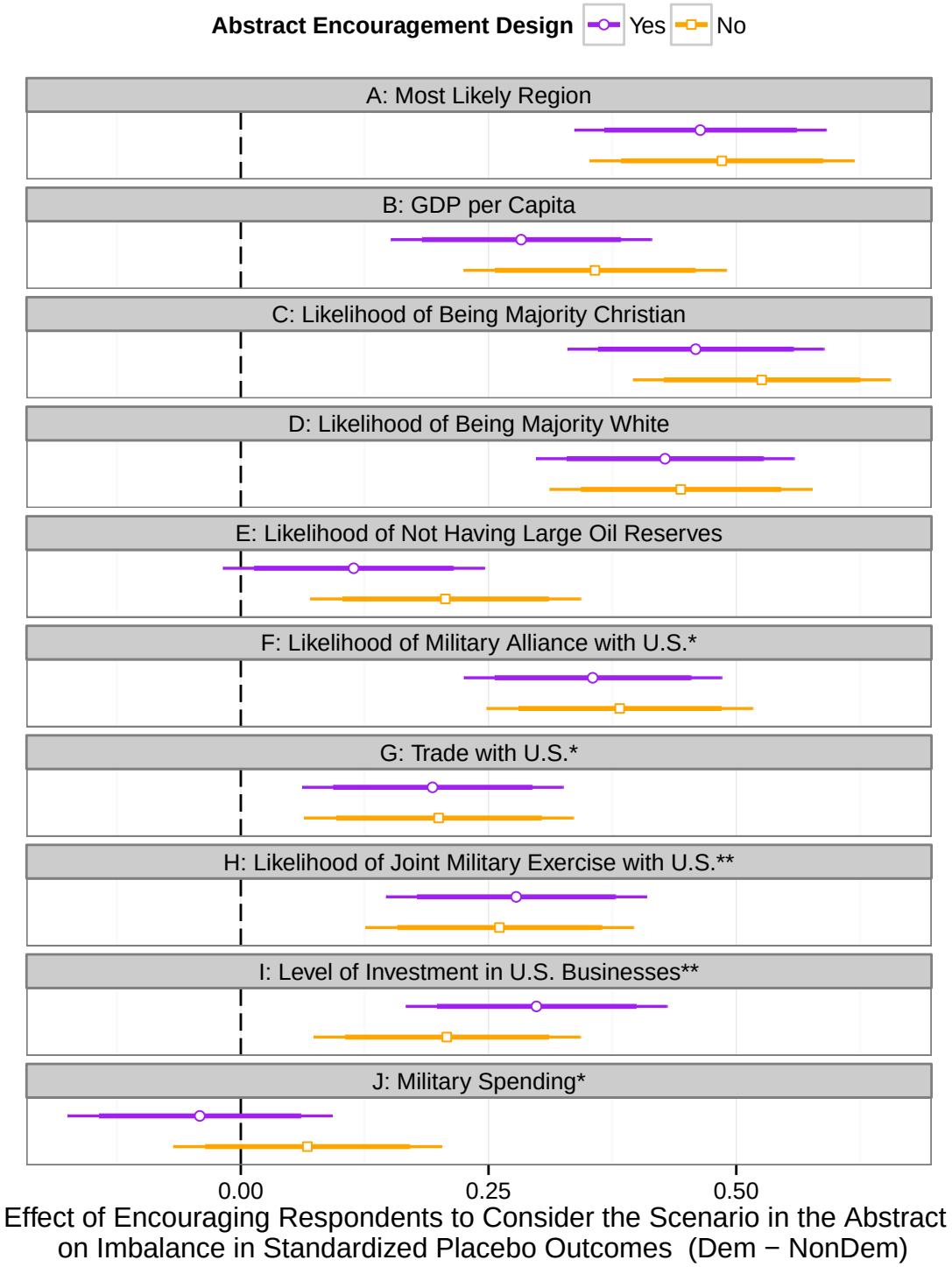
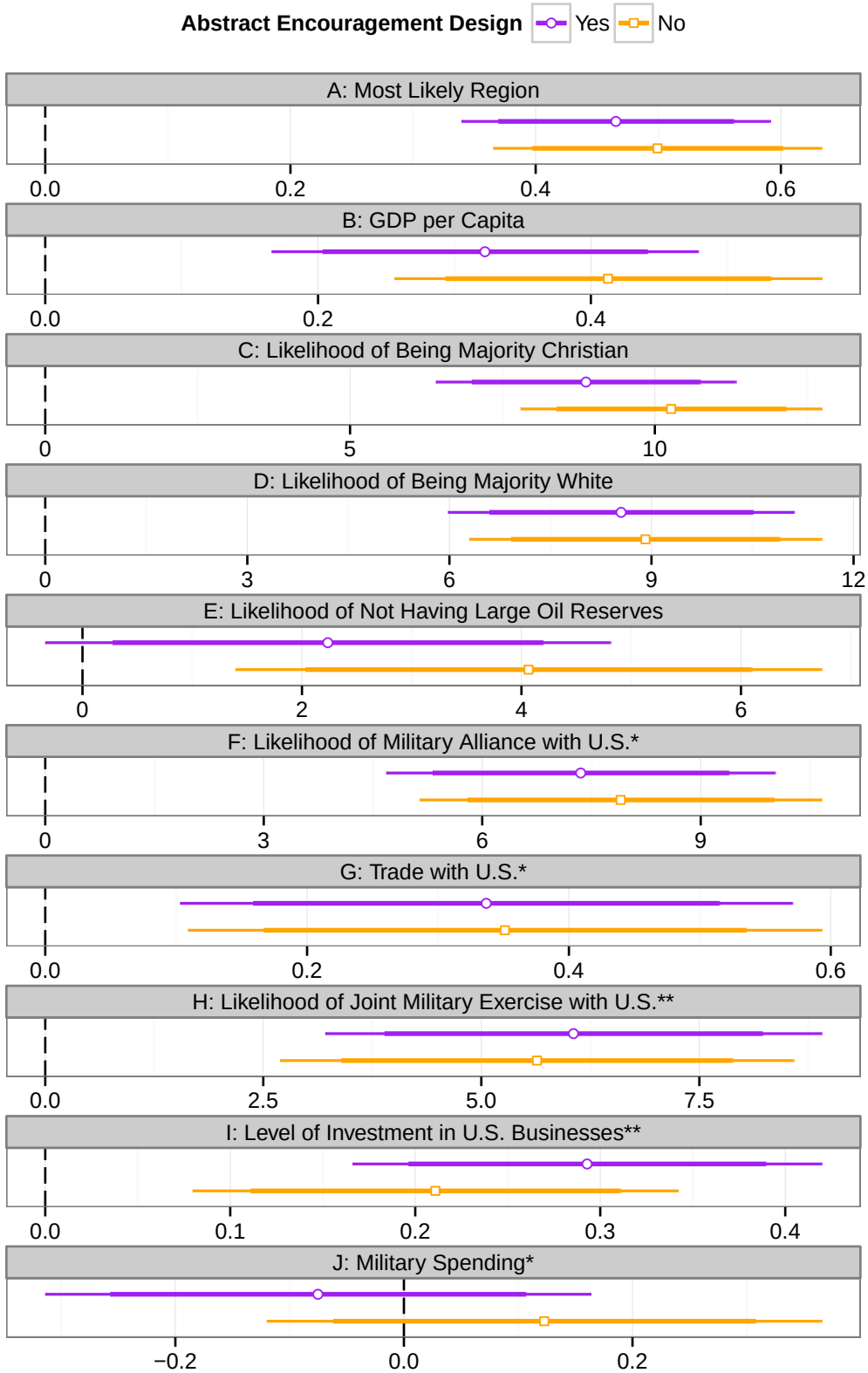


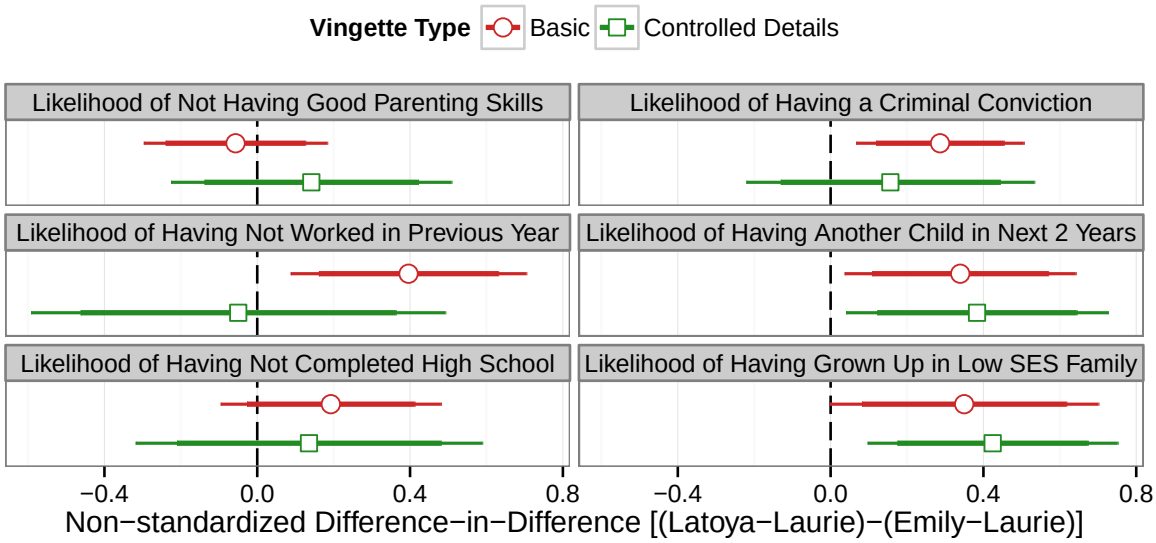
Figure 38: Effect of the Abstract Encouragement Design (Non-standardized)



E Replication and Expansion of DeSante (2013)

E.1 Placebo Test Results

Figure 39: Placebo Test Questions Results (Non-standardized)



F Latura's (2015) Survey Experiment

F.1 Test of the Survey

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

F.1.1 Basic Design Text

(*The following text is what respondents in the Basic Design see.*)

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.]

F.1.2 ENE Design Text

(*The following text is what respondents in the ENE Design see.*)

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.

F.1.3 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]

F.1.4 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]

F.2 Placebo Test Results

Figure 40: Placebo Test Questions Results (Non-standardized)