PA

Information Equivalence in Survey Experiments

Allan Dafoe[1,2], Baobao Zhang[1] and Devin Caughey[3]

[1] *Department of Political Science, Yale University, New Haven, CT 06520, USA. Email: allandafoe@gmail.com*
[2] *Governance of AI Program, University of Oxford, OX1 1PT, UK*
[3] *Department of Political Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

Abstract

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

Keywords: survey experiments, survey design, natural experiments, causal inference

1 Introduction

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

The appeal of survey experiments stems in large part from their combination of internal and external validity, which makes them a powerful tool for "inferring how public opinion works

Authors' note: Replication files for this paper can be downloaded from Dafoe, Zhang, and Caughey (2017). Further materials can be found at allandafoe.com/ie. The main studies reported in this paper have been preregistered and preanalysis plans have been posted. Superscripted capital letters indicate relevant portions of the Supplementary Information (see Section A, available at https://doi.org/10.1017/pan.2018.9). For helpful comments, we would like to thank Peter Aronow, Cameron Ballard-Rosa, Adam Berinsky, Matthew Blackwell, David Broockman, Alex Debs, Chris Fariss, Alan Gerber, Donald Green, Sophia Hatz, Dan Hopkins, Susan Hyde, Josh Kalla, Gary King, Audrey Latura, Jason Lyall, Neil Malhotra, Elizabeth Menninga, Nuno Monteiro, Brendan Nyhan, Jonathan Renshon, Bruce Russett, Cyrus Samii, Jas Sekhon, Maya Sen, Robert Trager, Mike Tomz, Jessica Weeks, Teppei Yamamoto, Sean Zeigler, Thomas Zeitzoff, and participants of the University of North Carolina Research Series, the Yale Institution for Social and Policy Studies Experiments Workshop, the Yale International Relations Workshop, the University of Konstanz Communication, Networks and Contention Workshop, the Polmeth 2014 and 2015 Summer Methods Meetings, the Survey Experiments in Peace Science Workshop, the West Coast Experiments Conference, and the Comparative Political Economy and Conjoint Analysis workshop at the University of Zurich. For support, we acknowledge the MacMillan Institute at Yale University and the National Science Foundation Graduate Research Fellowship Program. Yale IRB has granted exemption to the survey experiments reported in this paper under IRB Protocol # 1302011471.

Political Analysis (2018)
vol. 26:399–416
DOI: 10.1017/pan.2018.9

Published
3 August 2018

Corresponding author
Allan Dafoe

Edited by
Jonathan N. Katz

in the real world" (Gaines, Kuklinski, and Quirk 2007, 4). As the foregoing quotation suggests, substantive survey experiments are designed to shed light on the real-world effects of some attribute or factor, what Barabas and Jerit (2010) call the "natural treatment." In many survey experiments, the real-world effects of interest are *informational*—that is, they concern how people react to the content and format of information presented to them. For instance, how are attitudes toward anti-poverty programs affected by whether these programs are called "welfare" or "assistance to the poor" (Gilens 2002, 236)? How does the framing and sequencing of competing arguments influence support for anti-terrorism legislation (Chong and Druckman 2010)? How do the attributes included in profiles of immigrants affect support for granting them citizenship (Hainmueller, Hangartner, and Yamamoto 2015)? To learn about real-world informational effects such as these, survey experiments manipulate the information presented to survey subjects and compare the responses of subjects assigned to different informational conditions. Although informational survey experiments can be complicated by questions of external validity, causal mechanisms, and other issues, interpreting their results is greatly simplified by the fact that the natural treatment—the presentation of information—closely corresponds to what the experiment manipulates.

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

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

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

differences between experimental groups cannot be reliably attributed to the effects of the beliefs of interest.

Survey experimentalists recognize, of course, that the relationship between survey results and real-world phenomena is far from automatic. Indeed, seminar discussions of survey experiments frequently center on whether the estimated effects are due to the construct of interest or some other aspect of the manipulated text. Moreover, a few published works do reference the specific problem we describe.[2] These include our running example of Tomz and Weeks (2013), who note that previous survey experiments on the democratic peace failed to specify attributes of the scenario "that could confound the relationship between shared democracy and public support for war," such as whether "the country was also an ally, a major trading partner, or a powerful adversary" (849, 853; contrast with Mintz and Geva 1993; Johns and Davies 2012). According to our review of the survey-experimental literature, however, the IE assumption and the problems caused by its violation are not widely appreciated. Only 14% of scenario-based survey experiments in our review evince any awareness of the problem.[3] Nor has any work in political science systematically considered the issue of IE in survey experiments. Applied researchers have thus received little guidance on predicting IE violations, diagnosing them when they occur, or avoiding them in the first place.

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

We also evaluate three experimental designs that may help achieve information equivalence: abstract encouragement, covariate control, and embedded natural experiments, the last of which is our own invention. We find that *abstract encouragement*, which asks subjects to consider the scenario in the abstract rather than thinking of real-world examples, is not effective at reducing imbalance on background beliefs. *Covariate control* (CC), which entails specifying the values of background attributes in order to prevent respondents from updating about them, reduces imbalance only on attributes that are explicitly or implicitly controlled. The *embedded natural experiment* (ENE) design, which constructs a scenario in which the attribute of interest is randomly or haphazardly assigned, tends to reduce imbalance on all background characteristics. We draw empirical support for these conclusions from four survey experiments, focusing mainly on an extension of Tomz and Weeks's study of the democratic peace. We conclude by discussing the strengths and weaknesses of CC and ENE designs and offering recommendations for applied survey experimentalists.

2 Formal Exposition

In this section we formally define the real-world and survey quantities of interest (QOIs) and the role of information equivalence in linking them. For ease of exposition, we focus on cases where

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

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

the real-world QOI is the total causal effect of a single binary treatment, but our basic conclusions also hold under more general conditions (see SI, Appendix B).

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

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

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

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

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

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

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

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

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

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

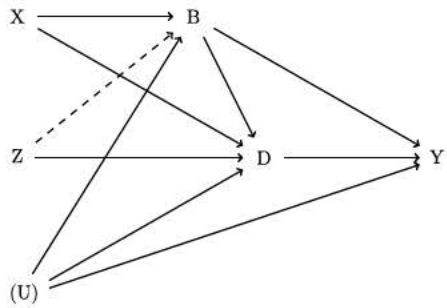


Figure 1. Graphical illustration of information equivalence in survey experiments, for the case in which B precedes D. Z denotes the survey manipulation, X other scenario details, B background beliefs, D beliefs about the causal factor of interest, Y the outcome, and U unobserved common causes of D,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

Under this assumption, the epistemic effect of interest is

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

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

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

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

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

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

The bottom line is that in order for the estimands identified by the survey experiment (the ITT and, if D is observed, the CACE) to entail conclusions about the epistemic QOI (the effect of D on Y, holding B constant), the four canonical assumptions of IV analysis (or assumptions at least as

strong) must be satisfied.[7] A crucial (and testable) implication of the IV assumptions—in particular, of the exclusion restriction ($\mathcal{A}4$)—is that different versions of the survey do not affect subjects' beliefs about background characteristics:

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

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

3 Predicting and Diagnosing IE Violations

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

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

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

Just as observational researchers validate their identification assumptions by conducting placebo tests of effects assumed to be zero (Sekhon 2009), so too should survey experimentalists

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

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

4 Preventing IE Violations

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

4.1 Abstract encouragement

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

4.2 Covariate control

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

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

4.3 Embedded natural experiments

The third strategy is to employ an *ENE*. This strategy is motivated by the realistic Bayesian model, which predicts that the survey manipulation will influence respondents' beliefs about background attributes unless they perceive the content of the manipulation to be statistically independent of—and thus to convey no information about—those attributes. In other words, a Bayesian will

not update their beliefs about background features of the scenario if and only if they believe the causal factor of interest was as good as randomly assigned in the scenario world.

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

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

More generally, ENEs may involve any treatment assignment mechanism that is at least approximately independent of background attributes. In many cases, these will involve incidents or phenomena that, if not strictly random, are at least accidental. Examples include the outcome of an assassination attempt (Jones and Olken 2009) or an episode in which two fighter jets either collide or barely miss each other (Dafoe, Hatz, and Zhang 2018). ENEs based on other quasi-experimental designs, such as regression discontinuity, are also possible. In practice, ENEs will fall somewhere on a spectrum of as-if randomness, just as observational natural experiments do (Dunning 2012).[E]

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

5 An Application to the Democratic Peace

We evaluate evidence of IE violations, and the effectiveness of the various strategies for mitigating them, using several applications.[8] The first and most elaborate is a replication and extension

8 Replication files for all the analyses in this paper can be downloaded from Dafoe *et al.* (2017).

of Tomz and Weeks's (2013) survey experiment on the mass basis for the democratic peace. Using placebo tests, we show that randomly manipulating whether a target country is described as democratic is not sufficient to prevent respondents from updating their beliefs about the background attributes of the country, potentially biasing the effect of interest. We further demonstrate that abstract encouragement does little to mitigate these violations of IE, and that CC does so only on attributes explicitly or indirectly controlled in the vignette. An ENE design is most effective at achieving IE. Figure 3 and Appendix D in the SI demonstrate the use of an IV estimator to estimate the CACE in this study, and discusses relevant assumptions and issues of interpretation.

5.1 Survey design

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

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

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

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

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

Finally, respondents read the conclusion of the scenario:

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

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

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

> (Z_{ENE}) Five years ago a country, Country A, was a fragile democracy. It had a democratically elected government, headed by a popular president. At the time, a well-researched U.S.

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

State Department report concluded that without this president, there was a very high probability that the country's military would overthrow the government to set up a dictatorship.

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

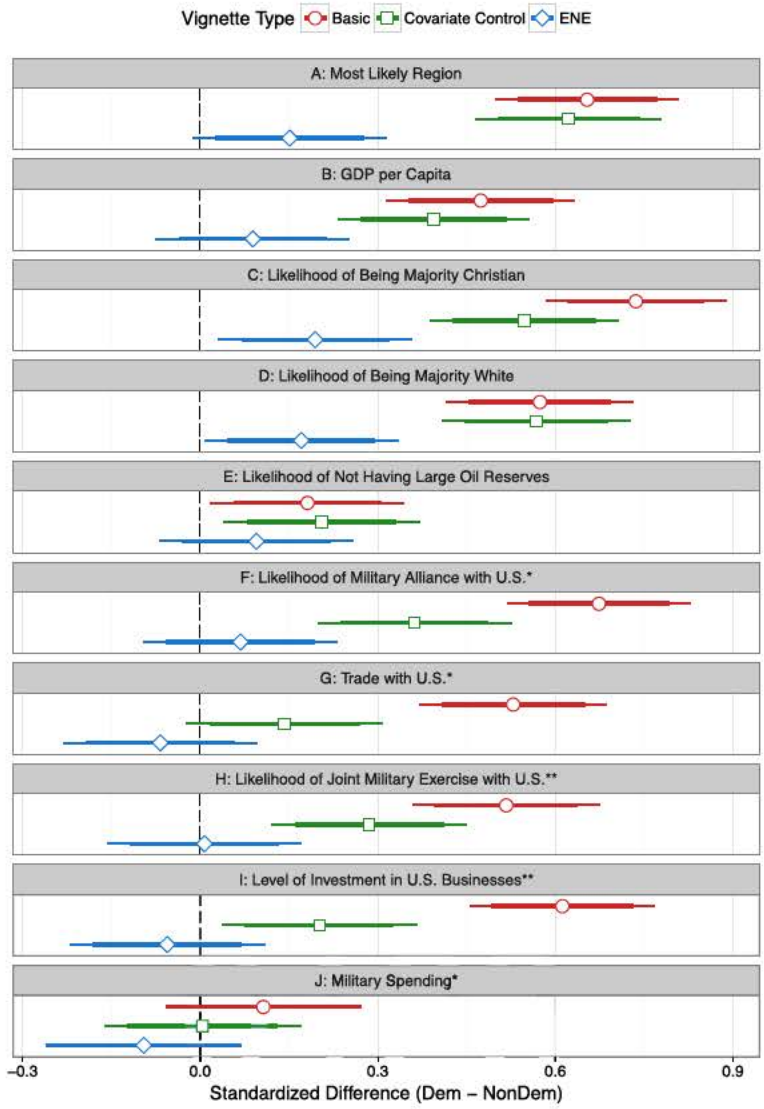
After reading the vignette, respondents were asked about their support for using force against the target country, as well as demographic questions and questions related to the placebos, potential mechanisms, and the treatment. We randomized the order of all these questions (we did not detect any relevant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

5.2 Placebo tests

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

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

The ENE design was by far the most effective at reducing imbalance on placebo attributes. For most placebos, the imbalance is much less severe than for the other two designs, and in no case was it detectably worse. Strikingly, even attributes that were *explicitly controlled* in the CC design were more balanced in the ENE design. This result is not a symptom of a weak manipulation, as the ENE manipulation's effect on perceived regime was nearly as great as the other designs (Figure 3; SI, Appendix D). Overall, the results suggest that just as natural experiments, when truly



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

Figure 2. Placebo tests by vignette type.

as-if random, tend to yield plausible causal inferences in observational studies, so too are ENEs a potentially effective strategy for credible causal inference in survey experiments.

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

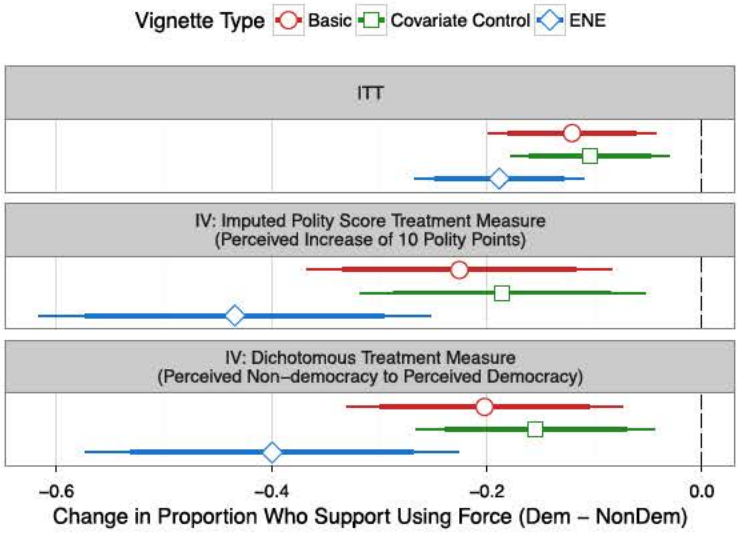


Figure 3. Effect estimates from different versions of the democratic peace experiment. The error bars represent the 95% and 99% confidence intervals.

6 Extensions to Other Studies

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

6.1 Effects of coercive harm

Dafoe, Hatz, and Zhang (2018) investigate whether harm experienced in a coercive context provokes resolve and desire for retaliation, through survey experiments fielded in China and the United States (SI, Appendix H). The scenario depicts China and the United States engaged in a tense dispute in the East China Sea. In the basic design of the survey given to American respondents (n = 705), the control describes the dispute, the treatment also describes China shooting down a US military plane for trespassing in Chinese airspace. In the ENE version (n = 731) the US plane is made to crash (or not) in an as-if random way, that is nevertheless part of the coercive context: "the Chinese plane was flying dangerous maneuvers around the American plane, making several close passes. On the third pass the planes [almost collided/collided]. . . ." Figure 4 shows how perceptions of hostile military intent were imbalanced in the basic design, but become balanced in the ENE design. The ENE design thus has a claim to better isolating the causal factor of interest: harm in a coercive context.

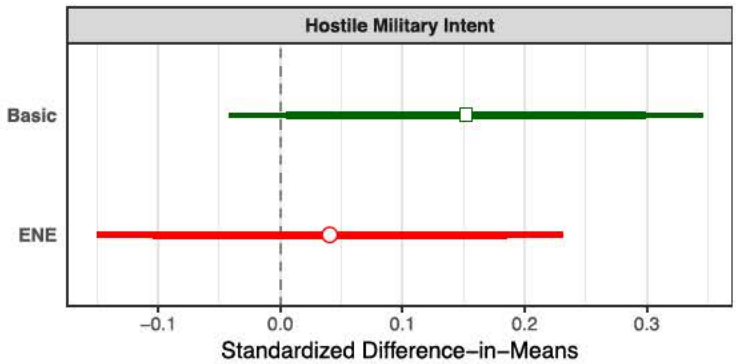


Figure 4. Placebo test results from Dafoe, Hatz, and Zhang (2018). The error bars represent the 95% and 99% confidence intervals.

6.2 Effects of subsidized childcare

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

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

6.3 Why is Latoya discriminated against?

Finally, we replicated and extended Desante's (2013) study of whether and why Americans are more willing to support welfare for people who are white than black. Our basic design manipulates the name of the welfare applicant (e.g., Emily vs. Latoya), and holds constant the number and age of the applicant's children. Following Desante (2013), our CC design additionally includes a "Worker Quality Assessment" (with values of "Poor" or "Excellent"). In doing so, the CC design hopes to rule out "principled conservative" reasons for discrimination, leaving only "racial animus" as the basis for discrimination. For placebo questions, we sought characteristics that would be the basis for "principled conservative" discrimination, which led us to a set of questions

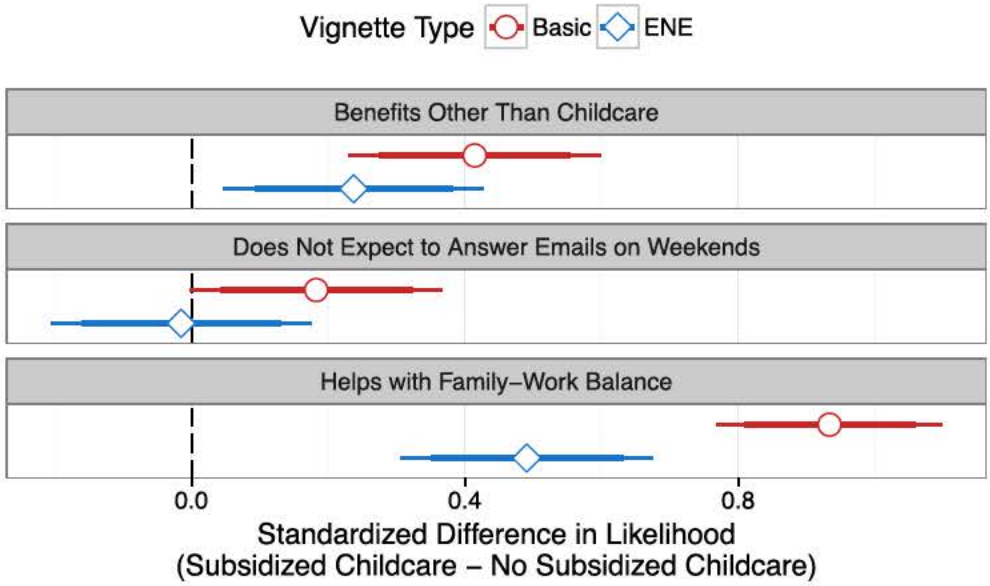


Figure 5. Placebo test results from Latura (2015). The error bars represent the 95% and 99% confidence intervals.

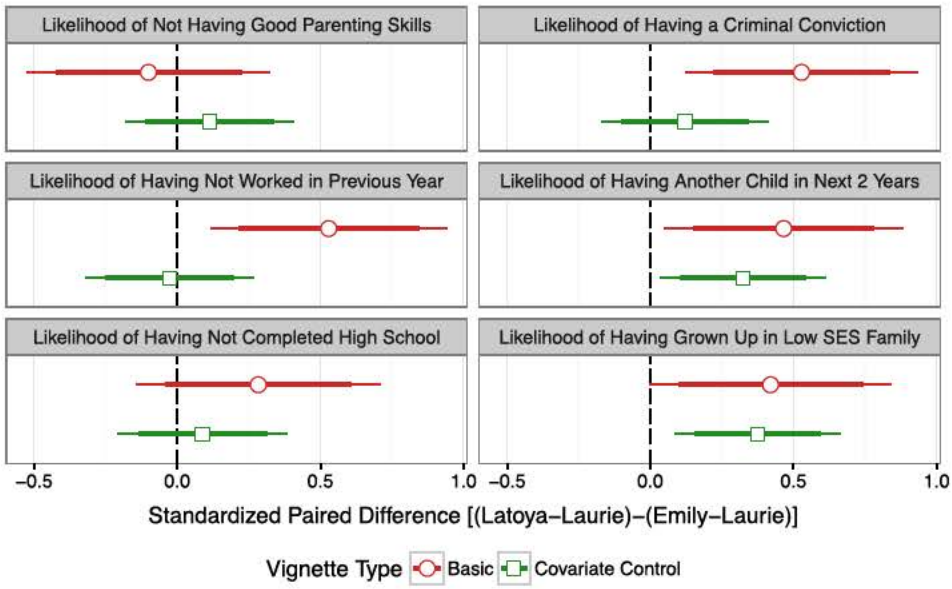


Figure 6. Placebo test results from the replication and expansion of Desante (2013). The error bars represent the 95% and 99% confidence intervals.

from the North Carolina welfare agency.[H] We also included two additional questions: whether the applicant grew up in a low socio-economic status (SES) family and whether the applicant is likely to have another child in the next two years.[I]

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

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

7 Limitations of Different Designs

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

7.1 Limitations of covariate control

While CC did not achieve IE in our examples, it did reduce imbalance on those background features that were specified. Is the solution then simply to devise extremely detailed scenarios that specify every possible background feature? One potential problem with this strategy is respondent exhaustion or satisficing (Krosnick 1999; but see Bansak *et al.* 2018). A more fundamental limit, however, is what might be termed the *plausibility constraint*: as the number of controls increases, so too does the probability of a vignette that is implausible to respondents. As in observational studies, the more variables we control for, the more likely it is for a counterfactual to go beyond the support of the data (King and Zeng 2006). There is, for example, simply no empirical referent for a Western European democracy that uses Sharia law for criminal proceedings. Researchers can prune away implausible vignettes (Hainmueller, Hopkins, and Yamamoto 2014, 20), but as the number of control variables increases the subset of plausible combinations will tend to become smaller.

A related strategy is to use a *proper-noun vignette*: specifying a real-world referent in the scenario, thus implicitly controlling for an almost infinite number of background attributes. For example, in another survey experiment we provided selective information about a country's past foreign-policy behavior; in one version the country was identified as Iran. We (Renshon, Dafoe, and Huth 2018) found that naming the country reduced imbalance on background attributes such as the country's regime type, but it did not eliminate it. This is likely to always be the case since the vignette is a hypothetical, allowing for the possibility of other unspecified changes in background covariates. Violations of IE will likely be more severe the less respondents know about the real-world referent, since then respondents will infer more about background features from the treatment prompt (Z). A variant of this strategy that avoids hypotheticals, which we label a *selective-history design* (Dafoe and Weiss 2018; Weiss and Dafoe 2018), entails selectively informing or reminding the respondent about certain facts of a historical episode. Such a strategy has promise, but it is limited by the kinds of scenarios generated by the real world and can still induce IE violations if respondents are not perfectly informed about history.

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

7.2 Limitations of embedded natural experiments

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

A second concern is that ENE designs only allow us to estimate a narrow estimand—the effects for a narrowly defined set of scenarios—and not the general causal estimand the researchers may have had in mind. In the democratic peace study, our ENE only allows us to estimate the effect

for the kinds of countries and manipulations that fit the ENE scenario. This is a *local* causal effect because it captures the average effect among only those countries that fit the ENE. Consistent with this concern, our respondents report perceiving the ENE scenario to be much more typical of the Middle East and North Africa than Western Europe (see SI, Appendix G, Figure 12). Relatedly, the effects identified by an ENE may be specific to the particular manipulation. For example, if Americans are especially concerned about leaders of fragile democracies, our democratic peace ENE estimand (the effect of democracy in countries prone to coup attempts) may be quite different from that in other kinds of countries. It is even possible for ENEs to introduce their own IE violations if the cause of interest in the scenario is "bundled" with other causes. For example, it may be that surviving an assassination attempt makes the surviving leader more sympathetic to subjects, in addition to changing respondents' beliefs about the country's regime type.

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

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

8 Conclusion and Recommendations

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

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

(1) **State your QOI and theorize about information equivalence.** Think clearly about what real-world counterfactual you are trying to reproduce. What set of background characteristics need to be held fixed for this to succeed? What background characteristics are correlated in the real world with treatment, and thus are most at risk of being influenced by your survey manipulation?
(2) **Measure your causal factor.** This can be used to evaluate the assumption of a monotonic first stage, to estimate complier average treatment effects, and to understand the kinds of variation in D that are informing your estimates.
(3) **Employ a credible design.** Find a credible hypothetical natural experiment that you can embed into your scenario, and for which the resulting causal effect is relevant.
(4) **Control covariates.** If you cannot employ an ENE, employ CC designs to rule out at least some sources of IE violations.
(5) **Diagnose violations of IE.** Employ placebo tests to evaluate whether IE seems plausible, and if not, why not.
(6) **Theorize the bias.** Formally or informally reason through the direction and size of biases likely to come from the violations of IE. A causal estimate will be more compelling if you can persuasively argue that the bias is likely to be small or in the opposite direction as your prediction.
(7) **Qualify your inferences.** Acknowledge the remaining risk of violations of IE. Recognize that your estimated causal effects are local to the kinds of scenarios that you presented and the respondents' inferences about the context of the scenario.

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

Supplementary Materials

For supplementary materials accompanying this paper, please visit https://doi.org/10.1017/pan.2018.9.

References

Acharya, Avidit, Matthew Blackwell, and Maya Sen. Forthcoming. Analyzing causal mechanisms in survey experiments. *Political Analysis*, to appear.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91(434):444–455.

Aronow, Peter M., and Cyrus Samii. 2016. Does regression produce representative estimates of causal effects? *American Journal of Political Science* 60(1):250–267.

Baker, Andy. 2015. Race, paternalism, and foreign aid: evidence from US public opinion. *American Political Science Review* 109(1):93–109.

Bansak, Kirk, Jens Hainmueller, Daniel J. Hopkins, and Teppei Yamamoto. 2018. Beyond the breaking point? Survey satisficing in conjoint experiments. *Political Analysis* 26(1):112–119.

Barabas, Jason, and Jennifer Jerit. 2010. Are survey experiments externally valid? *American Political Science Review* 104(2):226–242.

Brader, Ted, Nicholas A. Valentino, and Elizabeth Suhay. 2008. What triggers public opposition to immigration? Anxiety, group cues, and immigration threat. *American Journal of Political Science* 52(4):959–978.

Butler, Daniel M., and Jonathan Homola. 2017. An empirical justification for the use of racially distinctive names to signal race in experiments. *Political Analysis* 25(1):122–130.

Butler, Daniel M., and Eleanor Neff Powell. 2014. Understanding the party brand: experimental evidence on the role of valence. *Journal of Politics* 76(2):492–505.

Chong, Dennis, and James N. Druckman. 2010. Dynamic public opinion: communication effects over time. *American Political Science Review* 104(4):663–680.

Dafoe, Allan, and Jessica Weiss. 2018. Provocation, public opinion, and international crises: Evidence from China. http://www.allandafoe.com/china.

Dafoe, Allan, Sophia Hatz, and Baobao Zhang. Coercion and provocation. Unpublished working paper. http://www.allandafoe.com/provocation.
Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2015. Confounding in survey experiments. Paper presented at the Annual Meeting of The Society for Political Methodology, University of Rochester, Rochester, NY, July 23.
Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2017. Replication data for: Information equivalence in survey experiments, https://doi.org/10.7910/DVN/KVZXE8, Harvard Dataverse, V1, UNF:6:pUX5QK8MgtHBJ2cJQwYiyw==.
Desante, Christopher D. 2013. Working twice as hard to get half as far: race, work ethic, and America's deserving poor. *American Journal of Political Science* 57(2):342–356.
Dunning, Thad. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach*. New York: Cambridge.
Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk. 2007. The logic of the survey experiment reexamined. *Political Analysis* 15(1):1–20.
Gilens, Martin. 2002. An anatomy of survey-based experiments. In *Navigating Public Opinion: Polls, Policy, and the Future of American Democracy*, ed. Jeff Manza, Fay Lomax Cook, and I. Benjamin. New York: Oxford, pp. 232–250.
Hainmueller, Jens, and Michael J. Hiscox. 2010. Attitudes toward highly skilled and low-skilled immigration: evidence from a survey experiment. *American Political Science Review* 104(1):61–84.
Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto. 2015. Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences* 112(8):2395–2400.
Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. 2014. Causal inference in conjoint analysis: understanding multidimensional choices via stated preference experiments. *Political Analysis* 22(1):1–30.
Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011. Unpacking the black box of causality: learning about causal mechanisms from experimental and observational studies. *American Political Science Review* 105(4):765–789.
Johns, Robert, and Graeme A. M. Davies. 2012. Democratic peace or clash of civilizations? Target states and support for war in Britain and the United States. *Journal of Politics* 74(4):1038–1052.
Jones, Benjamin F., and Banjamin A. Olken. 2009. Hit or miss? The effect of assassinations on insitutions and war. *American Economic Journal: Macroeconomics* 1(2):55–87.
Kahneman, Daniel, and Amos Tversky. 1973. On the psychology of prediction. *Psychological Review* 80(4):237–251.
Kertzer, Joshua D., and Ryan Brutger. 2016. Decomposing audience costs: bringing the audience back into audience cost theory. *American Journal of Political Science* 60(1):234–249.
King, Gary, and Langche Zeng. 2006. The dangers of extreme counterfactuals. *Political Analysis* 14(2):131–159.
Krosnick, Jon A. 1999. Survey research. *Annual Review of Psychology* 50(1):537–567.
Latura, Audrey. 2015. Material and normative factors in womens professional advancement: experimental evidence from a childcare policy intervention. Paper presented at the American Politics Research Workshop, Harvard University, April 28. http://lists.fas.harvard.edu/pipermail/gov3004-list/attachments/20150427/ea95d274/attachment-0001.pdf.
Marsden, Peter V., and James D. Wright. 2010. *Handbook of Survey Research*. Bingley: Emerald Group Publishing.
Middleton, Joel A., Marc A. Scott, Ronli Diakow, and Jennifer L. Hill. 2016. Bias amplification and bias unmasking. *Political Analysis* 24(4):307–323.
Mintz, Alex, and Nehemia Geva. 1993. Why don't democracies fight each other? An experimental study. *Journal of Conflict Resolution* 37(3):484–503.
Mutz, Diana C. 2011. *Population-Based Survey Experiments*. Princeton, NJ: Princeton.
Renshon, Jonathan, Allan Dafoe, and Paul Huth. 2018. Leader influence and reputation formation in world politics. *American Journal of Political Science* 62(2):325–339.
Sekhon, Jasjeet S. 2009. Opiates for the matches: matching methods for causal inference. *Annual Review of Political Science* 12(1):487–508.
Sen, Maya, and Omar Wasow. 2016. Race as a 'bundle of sticks': designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science* 19:499–522.
Sher, Shlomi, and Craig R. M. McKenzie. 2006. Information leakage from logically equivalent frames. *Cognition* 101(3):467–494.
Sniderman, Paul M., and Douglas B. Grob. 1996. Innovations in experimental design in attitude surveys. *Annual Review of Sociology* 22:377–399.
Tomz, Michael, and Jessica L. Weeks. 2013. Public opinion and the democratic peace. *American Political Science Review* 107(4):849–865.
Weiss, Jessica, and Allan Dafoe. 2018. Authoritarian audiences and government rhetoric in international crises: Evidence from China. http://www.allandafoe.com/china.