Abstraction and Detail in Experimental Design

March 18, 2021

Ryan Brutger[1], Joshua D. Kertzer[2], Jonathan Renshon[3], Dustin Tingley[4] & Chagai M. Weiss[5]

ABSTRACT: Political scientists designing survey experiments often face the question of how abstract or detailed their experimental stimuli should be. Typically, this question is framed in terms of tradeoffs relating to experimental control and generalizability: the more context we introduce into our studies, the less control we have, and the more difficulty generalizing the results. Yet we have reasons to question this framing of the tradeoff, and there is relatively little systematic evidence experimenters can rely on when calibrating the degree of abstraction in their studies. We make two contributions. First, we provide a theoretical framework which identifies and considers the consequences of three dimensions of abstraction in experimental design: situational hypotheticality, actor identity, and contextual detail. Second, we replicate a range of classic vignette-based survey experiments from political science, varying these levels of abstraction. Our results suggest, that apart from a specific set of conditions, there are fewer tradeoffs between abstraction and detail in survey experiment design than political scientists often assume.

Word count: 9649 words

ACKNOWLEDGMENTS: This paper benefited from helpful feedback from audiences at MIT's Political Experiments Research Lab in 2014, MPSA 2019, APSA 2020 and the 2020 NYU Experiments Conference, Adam Berinsky, Adam Seth Levine, Jonathan Mummolo, Rich Nielsen, Anne Sartori, Jonathan Woon, and Teppei Yamamoto. Authors listed in alphabetical order.

[1]Assistant Professor, University of California, Berkley, Department of Political Science, Email: brutger@berkeley.edu. Web: https://sites.google.com/berkeley.edu/brutger/.

[2]Professor of Government, Department of Government, Harvard University. Email: jkertzer@gov.harvard.edu. Web: http:/people.fas.harvard.edu/~jkertzer/

[3]Associate Professor & Glenn B. and Cleone Orr Hawkins Chair, Department of Political Science, University of Wisconsin-Madison. Email: renshon@wisc.edu. Web: http://jonathanrenshon.net

[4]Professor, Department of Government, Harvard University. Email: dtingley@gov.harvard.edu. Web: https://scholar.harvard.edu/dtingley

[5]PhD Candidate, Department of Political Science, University of Wisconsin-Madison, Email cmweiss3@wisc.edu, Web: http://chagaimweiss.com

Survey experimentalists in political science often face a question about how abstract or detailed their experimental stimuli should be. This question is typically thought of in terms of tradeoffs between experimental control and generalizability. Some researchers prefer highly stylized experiments that are deliberately light on context, even though this comes at the expense of ecological validity and mundane realism (Morton and Williams, 2010, 313-14). While particularly popular in behavioral experiments seeking to test the predictions of formal models (e.g., Dickson, 2009; Dawes, Loewen and Fowler, 2011; Tingley and Walter, 2011; Kanthak and Woon, 2015; LeVeck and Narang, 2017), this tradition also arises in survey experiments as well (e.g., Renshon, 2015; Mutz and Kim, 2017).

Others prefer the use of rich and detailed vignette-based experiments (e.g., Rousseau and Garcia-Retamero, 2007; Brooks and Valentino, 2011; Druckman, Peterson and Slothuus, 2013; Teele, Kalla and Rosenbluth, 2018; Reeves and Rogowski, 2018). Rich and detailed stimuli are in some ways a response to the "major problem in public opinion and survey research": the "ambiguity that often arises when survey respondents are asked to make decisions and judgments from rather abstract and limited information" (Alexander and Becker, 1978, 103). The ability to generalize experimental findings to other contexts, and the degree to which an experiment triggers the psychological process that would occur in the "real world", are both thought to rise in proportion to the level of "realism" in a given vignette (Aguinis and Bradley, 2014, 361). Similarly, others argue that "concrete, realistic context" results in more "reliable assessments" of the dependent variables we care about (Steiner, Atzmüller and Su, 2016, 53).

Political scientists seeking to navigate these tradeoffs are usually exposed to one or the other of these schools of thought regarding experimental design, but have relatively little systematic evidence about how to choose between them. Some scholars advise that respondents perform better in more concrete and familiar settings (Reiley, 2015), while others worry that detail reduces experimental control (Camerer, 1997). Decisions regarding abstraction and detail are particularly important for the design of survey experiments because of their emphasis on vignettes (Gaines, Kuklinski and Quirk, 2007), but also arise in almost any experiment where researchers present respondents with information, whether in the lab (Renshon, 2015) or in the field (Karpowitz, Monson and Preece, 2017).

And yet, as a discipline we know relatively little about the tradeoffs inherent in abstract versus concrete experimental designs. Certainly, increasing "color in the laboratory" *may* trigger "un-

known (to the experimenter) impressions and memories of past experiences over which the experimenter has no control" (Friedman, Friedman and Sunder, 1994), but it is not obvious why sparse experiments would fare better in this respect. In fact, a review of the broader experimental literature suggests strong disagreement on which would be a bigger problem in terms of respondents "filling in the blanks": rich, detailed experiments (e.g., Friedman, Friedman and Sunder, 1994) or abstract, sparse studies (e.g., Alekseev, Charness and Gneezy, 2017). While others have noted that there is no "general theory that would give experimentalists guidance as to when stylization" might pose problems (Dickson, 2011, 61), and that this is "ultimately, an empirical issue that would have to be thrashed out by comparing data from abstract as well as contextually rich experiments" (Friedman, Friedman and Sunder, 1994, 53-4), there is surprisingly little systematic work that does so, forcing experimentalists in political science to rely on hunches and intuitions rather than systematic evidence and theoretical guidance.

In this article, we seek to make both a theoretical and an empirical contribution. First, we offer an overarching conceptual framework outlining three dimensions of abstraction implicated in survey experimental design: *situational hypotheticality*, *actor identity*, and *contextual detail*. We argue that there are certain types of questions where ethical or feasibility considerations mandate at least some form of hypotheticality or abstraction, while there are others where scholars have more leeway. Yet, in those cases where scholars do have leeway, we argue that the tradeoffs between abstraction and detail in experimental design are not as stark as political scientists often claim. Second, like other recent work seeking to subject conventional wisdom about experimental design principles to empirical scrutiny (Mullinix et al., 2015; Coppock, 2019; Mummolo and Peterson, 2019; Kertzer, 2020), we test our theoretical framework, replicating three well-known survey experiments in political science, and manipulating their levels of abstraction in three different ways. We find no evidence that situational hypotheticality changes the results experimenters obtain, an important finding as Institutional Review Boards (IRBs) and our field more broadly become increasingly concerned about the use of deception. Whether with politicians in American politics experiments, or countries in International Relations experiments, we generally find little evidence that varying the abstraction of actor identities changes experimental results, although cue-taking experiments that use real and highly salient cuegivers obtain stronger effects than those that use fake ones. And, we show that adding contextual detail to experimental vignettes attenuates the size of treatment effects because respondents are less likely to be able to recall the treatment. Altogether our results suggest that

there are fewer tradeoffs between abstraction and detail in survey experiment design than political scientists often assume.

1 *Abstraction and detail*

One of the many design choices political scientists face when designing survey experiments concerns the appropriate level of *abstraction* in their stimuli. There is a rich literature on abstraction in philosophy, psychology, and cognitive science, which often operationalizes abstraction in slightly different ways (e.g., Paivio, 1990; Colburn and Shute, 2007). For our purposes, we borrow from construal level theory in defining abstraction as a higher-level representation (Sartori, 1970, 1040-46; Trope and Liberman, 2003). It involves making "a distinction between primary, defining features, which are relatively stable and invariant, and secondary features, which may change with changes in context and hence are omitted from the higher-level representation" (Shapira et al., 2012, 231). As Table 1 shows, an abstract representation is sparse and decontextualized, reduced to the object's most central elements (e.g., "A nuclear weapon"), whereas a concrete representation is contextualized and rich in specific detail, including subordinate considerations (e.g., "North Korea's Hwasong-14 intercontinental ballistic missile").

Table 1: Conceptualizing abstraction

Abstract	**Concrete**
High level of construal	Low level of construal
Decontextualized	Contextualized
Primary features	Secondary features
Sparse	Complex
Superordinate elements	Subordinate elements

Modified from Trope and Liberman (2003, 405).

In specifying which elements of a construct are primary and which are secondary, the act of abstraction is inherently a theoretical phenomenon, rather than an empirical one. Although our interest in this article is in abstraction in experimental methods, similar questions also arise in formal modeling, and in quantitative and qualitative methods alike (Clarke and Primo, 2012). In addition, questions about the appropriate level of abstraction loom large in a variety of issues in experimental design: whether experiments should be "stylized" or "contextually rich" (Dickson, 2011; Kreps and Roblin, 2019), use real or hypothetical actors (McDonald, 2019; Nielson, Hyde and Kelley,

2019), and refer to imminent, future, or hypothetical situations. In this sense, experiments can be abstract or concrete along multiple dimensions at the same time. In the discussion below, we suggest that abstraction in experimental design can be conceptualized along at least three dimensions: situational hypotheticality, actor identity, and contextual detail.[1]

1.1 Situational hypotheticality

The first type of abstraction in experimental design concerns whether a scenario is described as hypothetical or not. The rationale for using hypothetical scenarios in survey experiments is simple: in their most stylized form, experimentalists make causal inferences by drawing comparisons between two different states of the world, randomly assigning participants to either a treatment condition, or control. Some experiments intervene by giving respondents in the treatment condition information about the world that they might not otherwise have (e.g., Butler, Nickerson et al., 2011; Raffler, 2019), but especially in survey experiments, experimentalists often manipulate features of the world itself. In order to manipulate features of the world in this manner, experimentalists must either engage in deception (showing respondents mock news articles purported to be real, e.g., Brader, Valentino and Suhay, 2008; Arceneaux, 2012), or find another way to justify—whether to respondents, or to Institutional Review Boards (IRBs)—why the scenario being described to respondents deviates from the one they are in.

One technique employed for this purpose is to explicitly describe the scenario as hypothetical: respondents in Boettcher (2004, 344), for example, are asked to "envision a hypothetical presidency apart from the current administration." Others implicitly invoke hypotheticality: respondents participating in conjoint experiments studying immigration preferences, for example (e.g., Hainmueller and Hopkins, 2015), are presumably not under the illusion that the immigrants they are being asked to choose between are real. Another widely used variant under the category of "implicit hypotheticality" is to describe a scenario as set in the future (e.g., Mattes and Weeks, 2019). This is often termed a *prospective* scenario, but ultimately the future setting is simply a mechanism to make the scenario implicitly hypothetical.

The rationale for these design choices is often not explicitly stated, but usually involves concerns that respondents will not take studies as seriously when scenarios are presented as explic-

[1] These three strike us as the most important dimensions to confront experimentalists designing their studies, but the list is not necessarily exhaustive.

itly hypothetical — the sense that researchers asking hypothetical questions will be rewarded with hypothetical answers (Converse and Presser, 1986, 23). Experimentalists operating out of an economics-style tradition tend to avoid both deception and situational hypotheticality in order to accentuate the effects of incentives (Morton and Williams, 2010). Yet, there is relatively little empirical work testing the conditions in which situational hypotheticality affects responses in political science experiments.

1.2 Actor identity

The second dimension of abstraction involves the identity of the actors invoked in experimental vignettes: are they real, or artificial? Some experimenters explicitly use real world actors in contexts ripped from the headlines, as in Boettcher and Cobb's (2006) study of how casualty frames shape support for the war in Iraq, or Evers, Fisher and Schaaf (2019), who experimentally investigate audience costs using Donald Trump and Barack Obama. In this sense, the artificiality of the actors in an experiment is distinct from the hypotheticality of the situations in which actors are embedded and, indeed, experimenters often use real world actors in hypothetical scenarios. For example, Kriner and Shen's (2014) casualty sensitivity experiments explore how many casualties Americans would be willing to bear in a series of "hypothetical" interventions in "real" countries. In this case, the military interventions are artificial and prospective, while the relevant target countries are real.

Moving up the ladder of abstraction, some experimenters describe hypothetical scenarios in artificial countries, in order to exert complete control over how much information participants bring to bear. For example, Brooks and Valentino (2011) describe a conflict between "Malaguay and Westria", and Rubenzer and Redd (2010) describe a crisis in the state of "Gorendy." Taking this approach a step forward, many experimentalists use unnamed countries, describing target states as "Country A" or "Country B" (Johns and Davies, 2012; Yarhi-Milo, Kertzer and Renshon, 2018), or simply referring to "A country" rather than providing a label (Tomz and Weeks, 2013).

Concerns about actor identity and hypotheticality are not limited to the subfield of international relations. In comparative politics, Banerjee et al. (2014) describe hypothetical representatives (running for office in hypothetical districts) to study the concerns of voters in rural India. "Hypothetical candidate" experiments are also a long-running feature in the study of American politics (as in Colleau et al., 1990; Kam and Zechmeister, 2013) — and are particularly common in conjoint experiments. In a meta-analysis of 111 studies of negative campaigning, Lau, Sigelman and Rovner

(2007) find that experiments featuring hypothetical candidates don't offer significantly different results from those featuring real ones. McDonald (2019), in contrast, argues that experiments on hypothetical candidates both increase cognitive burden and produce larger treatment effects than experiments on candidates about which respondents have strong priors.

As with the case of situational hypotheticality, the logic of using unnamed or hypothetical actors stems directly from the questions being tested. Political scientists turned to experimental methods to study the effects of candidate gender (Brooks and Valentino, 2011), for example, precisely because it is difficult to find two real-world candidates identical to one another on all dimensions other than their gender. The same is true in studies of race in politics (Burge, Wamble and Cuomo, 2020), or ethnicity (Dunning and Harrison, 2010). In an IR context, it is hard to think of two real-world countries that are identical in all respects but one, such that IR scholars interested in manipulating the effects of regime type, military capabilities, or foreign policy interests usually do so with fictional or hypothetical countries (e.g., Rousseau and Garcia-Retamero, 2007).

1.3 Contextual Detail

The third dimension of abstraction involves the amount of additional context provided in an experiment. Press, Sagan and Valentino (2013) present a lengthy newspaper article that provides participants with a large amount of context, as do experiments in American politics that generate fake campaign advertisements or news clips (Brader, Valentino and Suhay, 2008). In contrast, other experiments often present relatively little information (Tingley and Walter, 2011; Kanthak and Woon, 2015). This modeling decision is not limited to economics-style bargaining games: Trager and Vavreck (2011), for example, manipulate the President's strategy in a foreign policy crisis as well as information about the US domestic political environment, but as with most audience cost experiments, they say relatively little about the context of the intervention itself.

The argument usually offered in favor of contextual detail is that it increases realism and respondent engagement. Yet, apart from Kreps and Roblin (2019) and Bansak et al. (2020), there has been little empirical work adjudicating what the consequences of providing richer or sparser stimuli might be. Bansak et al. (2020) use a clever multi-stage conjoint design to first find "filler attributes" (information uncorrelated with the object of interest in the study) and then experimentally vary the amount of filler in the second stage, finding relatively stable treatment effects even with large numbers of filler items. Kreps and Roblin (2019) focus on treatment "formats" (such as

comparing short vignettes with longer mock news stories), finding that respondent attention (as a measure of satisficing) was unaffected by the presentational format.

This discussion suggests that what is often referred to as "contextual detail" is actually composed of at least three related dimensions. The first is simply the volume of information provided: more or less information can be provided in an experiment to add "realism." The second concerns *how* the information is presented, and here there have been examples of any number of treatment formats in experiments, from bullet-pointed vignettes (Tomz, 2007), to mock news reports (Druckman and Nelson, 2003; Valentino, Neuner and Vandenbroek, 2018). The third is the content of the information itself, which is orthogonal to its volume. Any bit of information may be classified as either what Bansak et al. (2020) call "filler" or its opposite, what we call "charged" content, which may interact with the treatment in some way and affect the results of a study through a mechanism other than simple respondent satisficing. If a President's "favorite highway" is filler, then Bansak et al. (2020) also show that other attributes (e.g., previous occupation and number of children) are associated with the object of interest and are thus ill-suited to be added simply to increase the "realism" of a vignette. But while they show that satisficing is less of a problem than we might expect once we introduce filler attributes, we are still largely in the dark with respect to understanding how the addition of charged (versus filler) content affects our interpretation of experimental results.

2 *Navigating the tradeoffs*

In sum, although political scientists tend to recognize that tradeoffs between abstract and concrete survey experiments exist, there is less certainty about how one should balance them. One method has been to run both abstract and concrete versions of an experiment to test whether the results hold (e.g., Herrmann, Tetlock and Visser, 1999; Rousseau and Garcia-Retamero, 2007; Berinsky, 2009; Renshon, Dafoe and Huth, 2018; Nielson, Hyde and Kelley, 2019), though this is less than ideal since adjusting levels of abstraction on multiple dimensions simultaneously provides limited insight regarding the specific dimension driving experimental outcomes.

There are some circumstances where for logistical or ethical reasons, experimenters will be constrained in terms of how abstract or concrete their stimuli will be. For example, researchers are limited in their ability to select real world actors when studying the effects of race and gender in

candidate selection, or the effects of country-level characteristics on foreign policy preferences. Additionally, there are experiments where some form of situational hypotheticality is required (often at the demand of IRBs) to avoid the use of deception, and some contexts where the use of deception raises ethical challenges: for example, telling respondents that a real-world political candidate is unethical (e.g., Butler and Powell, 2014).

In other cases, however, survey experimentalists have more of a choice when designing their studies. In the discussion below, we link each dimension of abstraction to questions about experimental control, on the one hand, and generalizability, on the other. Although political scientists often see these two principles as in tension with one another—associating the former with internal validity, and the latter with external validity—we argue that the implications of abstraction in experimental design for each principle are actually more complex. There are some instances where an increase in abstraction may enhance experimental control, and others where an increase in abstraction may come at the expense of experimental control; because experimentalists may not exercise as much control over their respondents as we like to think, more abstract stimuli may not necessarily be more generalizable. We suggest, then, that the tradeoff between abstract and concrete experimental designs represents something of a paradox: the circumstances in which experimentalists have the most leeway in terms of the abstraction of design choices may be the ones where the tradeoffs between different design choices are the least consequential.

2.1 Experimental Control

Experimenters seek to obtain "control" over the ways in which respondents construe the contextual features of vignettes, in order to ensure proper implementation of their experimental designs. When experimental vignettes provoke different reactions amongst different types of respondents—perhaps reactions the researcher never intended—experimenters can risk losing control over their study, raising concerns regarding internal validity. By varying the information provided along the three aforementioned levels of abstraction, experimenters can potentially shape the degree of control they obtain.

Yet, we argue that there are less to these tradeoffs than meets the eye. First, the relationship between abstraction and control varies based upon the dimension under investigation. Increasing contextual detail is often thought to enhance experimental control, by fixing the type and degree of information that all subjects share regarding an issue area. For example, when implementing

an endorsement experiment regarding a (fictional or real) immigration policy (Nicholson, 2012), researchers can provide detailed information regarding: who initiated the policy, when it comes into effect, and how it relates to previous policies. Presumably, these additions can ensure an informational common denominator, and avoid a situation in which respondents with different background knowledge construe the experimental vignette in diverging ways.

In contrast, increased detail in terms of actor identity is usually argued to reduce experimental control. In an international relations context, Herrmann, Tetlock and Visser (1999, 556) note that "the use of real countries [adds] a degree of realism…but it also sacrifice[s] a degree of experimental control. Affective reactions to the various countries may differ, and [characteristics of the countries] may not be perceived uniformly by all participants." In American politics, Reeves and Rogowski (2018, 428) write that "the use of hypothetical candidates comes at the cost of reducing the real-world attributes of the experiment, but this cost is offset by removing respondents from their feelings about any actual politician, which could serve as confounders." These examples suggest that by introducing real world actors and adding detail into vignettes, experimenters lose control over their respondents — the opposite of conventional wisdom about the effects of contextual detail.

More generally, it may be somewhat misleading to think that by turning from real to hypothetical actors, or from contextually sparse to rich vignettes, experimenters gain control over their study. Indeed, when presented with relatively pared down stimuli, participants often "fill in the blanks." For example, scenarios in which "a country sent its military to take over a neighboring country" in which the US is considering sending troops to repel the invader (e.g. Tomz, 2007) may lead some participants to think of the Gulf War. It is also possible that different respondents will exert diverging reactions to additional contextual detail, causing experimenters to lose, rather than gain control. Adopting an abstract design can thus both increase or decrease experimental control, such that the tradeoff here may not be as clean cut as experimentalists sometimes suggest.

Even if experimenters may have more leeway when choosing the appropriate level of abstraction for actor identity than is often claimed, this does not mean that all concrete actor identities are equally desirable. In particular, experimenters should attend to at least two considerations when choosing real world actors. The first is *schema consistency* (Hashtroudi et al., 1984): is the choice of actor reasonable given the scenario in which the actor is embedded? For example, in experimental scenarios in which a country is pursuing a nuclear weapons program (e.g., Tomz and

Weeks, 2013), experimental control decreases if the experimenter chooses a country that already has nuclear weapons (e.g., Russia), or a country that respondents think is unlikely to pursue them (e.g., Canada). If a schema-inconsistent actor is chosen, the respondent is less likely to believe the scenario or accept the treatment, thus weakening the treatment effect. The second is *treatment consistency*: if the treatment manipulates an attribute of an actor, are all of the levels of the attribute being manipulated seen as plausible by respondents? In candidate selection experiments, for example, it would be difficult to manipulate the policy stances of politicians on issues where they have already taken prominent positions. If respondents do not perceive the treatment as consistent with the identity of the actor, then the experimenter is likely to lose control since the respondent may not comply with the treatment, attenuating the treatment effect.

2.2 Generalizability

While experimental control is a fundamental aspect in designing vignettes, scholars may very well be concerned by other factors such as generalizability – the extent to which results from a given study speak to a broader set of real world scenarios. Like control, degrees of generalizability may be shaped by levels of abstraction in experimental design. Thus when framing an experiment as hypothetical or real, and when selecting particular actors, and levels of contextual detail, researchers may condition the degree to which their results generalize beyond a particular context.

Oftentimes, experimenters adopt unnamed actors in experimental vignettes in order to enhance generalizability. At least implicitly, the selection of an unnamed actor is motivated by the assumption that a researcher's quantity of interest is a main rather than a conditional effect: for example, when a researcher is interested in the effect of past behavior on forming reputations for resolve in general, not the effect of past behavior on forming reputations for resolve for Iran specifically (Renshon, Dafoe and Huth, 2018).

Yet, it is unclear that increased abstraction actually increases generalizability. First, when we generalize from these experiments to the problems in the real world that motivate us to conduct them in the first place, selecting unnamed actors may lead us to miss important sources of treatment heterogeneity, and may even make it harder to generalize results to any motivating real world cases. For example, because respondents are often "pre-treated" with partisan cues prior to participating in our studies (Gaines, Kuklinski and Quirk, 2007), experimenters might deliberately choose nonpartisan scenarios where these pretreatment effects are minimized, lest the effects of

partisanship swamp or overwhelm the treatments of interest. Yet if many political phenomena have a partisan hue, the absence of partisan dynamics in the experiment actually makes it harder to generalize these results (McDonald, 2019).

Similarly, the degree of contextual detail provided by experimenters might shape the extent that findings from an experiment can generalize to real world scenarios. If participants in experiments only receive two pieces of information, one of which is the treatment being randomly assigned, the relative "dosage" of the treatment is likely to be unrealistically high, and may not hold in a more naturalistic setting (Barabas and Jerit, 2010). In contrast, if the treatment is presented to participants embedded in a larger amount of information the treatment is likely to exert a (realistically) smaller effect.

In sum, although experimentalists frequently think about questions regarding experimental control and generalizability as two competing principles, the latter linked to abstract designs, and the former to concrete ones, it is not clear that the tradeoffs are actually as stark: adding contextual detail can increase control, but choosing real-world actors may lower it. We seek to evaluate these conjectures empirically. Specifically, by experimentally manipulating the situational hypotheticality, actor identity, and contextual detail of a series of popular survey experiments, we aim to determine if and how different forms of abstraction shape the results obtained. If introducing real actors or elaborate contextual detail systematically affects experimental control and generalizability, then one would expect to observe variation in outcomes across experiments varying in abstraction. If, however, the amount and type of detail across experiments only modestly affects the results, it would suggest that the tradeoffs between these design choices are somewhat overstated.

3 Research Design

To provide guidance for experimentalists on how abstract their survey experiment ought to be as well as how scholars should balance the potential tradeoffs associated with differing levels of abstraction, we fielded a series of survey experiments, each designed to address one of the dimensions of abstraction described earlier. Our study selection criteria sought to replicate studies that i) focused on core theoretical debates in political science, ii) had simple designs (so that we would be sufficiently powered to detect moderation effects), iii) uncovered a large and substantively meaningful effect, and iv) which were conductive to manipulating situational hypotheticality, actor iden-

tity, and contextual detail.

We focus on three experiments (depicted in Table 2), each of which features three levels of treatment: (1) the central treatments from the replicated studies, (2) contextual detail and actor identity treatments varying the amount of context or the names of the actors respondents are presented with, and (3) a situational hypotheticality treatment which describes experimental scenarios as either real, explicitly hypothetical, or implicitly hypothetical. An additional summary of the structure of our survey instrument is depicted in Appendix §1 and the details of each replication are contained in Appendix §2.

Our first study, the ELITE CUES experiment, replicates Nicholson (2012), which compares support for immigration policy amongst respondents receiving an in-party (or out-party) politician endorsement. In our replication, we updated the relevant salient cuegivers (Joe Biden or Donald Trump) and the substantive context of the experiment—protection for "Dreamers" in the U.S.—while adding actor identity treatments that vary whether the immigration reform endorsement is made by less salient partisan cuegivers (Senator Tom Carper of Delaware or Senator Mike Rounds of South Dakota), or by a fictional politician (Stephen Smith) whose partisanship we manipulate. This experiment therefore lets us explore the effects of varying actor identity in experimental design.

Our second study, the INGROUP FAVORITISM experiment, replicates Mutz and Kim (2017), which tests how manipulating the expected relative gains in a trade deal shapes public support. We use this study to explore the effects of additional contextual detail, randomly assigning respondents to either the original short vignette, or a more elaborate vignette which provides additional detail. Consistent with Bansak et al. (2020), we provide two types of additional context: "filler" context—peripheral information that increases the volume of text, but is not expected to interact with the treatment—and "charged" context that similarly increases the length of the stimulus, but which is more relevant to the treatment.

Our final study, the NUCLEAR WEAPONS experiment, replicates Press, Sagan and Valentino (2013), which tests how manipulating the relative effectiveness of nuclear weapons affects public support for nuclear attacks. We use this study to explore the effects of both the effects of contextual detail and actor identity, adding two additional treatment arms. First, we manipulate the vignette's context to either include: (1) Elaborate context (as in the original study) or (2) Reduced context. Second, we manipulate the identity of the actor in the dispute: (1) Syria (as in the original study),

(2) An unnamed country ("a foreign country"), (3) A fictitious country name ("Malaguay"), or (4) A real and schema-inconsistent country (Bolivia).[2] Following the main outcome variable for all three experiments, respondents were asked to complete a thought listing exercise and a factual manipulation check. These questions enable us to investigate *why* decisions about how abstract the stimuli are might moderate (or fail to moderate) treatment effects.

Throughout all of the studies we introduce a situational hypotheticality treatment (randomized at the subject-, not the study level) which refers to the depicted scenarios as either real, explicitly or implicitly hypothetical in order to test whether manipulating hypotheticality moderates the experimental findings.[3] The structure of the studies are depicted in Table 2. The IN-GROUP FAVORITISM and NUCLEAR WEAPONS experiments were fielded on a sample of $N = 4686$ respondents through Dynata in spring 2019. The ELITE CUES experiment was fielded on a sample of $N = 4070$ respondents through Lucid's "Theorem" respondent pool in spring 2020.[4] In Appendix §3, we report results of power simulations demonstrating that we are well powered to identify our quantities of interest.

	Elite Cues	**In-Group Favoritism**	**Nuclear Weapons**
Treatments from original study	1. No Endorsement 2. In-Party Cue 3. Out-Party Cue	1. US gains 1000 and other country gains 10 2. US gains 10 and other country gains 1000 3. US gains 10 and other country loses 1000	1. 45% Success for conventional attack 2. 90% Success for conventional attack
Actor identity and contextual detail treatments	If assigned to cue: 1. Real + High Salience (Donald Trump/Joe Biden) 2. Real + Low Salience (Mike Rounds/Tom Carper) 3. Fictional (Stephen Smith/Stephen Smith)	1. No Additional Context (original) 2. Filler Context 3. Charged Context	1. Extended Context (original) 2. Reduced context 1. Unnamed (foreign country) 2. Made up (Malaguay) 3. Real + Schema consistent (Syria) 4. Real + Schema inconsistent (Bolivia)
Situational hypotheticality treatment	Situation described as: 1. Implicitly hypothetical 2. Explicitly hypothetical 3. Real	Situation described as: 1. Implicitly hypothetical 2. Explicitly hypothetical	Situation described as: 1. Implicitly hypothetical 2. Explicitly hypothetical

Table 2: Summary of Treatments for 3 Studies

[2]The extent to which real countries are schema-consistent with a given experimental scenario is an empirical question. Appendix §2 describes a pilot study we fielded in order to rate the consistency of 11 possible countries with the behavior described in the vignette.

[3]In our first survey respondents were assigned to one of two conditions describing a situation as either implicitly or explicitly hypothetical. In our second survey respondents were assigned to one of three conditions describing a situation as either real, implicitly, or explicitly hypothetical.

[4]More details about each platform are available in Appendix §1.

4 *Results*

4.1 Replication of Original Study Results

In Figure 1 we present our initial replication of the three studies under investigation along with estimates from the original studies.[5] As expected, our In-group Favoritism study shows that respondents are more likely to support trade deals in which the U.S. gains more than a rival country. Our Elite Cues study demonstrates that respondents are less likely to support an immigration policy endorsed by an out-party politician. Finally, our Nuclear Weapons study suggests that respondents are more likely to support the use of nuclear weapons when they are described as more effective than conventional weapons. Taken together, the results in Figure 1 demonstrate our initial success in replicating our studies of interest. More important is how our additional treatments moderate the main results depicted above.

4.2 Situational Hypotheticality Effects

Does describing an experimental scenario as explicitly hypothetical, prospective, or real affect the results obtained in experimental designs? To answer this question, we administered our situational hypotheticality treatment which assigned respondents to introductions describing each experimental vignette as follows: in the Nuclear Weapons and In-group favoritism studies, we described experimental vignettes as either explicitly hypothetical or prospective—and thus implicitly hypothetical— while in our Elite cues experiment, we introduced experimental vignettes as either hypothetical, real, or without addressing hypothetically at all. Throughout all our studies, subjects were randomly assigned to a hypotheticality condition at the beginning of the survey instrument, which was held constant for all the studies that followed.

To examine the effect of this design choice, we use standard OLS models in which we interact the original treatment from a given study—e.g., in the elite cues experiment, whether an out-party politician is the endorser of the immigration reform policy—with our hypotheticality treatment. Figure 2 presents results in which our main quantity of interest is the interaction effect, representing the moderating effect of our hypotheticality treatment on the original treatments. In our elite cues replication, hypotheticality can take one of three values (explicitly hypothetical, im-

[5] We do not include the original data estimate for Mutz and Kim because the original study included a more complex design with the potential for each country to gain or lose 1, 10, 100, and 1000 jobs, in contrast to our simplified version.

Figure 1: Replication of ATEs from the three experiments

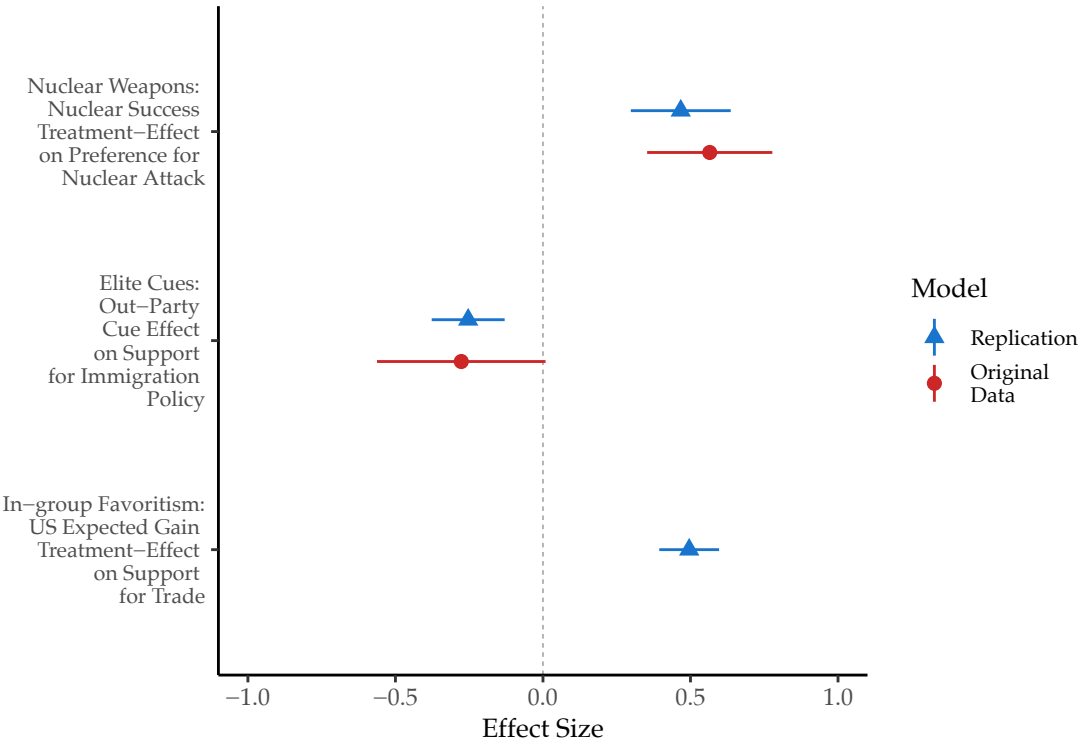

Figure 1 shows we successfully replicate the average treatment effects from the original studies. Point estimates and confidence interval are extracted from separate OLS models where original outcomes are predicted by treatments. When possible we benchmark our replication (Blue) to original studies (Red). We only analyze data from respondents exposed to the original format of the experiment, omitting respondents exposed to new variants of the experiment where we introduce diverging elements of abstraction or detail. All outcomes are standardized.

plicitly hypothetical, or real). However, we focus on comparing the real and explicitly hypothetical conditions, which are most distinct.[6]

As evident in Figure 2, framing an experimental vignette as explicitly hypothetical does not change the main findings from experimental studies. In all models, our situational hypotheticality treatment, and its interaction with original treatments are statistically and substantively insignificant. We interpret these results as evidence for the limited empirical consequences of design choices relating to situational hypotheticality.

Figure 2: No moderating effects of situational hypotheticality

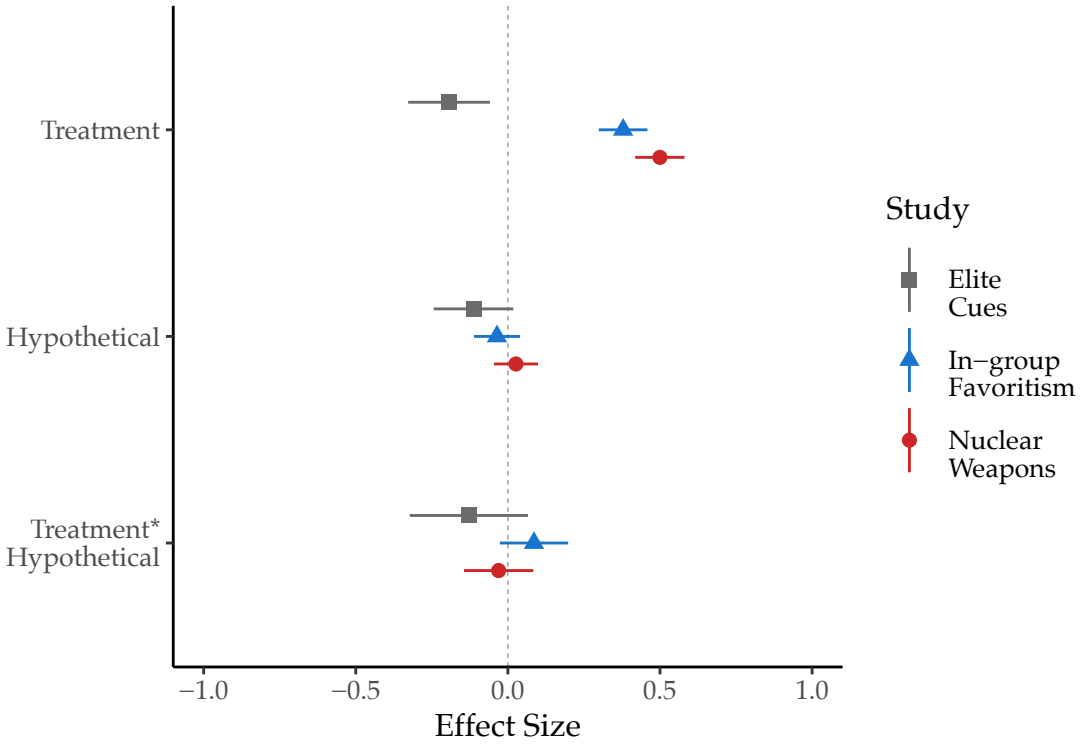

Figure 2 finds no evidence that situational hypotheticality significantly moderates our treatment effects in any of the three experiments. Point estimates and confidence intervals are extracted from three separate OLS models where original outcomes are predicted by original treatments interacted with the hypotheticality treatment. All outcomes are standardized.

[6]Comparisons of the explicit and implicit hypothetical conditions yield similar results.

Figure 3: Moderating effects of actor identity condition

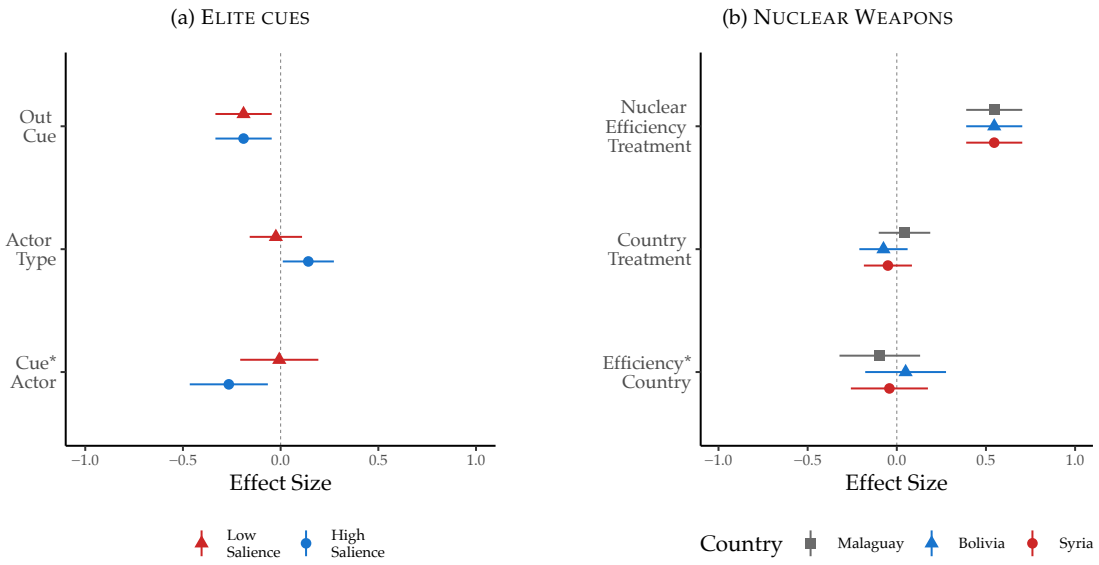

Figure 3 shows that manipulating country identity does not significantly moderate treatment effects in the NUCLEAR WEAPONS experiment, although we obtain slightly larger treatment effects in the ELITE CUES study when we use more salient cue-givers. Point estimates and confidence intervals are extracted from five separate OLS models where original outcomes are predicted by original treatments interacted with different actor identity conditions. Panel *a* compares made-up politicians with low salience (red) and high salience (blue) politicians. Panel *b* compares the unnamed country with a fake country (gray), schema inconsistent country (blue), and schema consistent country (red). All outcomes are standardized.

4.3 ACTOR IDENTITY EFFECTS

We now turn to an analysis of how actor identities of different levels of abstraction affect findings from experimental vignettes. In our NUCLEAR WEAPONS study, we randomized the target country as: unnamed (our baseline condition), fictional (Malaguy), real and schema inconsistent (Bolivia), or real and schema consistent (Syria). Similarly, in the ELITE CUES replication we randomized whether an out-party endorsement was by a made-up politician (Stephen Smith [D or R], our pooled baseline condition), a low salience politician (Senators Mike Rounds [R] or Tom Carper [D]), or a high salience politician (Donald Trump [R] or Joe Biden [D]).

We interact this actor identity treatment with each study's original treatment, and present results for both our ELITE CUES and NUCLEAR WEAPONS replications in Figure 3 (Panel *a* and *b* respectively). In these figures, our main quantity of interest is the interaction between the original treatment and our additional actor identity treatment.

As demonstrated in Figure 3, most actor identity conditions do not moderate the main treatment effects. For the most part, whether an actor is unnamed, fictional or real—and if real, schema-consistent or inconsistent—does not lead scholars to draw substantively different inferences or identify diverging effects, either in magnitude or direction. That said, in the left panel of Figure 3, we show that using high salience actors amplifies the endorsement treatment effects (when compared to baseline made-up actors).

There are at least three potential mechanisms to explain the actor identity results from the elite cue experiment. The first is cognitive burden. McDonald (2019) proposes a version of this hypothesis, arguing that survey experiments using hypothetical actors increase the cognitive burden on respondents, as measured by response latencies in survey questions. Yet as we show in Appendix §4, there is no significant effect of the actor identity treatment on response latency in our study, so it does not appear that moving from a hypothetical to a low or high salience actor alters cognitive burden amongst our respondents. The second potential mechanism is differential treatment recall: that respondents are better able to recall treatments from salient actors than non-salient ones. Yet, as Appendix §4 shows, we find no evidence that treatment recall rates significantly vary with the actor identity treatment. The third interpretation, which we believe is more consistent with our results, has to do with simple Bayesian models of persuasion: endorsement effects are stronger when the endorsement comes from a salient cuegiver because respondents are likely to have stronger priors about the cuegiver.[7]

4.4 Contextual Detail Effects

Lastly, we consider the moderating effects of contextual detail in Figure 4. We administered two versions of our context treatments. In the NUCLEAR WEAPONS experiment, respondents were either exposed to a reduced context vignette (baseline) or an original elaborate context vignette. In the IN-GROUP FAVORITISM experiment, respondents were either exposed to an original minimal context vignette (baseline), or an extended context vignette which included "filler" or "charged" additional context.

[7] In this sense, our findings offer helpful scope conditions for other experimental work arguing that using unnamed or hypothetical actors artificially inflates the size of treatment effects (McDonald, 2019): if the dependent variable involves measuring attitudes about an actor, a simple Bayesian framework would predict that the stronger the respondents' priors, the *less* they should update in response to new information about the actor. However, if the dependent variable involves measuring attitudes about a policy, that same Bayesian framework would predict that the stronger the respondents' priors about the policy's endorser, the *more* they should update in response to information about the cuegiver.

Figure 4: Adding contextual detail attenuates treatment effects

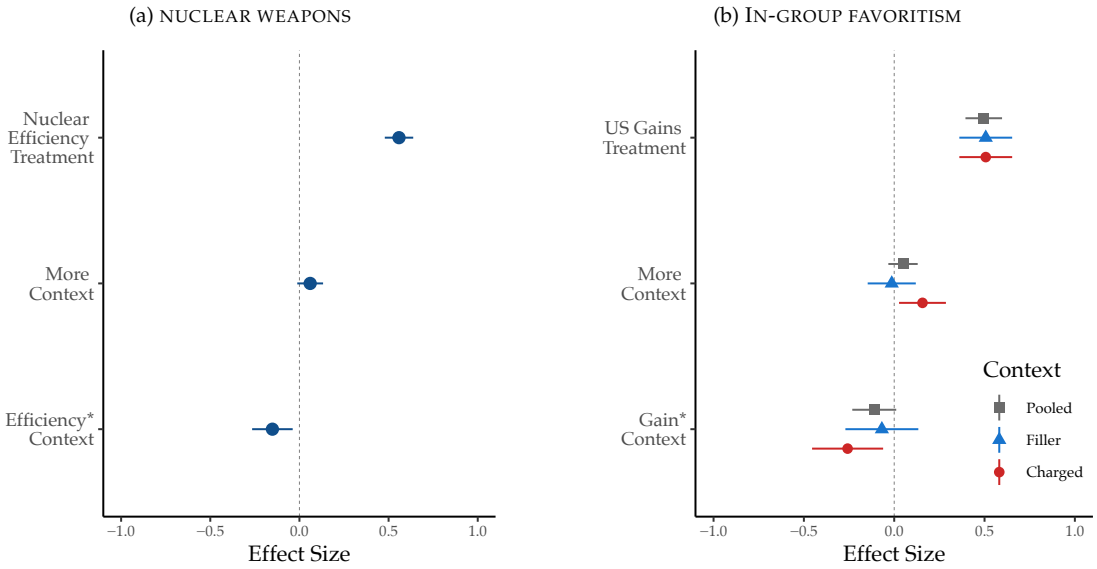

Figure 4 shows that adding contextual detail to studies weakens the treatment effects. Point estimates and confidence intervals are extracted from three separate OLS models where original outcomes are predicted by original treatments interacted with study level context. In panel *a*, a baseline reduced-context condition is compared with the original elaborate-context condition used in the original Press, Sagan, and Valentino. In panel *b*, we compare a baseline reduced context vignette with elaborate context conditions which are either non-innocuous (blue) or innocuous (red). We also consider a pooled model of both types of experimental context (gray). All outcomes are standardized.

As demonstrated in Figure 4(a), exposing respondents to the original rich experimental vignette in the NUCLEAR WEAPONS experiment has a negative moderating effect on the study's main treatment. Put differently, extended experimental vignettes seem to dampen the original treatment (nuclear effectiveness), but this moderating effect does not lead scholars to draw opposite inferences, but rather, just estimate more conservative treatment effects.

Figure 4(b) provides us with further insight into the moderating effects of contextual detail on main treatments. In this panel, we consider the general effect of adding contextual detail to experimental vignettes (grey - pooled model), as well as the particular effects of adding either "filler" or "charged" context. These results further suggest that adding contextual detail to experimental vignettes will dampen treatment effects. Indeed, the moderating effect of extended contextual detail (in relation to a baseline minimal context condition), when pooling together both "filler" and "charged" context conditions, approaches statistical significance ($p < 0.08$). As evident in Figure 4(b) this effect is driven by the "charged" context condition, which in and of itself has a statistically significant impact on the size (but not direction) of main treatment effects.

To better understand why adding contextual detail to experimental vignettes dampens treatment effects, we consider the effects of our contextual detail treatment on treatment recall success. To do so, we regress respondents' recall success of the original study-level treatments (Nuclear attack effectiveness in the NUCLEAR WEAPONS study and expected consequences of trade in the IN-GROUP FAVORITISM study) on respondents' contextual detail condition. Figure 5 demonstrates that increased context in experimental design hinders respondents' ability to successfully recall the treatment condition to which they were assigned.

5 Concluding Thoughts

We began this paper by calling attention to a significant problem faced by political scientists who seek to test their theories using survey experiments: in most cases, they have a wide degree of latitude in how to design the experimental stimuli and must make choices about whether to use real actor names or make them up (or leave them un-named), whether to add rich, contextual detail (and if so, how much, and what kind), how to present the information in the experiment (whether explicitly hypothetical, implicitly hypothetical, or as real), whether to use deception, and so on. In confronting the issues raised by these "design degrees of freedom," scholars have no

Figure 5: Contextual Detail Effects on Treatment Recall Success

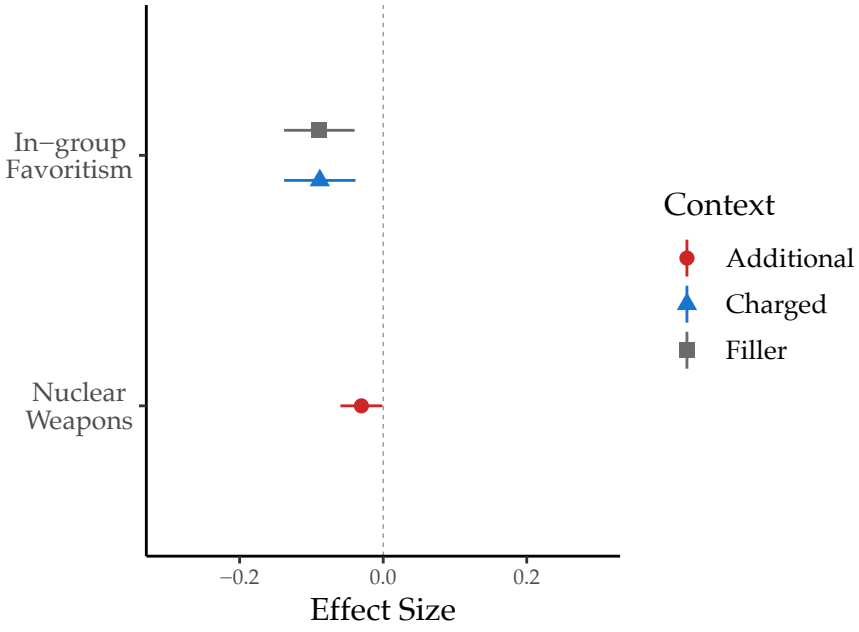

Figure 5 demonstrates how adding contextual detail negatively affects treatment recall. Point estimates and confidence intervals are extracted from three separate OLS models where a binary treatment recall success variable is predicted by the context condition to which respondents were assigned. The NUCLEAR WEAPONS model compares recall rates of respondents assigned to a baseline reduced context conditions, with respondents assigned to extended context condition. IN-GROUP FAVORITISM models, compare respondents assigned to a minimal baseline condition, with respondents assigned to filler or charged conditions. All outcomes are standardized.

shortage of folk wisdom to fall back on from their peers, mentors and textbooks, but the "conventional wisdom" on which they can rely is either nonexistent or contradictory. One thing nearly everyone agrees upon, however, is that—whatever our researcher chooses—they will face a trade-off between experimental control and generalizability. However, despite a recognition that these questions are, ultimately, subject to study and research like many other problems (e.g., Friedman, Friedman and Sunder, 1994), there is little in the way of theoretical frameworks or empirically-minded guidance for researchers who face these issues.

Our contribution is twofold. First, we provided a conceptual framework that helps to make sense of the many choices that experimentalists face in terms of the degree of abstraction or concreteness of their designs. In particular, our framework draws from construal level theory to outline three dimensions of abstraction—situational hypothetically, actor identity and contextual detail. Most importantly, our framework and theoretical discussion of the implications of each of these three dimensions of abstraction for internal and external validity help to clarify a key point: the oft-remarked upon tradeoff between experimental control and generalizability is not nearly as stark as is often assumed and in some cases is not a direct trade-off at all. Abstraction may in some cases enhance, rather than decrease, experimental control, which, in any case, experimentalists have less of than they realize in many cases.

Empirically, we test our theoretical framework through a replication and extension of three well-known vignette-based survey experiments in political science (Nicholson, 2012; Press, Sagan and Valentino, 2013; Mutz and Kim, 2017). To each of these, we add our layers of experimental manipulations to test the implications of abstraction in experimental design. In our ELITE CUES study, we manipulate the actor identity of the politician presented in the vignette; to the IN-GROUP FAVORITISM study's relatively sparse design we add two types of context ("filler" and "charged") and to the NUCLEAR WEAPONS experiment we add manipulations of both context and actor identity. In addition, for all three experiments, we manipulate the degree of situational hypothetically, presenting scenarios as either real, implicitly hypothetical, explicitly hypothetical, or without any mention of hypotheticality.

Our empirical results suggest reasons for optimism. Framing a study as hypothetical or real does not make any substantial difference, failing to affect any of the main findings from the three replicated studies. This suggests that the difficult ethical decisions about whether or not to use deception in order to increase respondent engagement may in many cases be unnecessary, adding

empirical weight to an important normative debate in the field. We examined contextual detail in two ways: adding two types of context in our IN-GROUP FAVORITISM study and and subtracting context from our NUCLEAR WEAPONS experiment. Our results are consistent across both studies: although additional context leads to more conservative estimates of treatment effects, it should not affect the rate of Type 2 errors in well-powered studies, and that context dampens treatment effects by hindering respondents' ability to successfully recall the main treatment. Choosing the appropriate level of contextual detail in experimental work thus depends on the purpose of the study: if the purpose is to demonstrate that an effect exists, a sparser experimental design better enables researchers to identify the mechanism, but if the purpose is instead to understand how important an effect might be relative to other considerations, or whether respondents in a more naturalistic setting would be likely to receive the treatment (Barabas and Jerit, 2010), a more contextually-rich design may be beneficial.

We also investigated the effects of varying the level of abstraction of the actors in the experiments. We manipulated actor identity in the NUCLEAR WEAPONS experiment by exposing respondents to conditions in which the country was either unnamed, fictional, or real and either consistent with the main thrust of the scenario or not. In the elite cues experiment, actor identity was manipulated using made-up, low-salience, or high-salience cue-givers. Across both experiments, which considered different types of actors (i.e. countries or politicians), most actor-related design choices did not matter, in that the interaction between the actor identity treatment and the main treatment was not statistically significant. The sole exception is that more salient politicians make more effective cuegivers than fictional actors do. We also consider the extent to which different dimensions of our framework (contextual detail and actor identity) interact to moderate experimental findings, shown in Appendix §5. We find little support for this notion, further enhancing our intuition that decisions around actor identities do not substantively moderate experimental findings.

In line with other recent work seeking to subject widely held assumptions about experimental methods to empirical scrutiny (Mullinix et al., 2015; Coppock, 2019; Mummolo and Peterson, 2019; Kertzer, 2020), we find limited empirical support to substantiate commonly held concerns regarding the consequences of these design choices for the substantive interpretation of experiments in political science. Our conceptual framework clarifies where, when, and how researchers might have discretion in selecting particular levels of abstraction in their experimental stimuli. However, somewhat ironically, our evidence suggests that in cases where researchers have discretion over de-

sign choices relating to abstraction, their choices bear limited empirical consequences. Our findings do not imply that levels of abstraction never moderate average treatment effects from experiments, but that they seem to do so in a manner that does not impact the substantive interpretation of any given experiment.

References

Adida, Claire L. 2015. "Do African voters favor coethnics? Evidence from a survey experiment in Benin." *Journal of Experimental Political Science* 2(1):1–11.

Aguinis, Herman and Kyle J Bradley. 2014. "Best practice recommendations for designing and implementing experimental vignette methodology studies." *Organizational Research Methods* 17(4):351–371.

Alekseev, Aleksandr, Gary Charness and Uri Gneezy. 2017. "Experimental methods: When and why contextual instructions are important." *Journal of Economic Behavior & Organization* 134:48–59.

Alexander, Cheryl S and Henry Jay Becker. 1978. "The use of vignettes in survey research." *Public opinion quarterly* 42(1):93–104.

Arceneaux, Kevin. 2012. "Cognitive Biases and the Strength of Political Arguments." *American Journal of Political Science* 56(2):271–285.

Banerjee, Abhijit, Donald P Green, Jeffery McManus and Rohini Pande. 2014. "Are poor voters indifferent to whether elected leaders are criminal or corrupt? A vignette experiment in rural India." *Political Communication* 31(3):391–407.

Bansak, Kirk, Jens Hainmueller, Daniel J. Hopkins and Teppei Yamamoto. 2020. "Beyond the breaking point? Survey satisficing in conjoint experiments." *Political Science Research and Methods* Forthcoming:1–19.

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

Berinsky, Adam J. 2009. *In time of war: Understanding American public opinion from World War II to Iraq*. Chicago, IL: University of Chicago Press.

Boettcher, III, William A. 2004. "The prospects for prospect theory: An empirical evaluation of international relations applications of framing and loss aversion." *Political Psychology* 25(3):331–362.

Boettcher III, William A and Michael D Cobb. 2006. "Echoes of Vietnam? Casualty framing and public perceptions of success and failure in Iraq." *Journal of Conflict Resolution* 50(6):831–854.

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

Brooks, Deborah Jordan and Benjamin A Valentino. 2011. "A war of one's own: Understanding the gender gap in support for war." *Public Opinion Quarterly* 75(2):270–286.

Burge, Camille, Julian J. Wamble and Rachel Cuomo. 2020. "A Certain Type of Descriptive Representative? Understanding How the Skin Tone and Gender of Candidates Influences Black Politics." *Journal of Politics* Forthcoming.

Butler, Daniel M, David W Nickerson et al. 2011. "Can learning constituency opinion affect how legislators vote? Results from a field experiment." *Quarterly Journal of Political Science* 6(1):55–83.

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

Camerer, Colin. 1997. Rules for experimenting in psychology and economics, and why they differ. In *Understanding Strategic Interaction*. Springer pp. 313–327.

Clarke, Kevin A. and David M. Primo. 2012. *A Model Discipline: Political Science and the Logic of Representations*. Oxford University Press.

Colburn, Timothy and Gary Shute. 2007. "Abstraction in computer science." *Minds and Machines* 17(2):169–184.

Colleau, Sophie M, Kevin Glynn, Steven Lybrand, Richard M Merelman, Paula Mohan and James E Wall. 1990. "Symbolic racism in candidate evaluation: An experiment." *Political Behavior* 12(4):385–402.

Converse, Jean M and Stanley Presser. 1986. *Survey questions: Handcrafting the standardized question-*

naire. SAGE Publications.
Coppock, Alexander. 2019. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods* 7(3):613–628.
Dawes, Christopher T, Peter John Loewen and James H Fowler. 2011. "Social preferences and political participation." *The Journal of Politics* 73(3):845–856.
Dickson, Eric S. 2009. "Do Participants and Observers Assess Intentions Differently During Bargaining and Conflict?" *American Journal of Political Science* 53(4):910–930.
Dickson, Eric S. 2011. Economics vs. Psychology Experiments: Stylization, Incentives, and Deception. In *Handbook of Experimental Political Science*, ed. James N. Druckman, Donald P. Green, James H. Kuklinski and Arthur Lupia. Cambridge University Press.
Druckman, James N, Erik Peterson and Rune Slothuus. 2013. "How elite partisan polarization affects public opinion formation." *American Political Science Review* 107(1):57–79.
Druckman, James N and Kjersten R Nelson. 2003. "Framing and deliberation: How citizens' conversations limit elite influence." *American Journal of Political Science* 47(4):729–745.
Dunning, Thad and Lauren Harrison. 2010. "Cross-cutting cleavages and ethnic voting: An experimental study of cousinage in Mali." *American Political Science Review* 104(1):21–39.
Evers, Miles M, Aleksandr Fisher and Steven D Schaaf. 2019. "Is There a Trump Effect? An Experiment on Political Polarization and Audience Costs." *Perspectives on Politics* 17(2):433–452.
Friedman, Sunder, Daniel Friedman and Shyam Sunder. 1994. *Experimental methods: A primer for economists*. 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.
Hainmueller, Jens and Daniel J Hopkins. 2015. "The hidden american immigration consensus: A conjoint analysis of attitudes toward immigrants." *American Journal of Political Science* 59(3):529–548.
Hashtroudi, Shahin, Sharon A Mutter, Elizabeth A Cole and Susan K Green. 1984. "Schema-consistent and schema-inconsistent information: Processing demands." *Personality and Social Psychology Bulletin* 10(2):269–278.
Herrmann, Richard K, Philip E Tetlock and Penny S Visser. 1999. "Mass public decisions on go to war: A cognitive-interactionist framework." *American Political Science Review* 93(3):553–573.
Johns, Robert and Graeme AM Davies. 2012. "Democratic peace or clash of civilizations? Target states and support for war in Britain and the United States." *The Journal of Politics* 74(4):1038–1052.
Kam, Cindy D and Elizabeth J Zechmeister. 2013. "Name recognition and candidate support." *American Journal of Political Science* 57(4):971–986.
Kanthak, Kristin and Jonathan Woon. 2015. "Women Don't Run? Election Aversion and Candidate Entry." *American Journal of Political Science* 59(3):595–612.
Karpowitz, Christopher F, J Quin Monson and Jessica Robinson Preece. 2017. "How to elect more women: Gender and candidate success in a field experiment." *American Journal of Political Science* 61(4):927–943.
Kertzer, Joshua D. 2020. "Re-assessing Elite-Public Gaps in Political Behavior." *American Journal of Political Science* Forthcoming.
Kreps, Sarah and Stephen Roblin. 2019. "Treatment format and external validity in international relations experiments." *International Interactions* Forthcoming.
Kriner, Douglas L and Francis X Shen. 2014. "Reassessing American casualty sensitivity: The mediating influence of inequality." *Journal of Conflict Resolution* 58(7):1174–1201.
Lau, Richard R, Lee Sigelman and Ivy Brown Rovner. 2007. "The effects of negative political campaigns: a meta-analytic reassessment." *Journal of Politics* 69(4):1176–1209.
LeVeck, Brad L. and Neil Narang. 2017. "The Democratic Peace and the Wisdom of Crowds." *International Studies Quarterly* 61(4):867–880.
Mattes, Michaela and Jessica L. P. Weeks. 2019. "Hawks, Doves and Peace: An Experimental Approach." *American Journal of Political Science* 63(1):53–66.

McDonald, Jared. 2019. "Avoiding the Hypothetical: Why "Mirror Experiments" are an Essential Part of Survey Research." *International Journal of Public Opinion Research* Forthcoming.

Morton, Rebecca B and Kenneth C Williams. 2010. *Experimental political science and the study of causality: From nature to the lab*. New York, NY: Cambridge University Press.

Mullinix, Kevin J, Thomas J Leeper, James N Druckman and Jeremy Freese. 2015. "The generalizability of survey experiments." *Journal of Experimental Political Science* 2(2):109–138.

Mummolo, Jonathan and Erik Peterson. 2019. "Demand effects in survey experiments: An empirical assessment." *American Political Science Review* 113(2):517–529.

Mutz, Diana C and Eunji Kim. 2017. "The impact of in-group favoritism on trade preferences." *International Organization* 71(4):827–850.

Nicholson, Stephen P. 2012. "Polarizing cues." *American journal of political science* 56(1):52–66.

Nielson, Daniel L., Susan D. Hyde and Judith Kelley. 2019. "The elusive sources of legitimacy beliefs: Civil society views of international election observers." *The Review of International Organizations* 14(4):685–715.

Paivio, Allan. 1990. *Mental representations: A dual coding approach*. New York, NY: Oxford University Press.

Press, Daryl G, Scott D Sagan and Benjamin A Valentino. 2013. "Atomic aversion: Experimental evidence on taboos, traditions, and the non-use of nuclear weapons." *American Political Science Review* 107(1):188–206.

Raffler, Pia. 2019. "Does political oversight of the bureaucracy increase accountability? Field experimental evidence from an electoral autocracy." Working paper.

Reeves, Andrew and Jon C. Rogowski. 2018. "The Public Cost of Unilateral Action." *American Journal of Political Science* 62(2):424–440.

Reiley, David. 2015. The lab and the field: empirical and experimental economics, by David Reiley. In *Handbook of experimental economic methodology*, ed. Guillaume R Fréchette and Andrew Schotter. Oxford University Press, USA pp. 410–412.

Renshon, Jonathan. 2015. "Losing Face and Sinking Costs: Experimental Evidence on the Judgment of Political and Military Leaders." *International Organization* 69(3):659–695.

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.

Rousseau, David L and Rocio Garcia-Retamero. 2007. "Identity, power, and threat perception: A cross-national experimental study." *Journal of Conflict Resolution* 51(5):744–771.

Rubenzer, Trevor and Steven B Redd. 2010. "Ethnic minority groups and US foreign policy: examining congressional decision making and economic sanctions." *International Studies Quarterly* 54(3):755–777.

Sartori, Giovanni. 1970. "Concept Misformation in Comparative Politics." *American Political Science Review* 64(4):1033–1053.

Shapira, Oren, Nira Liberman, Yaacov Trope and SoYon Rim. 2012. Levels of mental construal. In *SAGE Handbook of Social Cognition*, ed. Susan T. Fiske and C Neil Macrae. SAGE Publications pp. 229–250.

Steiner, Peter M, Christiane Atzmüller and Dan Su. 2016. "Designing valid and reliable vignette experiments for survey research: A case study on the fair gender income gap." *Journal of Methods and Measurement in the Social Sciences* 7(2):52–94.

Teele, Dawn Langan, Joshua Kalla and Frances Rosenbluth. 2018. "The Ties That Double Bind: Social Roles and Women's Underrepresentation in Politics." *American Political Science Review* 112(3):525–541.

Tingley, Dustin H and Barbara F Walter. 2011. "The effect of repeated play on reputation building: an experimental approach." *International Organization* 65(2):343–365.

Tomz, Michael. 2007. "Domestic audience costs in international relations: An experimental approach." *International Organization* 61(4):821–840.

Tomz, Michael R and Jessica LP Weeks. 2013. "Public opinion and the democratic peace." *American*

political science review 107(4):849–865.
Trager, Robert F and Lynn Vavreck. 2011. "The political costs of crisis bargaining: Presidential rhetoric and the role of party." *American Journal of Political Science* 55(3):526–545.
Trope, Yaacov and Nira Liberman. 2003. "Temporal Construal." *Psychological Review* 110(3):403–421.
Valentino, Nicholas A, Fabian G Neuner and L Matthew Vandenbroek. 2018. "The changing norms of racial political rhetoric and the end of racial priming." *The Journal of Politics* 80(3):757–771.
Yarhi-Milo, Keren, Joshua D. Kertzer and Jonathan Renshon. 2018. "Tying Hands, Sinking Costs, and Leader Attributes." *Journal of Conflict Resolution* 62(10):2150–2179.

Abstraction and Detail in Experimental Design:
Supplementary appendix

Contents

1 *Survey Overview*

The three experiments analyzed in our main text were embedded in two separate surveys, implemented in Spring 2019, and Spring 2020. Specifically, our NUCLEAR WEAPONS and IN-GROUP FAVORITISM experiments were fielded in Spring 2019, followed by a second study in Spring 2020 in which we fielded the ELITE CUE experiment. The implementation process of these studies followed a simple and common procedure further detailed in Figure 1.1.

1. **Informed consent and screening:** Respondents are asked to consent to the study, and are screened out if they are located outside of the US or are using a mobile device to answer the survey.
2. **Assignment to situational hypotheticality treatment:** Respondents are assigned to either an explicitly or implicitly hypothetical framing in our first wave. In our second study we randomized whether scenarios were described as real, or hypothetical or whether no description of hypotheticality was mentioned in the introduction the experiment. This treatment varies across respondents but remains constant across all studies for a given respondent. To strengthen this treatment, the emphasis on hypotheticality recurs in follow up questions that mention the experimental scenario.
3. **Assignment to order of experiments:** In both studies we randomized the order of studies to avoid ordering effects.
4. **Assignment to original study-level treatments:** Respondents are randomly assigned to the original conditions of studies. Unlike the assignment of the hypotheticality treatment, this assignment is independent across all studies.
5. **Assignment to contextual detail/actor identity treatments:** Respondents are randomly assigned versions of the original studies that vary in their amount of contextual detail, and in the identities of the actors in the scenarios. Unlike the situational hypotheticality treatment, this assignment is independent across all studies.
6. **Pre-Treatment Covariate Collection:** Respondents answered a battery of pre-treatment covariates, which we will employ in future analyses.
7. **Experiment completion:** Respondents participate in experiments and respond to our main outcome measures detailed below. Outcomes include original survey items as well as additional questions which investigate respondents' attention to the general vignette context and treatment.
8. **Additional Demographic and individual difference batteries:** Respondents respond to covariate batteries relating to: Foreign policy attitudes, cooperative internationalism, need for cognition, cognitive reflection (Thomson and Oppenheimer, 2016), political knowledge (Clifford and Jerit, 2016), and demographics.

Figure 1.1: Overview of Study Protocol

1.1 SAMPLE INFORMATION

Our first survey, in which we embedded the NUCLEAR WEAPONS and INGROUP FAVORITISM experiments, were implemented with Dynata (formerly known as Survey Sampling International (SSI)). Recent studies in political science have employed this platform for experimental research (see e.g. Kam (2012); Malhotra, Margalit and Mo (2013); Brutger and Rathbun (2020)). In Table 1, we report descriptive statistics of our sample, including basic demographics, and all variables employed in our analyses. Our ELITE CUE study was embedded in a second survey, implemented with Lucid. Recent investigations suggest that Lucid is a suitable platform for implementing survey experiments in the U.S. context (Coppock and McClellan, 2019) (For additional political science studies implemented with Lucid, see: Tomz and Weeks (2020); Hill and Huber (2019); Orr and Huber (2020)). We present additional descriptive statistics for our Lucid sample in Table 2.

Table 1: Descriptive Statistics: INGROUP FAVORITISM & NUCLEAR WEAPONS experiments

Statistic	N	Mean	St. Dev.	Min	Max
Age	4,311	50.807	17.322	0.000	99.000
Male	4,330	0.469	0.499	0.000	1.000
Female	4,330	0.525	0.499	0.000	1.000
Education	4,317	3.645	1.650	1.000	8.000
White	4,320	0.797	0.403	0.000	1.000
Black	4,320	0.082	0.274	0.000	1.000
Hispanic	4,320	0.043	0.203	0.000	1.000
Asian	4,320	0.050	0.218	0.000	1.000
Democrat	4,330	0.343	0.475	0.000	1.000
Republican	4,330	0.305	0.461	0.000	1.000
Independent	4,330	0.274	0.446	0.000	1.000

2 *Study Instrumentation*

2.1 ELITE CUES EXPERIMENT

The ELITE CUES experiment replicates Nicholson's (2012) study of elite cues about immigration reform in the United States, to explore the effects of actor identity in experimental design.[1] Nicholson's original study examined the effect of in/out party endorsements on partisan opinion in the context of a proposal to reform U.S. immigration policy that centered on a "path to citizenship"

[1]While Nicholson's study includes several experiments, considering different policies and cue-givers, we focus on the immigration policy experiment endorsed by politicians (rather than parties).

Table 2: Descriptive Statistics - ELITE CUES experiment

Statistic	N	Mean	St. Dev.	Min	Max
Age	4,030	45.190	17.301	1.000	98.000
Male	4,026	0.474	0.499	0.000	1.000
Female	4,026	0.517	0.500	0.000	1.000
Education	3,997	4.588	1.945	1.000	8.000
White	4,028	0.724	0.447	0.000	1.000
Black	4,028	0.117	0.321	0.000	1.000
Hispanic	4,028	0.072	0.259	0.000	1.000
Asian	4,028	0.042	0.201	0.000	1.000
Democrat	4,026	0.349	0.477	0.000	1.000
Republican	4,026	0.343	0.475	0.000	1.000
Independent	4,026	0.233	0.423	0.000	1.000

and used high-salience real actors: Barack Obama or John McCain. In our replication, we updated the relevant salient cuegivers (Joe Biden or Donald Trump), while also adding additional actor identity treatments that vary whether the immigration reform endorsement is made by less salient partisan cuegivers (Senator Tom Carper of Delaware or Senator Mike Rounds of South Dakota), or by a fictional politician (Stephen Smith) whose partisanship we manipulate.[2] In each condition respondents were told whether the endorser was a Republican or Democrat and for the fictional endorser — Stephen Smith — the partisan affiliation was randomized. Respondents then indicated their support for the immigration reform policy. Following the main outcome variable, respondents were asked to think about the situation again then asked to complete a thought listing exercise and a factual manipulation check (whether the policy was endorsed by a member of a particular party or not endorsed by anyone). These latter questions enable us to determine how actor identities affect respondents comprehension and recall of the general experimental scenario as well as the treatment.

To replicate the main results presented in Nicholson (2012), all subjects read the following introduction, followed by a vignette whose features randomly varied across respondents:[3]

> There is much concern about immigration policy in American Politics. We are going to describe a situation / We are going to describe a real situation / We are going to describe a hypothetical situation.

[2] Additionally, we update the substantive context of the experiment to focus on protection for "Dreamers" in the U.S.

[3] Note that underlined aquamarine text signifies our hypotheticality treatment, and *italicized blue text* signifies the original study's treatment, which we extended to include diverging types of actor identities (made up, low salience, high salience).

Some parts of the description may strike you as important; other parts may seem unimportant. Please read the details very carefully. After describing the situation, we will ask your opinion about a policy option.

As you know, there has been a lot of talk about immigration reform policy in the news. One proposal *Empty / , backed by Democrat Joe Biden / , backed by Republican Donald Trump / , backed by Republican Mike Rounds / , backed by Democrat Tom Carper / , backed by Democrat Stephen Smith / , backed by Republican Stephen Smith*. provided protections for Dreamers-including legal status and a path to legal citizenship for some of them.

The main DV for this study was "What is your view of this immigration policy?" Response options ranged from 1 (strongly support) to 5 (strongly oppose). After collecting our main outcome variable we further ask respondents:

When you think about the situation / real situation / hypothetical situation you just read, what features of the situation / real situation / hypothetical situation come to mind? Please list these thoughts or considerations below.

Simply write down the first thought that comes to mind in the first box, the second in the second box, and so on. Please put only one idea or thought in a box.

We've deliberately provided more boxes below than we think most people will need, just so you have plenty of room.

Following the thought listing exercise detailed above, we directly investigate respondents' attention to their main treatment condition. To do so, we ask the following question:

Think back to the **most recent** scenario described to you earlier in the survey. Was the immigration policy described, endorsed by a member of the Democratic party, the Republican party, an independent candidate, or no one at all.
possible responses include:

- Endorsed by a member of the Democratic party
- Endorsed by a member of the Republican party

- Endorsed by an independent candidate
- Not endorsed by anyone

2.2 IN-GROUP FAVORITISM EXPERIMENT

The INGROUP FAVORITISM experiment replicates portions of Mutz and Kim's (2017) investigation of American trade preferences to study the effects of additional contextual detail. In replicating their basic framework, we focus on a common decision experimentalist grapple with when designing instruments: how much contextual detail should vignettes include? We do so by randomly assigning respondents to either the original short vignette, or a more elaborate vignette which provides further detail on the experimental scenario. Consistent with Bansak et al. (2020), we provide two types of additional context. The first is "filler" context, with peripheral information that increases the volume of text respondents are presented with, but is not expected to interact with the treatment. The second is "charged" context that similarly increases the length of the stimulus, but which is more relevant to the treatment. In so doing, we test how additional information that is either likely or unlikely to interact with the study's main treatment moderates the original findings.

In particular, when implementing our study, we consider how providing respondents with increased context moderates the main identified treatment effect. Thus we manipulate the context in the experimental vignette to include either: (1) no additional context, (2) filler context which is *unlikely* to interact with treatment, or (3) charged context which is *likely* to interact with treatment. Apart from our contextual detail treatment, we follow a simplified version of the procedure implemented in Mutz and Kim (2017). In a similar fashion to our ELITE CUES replication, we provide respondents with a thought listing exercise as well as a factual manipulation check. Doing so enables us to test whether increased contextual detail affects respondents' comprehension of experimental scenarios and treatments.

To replicate the main results of Mutz and Kim (2017), we present all subjects with the following introduction, along with a vignette whose contents randomly varied across respondents:

> There is much concern these days about intentional trade and job security. We are going to describe a hypothetical situation / situation the United States could face in the future. Some parts of the description may strike you as important; other parts may seem unimportant. Please read the details very carefully. After describing the

situation, we will ask your opinion about a policy option.

Here is the hypothetical situation / situation:

The United States is considering a trade policy that would have the following effects:

For each **1,000 / 10** people in the U.S. who gain a job and can now provide for their family, **10 / 1000** people in a country that we trade with will **gain new jobs and now be able to provide for their family / lose jobs and will no longer be able to provide for their family**.[a]

Additional context:

None

Filler Context: If approved, this policy will be implemented within the next two years. As part of the implementation process, a commission of government officials and bureaucrats will outline the financial implications of the policy and provide guidance to businesses on how the new agreement affects them. Lastly, a team comprised of bureaucrats from both countries will oversee the policy implementation process which is expected to last two years.

Over the past 20 years, the trade volume between the United States and this country has been steadily increasing. There have been some years where the volume of trade has increased rapidly, while other years it has been somewhat slower. Throughout the past 20 years, both countries have signed several agreements, which were implemented in good faith. Both countries export and import a wide range of products, which will be covered by the terms of the new agreement if it is approved.

Charged Context: If approved, this policy will be implemented within the next two years. Analysis of the agreement has determined that it will dramatically increase

trade between the countries. This has the potential to create new business opportunities in both countries, but may also make it harder for some companies to compete. Lastly, a team comprised of bureaucrats from both countries will oversee the policy implementation process which is expected to last two years.

Over the past 20 years, the trade volume between the United States and this country has been steadily increasing. More specifically, U.S. goods and service trade with this country totaled an estimated $258.7 billion in 2018. Exports were $121 billion; imports were $137.7 billion. The U.S. goods and services trade deficit with the country was $47.5 billion in 2018. Throughout the past 20 years, both countries have signed several agreements, which were implemented in good faith.

[a] Possible combinations are: 1,000 - 10 - gain, 10 - 1,000 - gain, 10 - 1000 - lose.

We use the following item to create the main DV of our study: "Would you be likely to support this trade policy or oppose this trade policy?" The possible answers to this questions are: Support or oppose. Conditional on expressing a general policy preference respondents are further asked: "Are you strongly opposed / supportive of this new trade policy or somewhat opposed / supportive of this new trade policy?" The possible answers to this questions are: Somewhat support / oppose or strongly support / oppose. From this question we devise a 1-4 scale ranging from (1) strongly oppose to (4) strongly support which represents our main dependent variable.

After collecting our main outcome variable we further ask respondents:

When you think about the scenario / hypothetical scenario you just read, what features of the scenario / hypothetical scenario come to mind? Please list these thoughts or considerations below.

Simply write down the first thought that comes to mind in the first box, the second in the second box, and so on. Please put only one idea or thought in a box.

We've deliberately provided more boxes below than we think most people will need, just so you have plenty of room.

Following the thought listing exercise detailed above, we directly investigate respondents' attention to their main treatment condition. To do so, we ask the following question:

Think back to the trade policy that was described to you earlier in the survey. Will our trading partner benefit more than the US, will the US benefit more than the trading partner, or will they be impacted equally?

possible responses include:

- The trading partner will benefit more than the US
- The US will benefit more than trading
- Both countries will benefit equally

2.3 NUCLEAR WEAPONS EXPERIMENT

The NUCLEAR WEAPONS experiment replicates Press, Sagan and Valentino's (2013) examination of norms against the use of nuclear weapons in public opinion, to study the effects of both actor identity and contextual detail in tandem. The original study investigated whether normative prohibitions against the use of nuclear weapons were a factor in the U.S. public's preferences about whether and how to use force in world politics. It did so by randomizing the relative probability of success for conventional attacks relative to nuclear attacks.[4]

We used our replication to consider the joint effects of contextual detail and actor identity, adding two additional treatment arms to the original study on nuclear aversion. More specifically, we manipulate the vignette's context to either include: (1) elaborate context (as in the original study) or (2) reduced context. We also consider four alternatives to country names, which include: (1) Syria (as in the original study), (2) an unnamed country ("a foreign country"), (3) a fictitious country name ("Malaguay"), or (4) a real and schema-inconsistent country (Bolivia). The extent to which real countries are schema-consistent with a given experimental scenario is an empirical question. Therefore, we fielded a pilot study on a sample of about 600 American adults recruited on Amazon Mechanical Turk, in which we described the experimental scenario in the NUCLEAR WEAPONS experiment in its un-named country format. We then presented the study's main outcome questions, and asked respondents to rate the likelihood that each of eleven different countries would be the actor in each scenario.[5] After the main outcome measure, we present respondents with a thought listing exercise and factual questions relating to the main treatment, as detailed in

[4] We simplified the original design to only include two treatment-levels for the probability of success, as as detailed in Appendix §2.3.

[5] For more information regarding our pretest procedure see Appendix §3.

Appendix §2.3.

To replicate the main results in Press, Sagan and Valentino (2013), we present all subjects with the following text:

> There is much concern these days about the spread of nuclear weapons. We are going to describe a hypothetical situation / situation the United States could face in the future. Some parts of the description may strike you as important; other parts may seem unimportant. Please read the details very carefully. After describing the situation, we will ask your opinion about a policy option.
>
> **Joint Chiefs Report Concludes Nuclear and Conventional Options for Destroying Al Qaeda Nuke Lab Equally Effective / Joint Chiefs Say U.S. Nuclear Options Offers Dramatically Increased Chances of Destroying Nuke Lab**
>
> *Expected Civilian Casualties, Physical Destruction Equivalent for Both Options / Chiefs Conclude Nuclear Option Has 90% Chance of Success, Conventional Only 45%*
>
> The Associated Press
>
> A report from *General Martin Dempsey, Chairman of the Joint Chiefs of Staff, / the Joint Chiefs of Staff* to the President **concludes that military strikes using nuclear or conventional weapons would be "equally effective" / concludes that nuclear weapons would be "dramatically more effective" than conventional strikes** in destroying an Al Qaeda nuclear weapons facility in *Syria / Malaguay / the country / Ecuador*.
>
> The report compares two American military options, a conventional strike using nearly one hundred conventionally-armed cruise missiles, and an attack using two small, nuclear-armed cruise missiles. **The report estimates that both options have a 90 percent chance of successfully destroying the Al Qaeda nuclear weapons lab / the conventional strike has a 45 percent chance of successfully destroying the atomic bomb lab while nuclear weapons increase the chances of success to approximately 90 percent**. *Empty* / Syria / Malaguay / the country / Ecuador has refused to allow international inspectors access to the facility.
>
> *The Joint Chief's assessment comes two weeks after Russian intelligence agents intercepted a shipment of centrifuges and low-enriched uranium which could be used to produce nuclear*

weapons. The bomb-making equipment was being smuggled out of Russia to an Al Qaeda facility located near a remote town in the north of Syria / Malaguay / the country / Ecuador. *The suspects in the smuggling operation were employed at a Russian nuclear lab. The smugglers confirmed under questioning that other shipments of centrifuges and low-enriched uranium had already been delivered to the Al Qaeda base, where the centrifuges are being used to make fuel for a nuclear bomb. The smugglers stated that there will be enough bomb grade material produced for at least one weapon within two weeks.* Syria / Malaguay / the country / Ecuador *has refused to allow international inspectors access to the facility./ Empty*

The Joint Chiefs' report to the President does not recommend a specific course of action, *However, it concludes that "because the Al Qaeda facility is comprised of a series of deeply buried bunkers, a strike would require either large numbers of conventional missiles, or two nuclear weapons, to destroy the facility." / but concludes that destroying the facility would require either large numbers of conventional missiles, or two nuclear weapons.*

Either option would have roughly a ninety percent chance of success, according to the report. / According to the report, because of the facility's depth, nuclear weapons would be far more effective for destroying the target.

The report was leaked to the Associated Press by a high-ranking administration official involved in planning the strike. According to the official, the centrifuges and nuclear materials are too large to be moved without detection. / Empty The US intelligence official stated that he has high confidence that Al Qaeda is within two weeks of producing an operational bomb. *After that, the official said, "all bets are off." According to Dr. David Wright, a nuclear expert at the Union of Concerned Scientists, an independent think-tank based in Washington, D.C., "If a bomb of this size exploded in New York City, it could easily kill 50,000 to 70,000 people." / ; estimates suggest that if a bomb of this size exploded in New York City, it could easily kill 50,000 to 70,000 people.*

The report states that the remote location of the Al Qaeda facility should limit civilian fatalities in Syria / Malaguay / the country / Ecuador *for either option. Because many conventional weapons would be required to destroy the Al Qaeda base, the report estimates that "the two options would kill approximately the same number of* Syrian / Malaguayian / foreign / Ecuadorian *civilians" ; about 1,000, including immediate deaths and long term consequences*

of the conventional and nuclear strike. As both options will rely on cruise missiles launched from U.S. naval vessels, the report concludes that "no U.S. military personnel are at risk in either operation." / The report states that Syrian / Malaguayan / the country's / Ecuadorian *civilian fatalities would be limited to about 1,000 for either option, including immediate deaths and long term consequences of the conventional and nuclear strike. No U.S. military personnel would be at risk in either operation.*

Target: Al Qaeda Nuclear Weapons		
	U.S Nuclear Strike	**U.S Conventional Strike**
Probability of Success	90%	**90% / 45%**
Estimated ***Syrian / Malaguayian / Foreign / Ecuadorian* Civilian Deaths**	1,000	1,000
IF U.S. STRIKE FAILS 50,000 - 70,000 US. CIVILIAN FATALITIES		
Chart from Joint Chief's report describing nuclear and conventional options for strike on Al Qaeda nuclear lab		

After reading the scenario, respondents are asked:

> Given the facts described in the article, if the United States decided to conduct a nuclear strike to destroy the Al Qaeda base, how much would you approve or disapprove of the U.S. military action? Given the facts described in this article, if the United States decided to conduct a conventional strike to destroy the Al Qaeda base, how much would you approve or disapprove of the U.S. military action?

For each question, respondents state their approval on a seven point scale ranging from strongly disapprove (1) to strongly approve (7). The are also asked:

> If you had to choose between one of the two U.S. military options described in the article, would you prefer the nuclear strike or the conventional strike?
>
> - strongly prefer the conventional strike;

- somewhat prefer the conventional strike;
- somewhat prefer the nuclear strike;
- strongly prefer the nuclear strike.

Like Press, Sagan, and Valentino, we use these three questions as our main dependent variables. We further include a question from the original instrument, which is directed towards respondents who stated their preference for conventional attacks. The question asks:

You said you preferred a conventional strike by the United States. Which of the following is the most important reason why you did not prefer the nuclear strike? Please select one.

- Using nuclear weapons increased the expected number of Syrian civilian fatalities in the operation;
- somewhat prefer the conventional strike;
- Using nuclear weapons is morally wrong;
- Using nuclear weapons in this situation might encourage other states or terrorist groups to use nuclear weapons against the U.S. or our allies in the future;
- Using nuclear weapons in this situation would damage America?s reputation with other countries;
- Using nuclear weapons did not provide a significant advantage over conventional weapons in destroying the target;
- Civilized countries don?t use nuclear weapons.

Lastly, we implement a similar set of post-treatment questions, to determine how country names and context impact respondents ability to recall the main treatment. These question include a recall survey item and a factual question regarding the treatment which are detailed below:

When you think about the scenario / hypothetical scenario you just read, what features of the scenario / hypothetical scenario come to mind? Please list these thoughts or considerations below.

Simply write down the first thought that comes to mind in the first box, the second in the second box, and so on. Please put only one idea or thought in a box.

We've deliberately provided more boxes below than we think most people will need, just so you have plenty of room.

Think back to the scenario described to you earlier in the survey. What is the relation between the probability of success for nuclear and conventional attacks?
possible responses include:

- Nuclear attacks will be more successful than conventional attacks
- Conventional attacks will be more successful than nuclear attacks
- Conventional and nuclear attacks have similar probabilities of success

3 Power Calculations

In our experiments we have two sets of quantities of interest: the study-level treatment effects (e.g. in the NUCLEAR WEAPONS experiment, whether nuclear weapons are equally effective or dramatically more effective than conventional strikes), and interaction effects between the study-level treatments and our design treatments (e.g. whether the scenario is described as explicitly hypothetical or not). In order to ensure that these interaction effects are sufficiently powered, in this section, we consider the statistical power of our experimental design to detect theoretically meaningful moderating effects of different design choices. To do so, we focus on the NUCLEAR WEAPONS experiment), because it has the largest number of experimental cells, due to the fact that the country-name treatment includes four design-choice conditions: i) no-name, ii) made-up name, iii) schema-inconsistent name, and iv) schema-consistent name. In each of our main models, we compare the original study's average-treatment effect under the no-name condition, with one of

the other country conditions. This effectively leads us to estimate models with approximately 1000 observations, in which our quantity of interest is the effect of nuclear effectiveness, conditional on country name choice.

Our key question is whether we are sufficiently powered to precisely estimate η, in the model depicted in equation 1. Specifically, we want to ensure that if altering country names in a given experiment (i.e. shifting γ_{design} from 0 to 1) shapes a study's average treatment effect, we would be sufficiently powered to detect it (formally denoted as $\eta(\beta_{treatment} * \gamma_{design})$).

$$y_i = \alpha + \beta_{treatment} + \gamma_{design} + \eta(\beta_{treatment} * \gamma_{design}) + \epsilon_i \tag{1}$$

We use our data, as well as simulation procedure in the `R` package `DeclareDesign` to address this concern. Specifically, we declare a model by specifying three quantities: i) the average treatment effect of the nuclear weapon study (nuclear effectiveness), ii) the average treatment effect of a country name choice (describing the country as Syria rather than an unnamed country), and iii) the interaction between each treatment.

In Figure 3.2, we report the main results from the `DeclareDesign` diagnosis based on 1,000 simulations. We consider five different interaction estimands:

- A negative interaction effect of -0.04 – based on the coefficient from our original model, estimating the moderating effect of Syria country name.
- A negative interaction effect of -0.13 – Resembling an attenuation equivalent to a 25% decrease in the original study's ATE.
- A negative interaction effect of -0.27 – Resembling an attenuation equivalent to a 50% decrease in the original study's ATE.
- A negative interaction effect of -0.40 – Resembling an attenuation equivalent to a 75% decrease in the original study's ATE.
- A negative interaction effect of -0.54 – Resembling a full attenuation of the original study's ATE.

The results reported in Figure 3.2 suggest that even with a sample of 400 subjects, we would be able to identify an interaction effect that fully attenuates our main treatment (Pink line). Note

Figure 3.2: Power Calculations

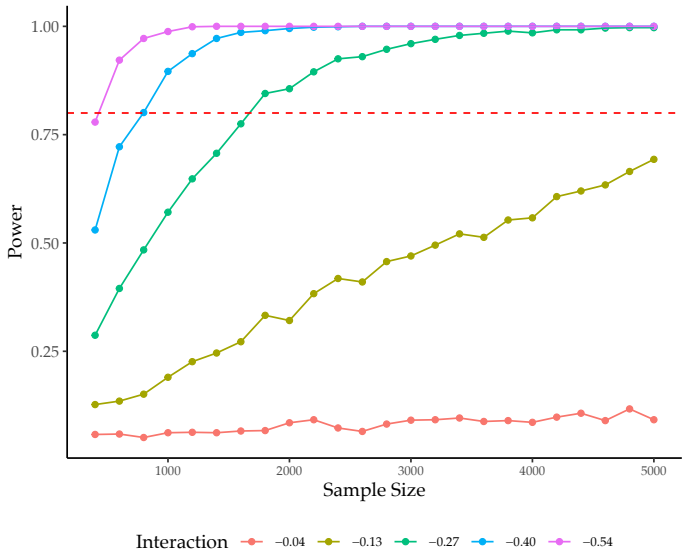

Figure 3.2 demonstrates our power to detect different interaction effect sizes, conditional on sample size.

that throughout the paper, all models include at least 1,000 subjects per comparison. Accordingly, we are relatively well-powered to detect moderating effects which attenuate our main average treatment effect by 50%-75% (blue and light-green line). That said, our ability to detect smaller attenuating effects — such as a 7%-25% attenuation in our main treatment is relatively limited (see dark green and orange lines).

Overall, the results of this exercise are encouraging. Even in our models with the smallest number of observations, we are well powered to detect design-moderating-effects which would lead scholars to draw directionally different conclusions. More so, we are well powered to detect attenuating effects that reduce (increase) main effects substantively (i.e. halving or doubling the original effect size), without changing the direction of a given average treatment effect.

4 *Pretest Procedure*

On March 18, 2019 we fielded a survey on a sample of 600 American adults recruited using Amazon Mechanical Turk to test the schema consistency of 11 different countries with the experimental scenarios presented in the original Press, Sagan and Valentino (2013) study on US policy towards

the development of nuclear attacks in foreign countries.[6] We also considered the schema consistency of countries for another replication of Tomz and Weeks (2013) democratic peace experiment, which we discuss in other work.

Our survey started off by requesting informed consent and screening out respondents located outside the US or respondents accessing the survey through non-desktop devices. Following this screening procedure, respondents were presented with the two experimental scenarios and their associated outcome questions. We randomized the sequencing of scenarios to avoid ordering effects. In addition, since both experiments relate to foreign policy and nuclear weapons, following the first scenario we emphasized that the second scenario describes a different situation.

To ensure the comparability of our pre-test and main study, we randomized all original study-level treatments apart from country name which was held constant at the unnamed country condition. After completing each scenario respondents were presented with a matrix of eleven countries, and asked: "On a scale of 1-5, where 1 is very unlikely and 5 is very likely, How likely is it that the above scenario describes the following countries?" The countries included in our pre-test were:

Egypt, Iran, Ecuador, Bolivia, Sudan, Vietnam, Turkey, Ethiopia, Kyrgyzstan, Malaysia, Syria

Parallel analysis suggests the likelihood ratings load onto three factors; principal axis factoring with oblimin rotation suggests the following three clusters:[7]

- **Countries outside the Middle East:** Ecuador, Bolivia, Vietnam, Ethiopia, Kyrgyzstan, Malaysia
- **Middle Eastern Adversaries:** Iran and Syria
- **Middle Eastern Others:** Egypt and Turkey

We therefore build (here, and in other work) on this clustering to inform our selection of country names, selecting Iran and Syria as schema consistent countries, and Ecuador and Bolivia as schema inconsistent countries.

[6] For recent articles fielded in political science journals using Amazon Mechanical Turk, see Brutger and Kertzer (2018); Tingley and Tomz (2014); Huff and Kertzer (2018); Renshon, Dafoe and Huth (2018).

[7] The model fit of a three-factor solution is good. For Tomz and Weeks: RMSEA=0.047, TLI=0.976; for Press, Sagan and Valentino: RMSEA=0.055, TLI=0.963.

5 *Actor Identities and Cognitive Burden and Treatment Recall*

In this section we present results of additional analyses relating to the ELITE CUE experiment. Specifically, we consider how the salience of an elite cue-giver, influences cognitive burden during the experimental procedure (measured by response latency), as well as treatment recall. Generally, we do not find evidence that actor type (made-up, low-salience, high-salience) impacts cognitive burden or treatment recall.

Figure 5.3: Actor Identity Effects on Response Times (ELITE CUE Experiment)

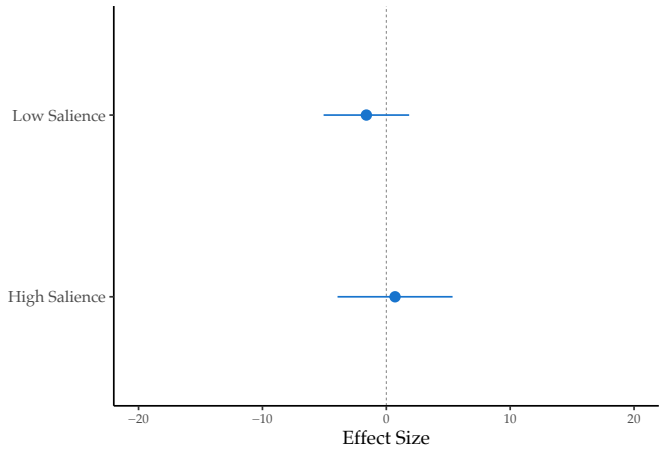

Figure 5.3 demonstrates that switching from a hypothetical actor to a low or high salience actor does not impact the cognitive burden of respondents (measured by response latency). Point estimates and corresponding confidence intervals are extracted from separate OLS models where the dependent variable–response time for the primary outcome measure– is regressed over a indicator taking the value of one for respondents assigned to high or low salience (rather than made up) condition. Sample Size for model comparing un-named and High-Salience actors is $n = 2428$. Sample Size for model comparing un-named and Low-Salience actors is $n = 2435$.

6 *Moderating Effects of Country Name Conditional on Contextual Detail*

Throughout the paper, we consider the moderating effects of design choices individually. However, one may wonder whether the consequences of different decisions regarding varying levels of design choices have interactive moderating effects on main treatments. To address this question, we leverage our NUCLEAR WEAPONS replication, in which we randomized both actor identity and contextual detail.

In figure 6.5, we present models where we consider the moderating effects of country names

Figure 5.4: Actor Identity Effects on Treatment Recall (ELITE CUE Experiment)

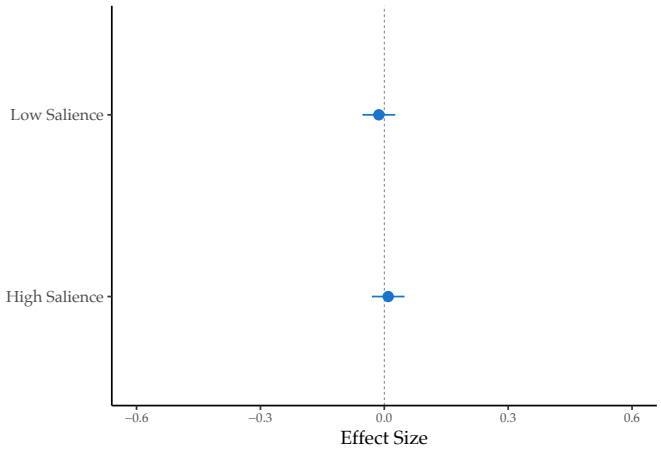

Figure 5.4 demonstrates that switching from a hypothetical actor to a low or high salience actor does not impact respondents' ability to correctly recall treatment. Point estimates and corresponding confidence intervals are extracted from separate OLS models where the dependent variable (correctly responding to the treatment recall question), is regressed over the actor identity treatment.

on original average treatment effects for two experimentally assigned sub-groups receiving either low or highly detailed vignettes. Generally, our findings suggest that the moderating effects of country names on original average treatment effects are not conditioned by the level of detail in an experimental vignette. However, we do find some evidence that adapting real world countries might have a small attenuating effect when context is low. That said, this conditional moderating effect, which approaches conventional levels of statistical significance ($p < 0.08$) will not lead experimenters to draw substantively different inferences.

Figure 6.5: Moderating Effects of Country Name by Contextual Detail Subsamples

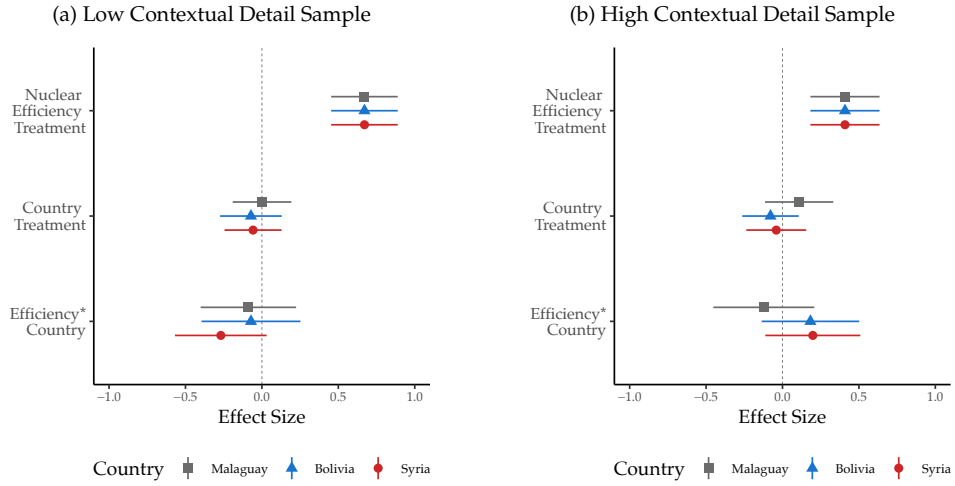

Figure 6.5 shows that different country names do not moderate average treatment effects in diverging and substantively significant ways across low and high contextually detailed vignettes in the NUCLEAR WEAPONS experiment. In each panel, point estimates and corresponding confidence intervals are extracted from three separate OLS models where original outcomes are predicted by original treatments interacted with country names. In all models across both panels un-named countries are the reference category.

References

Bansak, Kirk, Jens Hainmueller, Daniel J. Hopkins and Teppei Yamamoto. 2020. "Beyond the breaking point? Survey satisficing in conjoint experiments." *Political Science Research and Methods* Forthcoming:1–19.

Brutger, Ryan and Brian Rathbun. 2020. "Fair Share?: Equality and Equity in American Attitudes towards Trade." *International Organization* Forthcoming.

Brutger, Ryan and Joshua D. Kertzer. 2018. "A Dispositional Theory of Reputation Costs." *International Organization* 72(3):693–724.

Clifford, Scott and Jennifer Jerit. 2016. "Cheating on political knowledge questions in online surveys: An assessment of the problem and solutions." *Public Opinion Quarterly* 80(4):858–887.

Coppock, Alexander and Oliver A McClellan. 2019. "Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents." *Research & Politics* 6(1):2053168018822174.

Hill, Seth J and Gregory A Huber. 2019. "On the Meaning of Survey Reports of Roll-Call "Votes"." *American Journal of Political Science* 63(3):611–625.

Huff, Connor and Joshua D. Kertzer. 2018. "How the Public Defines Terrorism." *American Journal of Political Science* 62(1):55–71.

Kam, Cindy D. 2012. "Risk Attitudes and Political Participation." *American Journal of Political Science* 56(4):817–836.

Malhotra, Neil, Yotam Margalit and Cecilia Mo. 2013. "Economic Explanations for Opposition to Immigration: Distinguishing between Prevalence and Conditional Impact." *American Journal of Political Science* 57(2):391–410.

Mutz, Diana C and Eunji Kim. 2017. "The impact of in-group favoritism on trade preferences." *International Organization* 71(4):827–850.

Nicholson, Stephen P. 2012. "Polarizing cues." *American journal of political science* 56(1):52–66.

Orr, Lilla V and Gregory A Huber. 2020. "The policy basis of measured partisan animosity in the united states." *American Journal of Political Science* 64(3):569–586.

Press, Daryl G, Scott D Sagan and Benjamin A Valentino. 2013. "Atomic aversion: Experimental evidence on taboos, traditions, and the non-use of nuclear weapons." *American Political Science Review* 107(1):188–206.

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.

Thomson, Keela S and Daniel M Oppenheimer. 2016. "Investigating an alternate form of the cognitive reflection test." *Judgment and Decision Making* 11(1):99.

Tingley, Dustin and Michael Tomz. 2014. "Conditional Cooperation and Climate Change." *Comparative Political Studies* 47(3):344–368.

Tomz, Michael and Jessica LP Weeks. 2020. "Public opinion and foreign electoral intervention." *American Political Science Review* 114(3):856–873.

Tomz, Michael R and Jessica LP Weeks. 2013. "Public opinion and the democratic peace." *American political science review* 107(4):849–865.