Abstraction and Detail in Experimental Design[*]

Ryan Brutger[†], Joshua D. Kertzer[‡], Jonathan Renshon[§], Dustin Tingley[††] & Chagai M. Weiss[‡‡]

Running header: Abstraction in Experimental Design

Keywords: experiments; survey experiments; abstract; abstraction; vignette; generalizability

[*] 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, Soubhik Barari, Adam Seth Levine, Jonathan Mummolo, Rich Nielsen, Anne Sartori, Jonathan Woon, and Teppei Yamamoto. Authors listed in alphabetical order.

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

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

[§] 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

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

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

ABSTRACT: Political scientists designing 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 introduced into studies, the less control, and the more difficulty generalizing the results. Yet, we have reason to question this tradeoff, and there is relatively little systematic evidence to rely on when calibrating the degree of abstraction in 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 and extend a range of survey experiments, varying these levels of abstraction. We find no evidence that situational hypotheticality substantively changes experimental results, but increased contextual detail dampens treatment effects and the salience of actor identities moderates results in specific situations.

VERIFICATION MATERIALS: The data and materials required to verify the computational reproducibility of the results, procedures and analyses in this article are available on the American Journal of Political Science Dataverse within the Harvard Dataverse Network, at: `https://doi.org/10.7910/DVN/XGXEN6`

Word count: 9,922 words

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, but also has implications for construct validity. 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., 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).[1] 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).

[1] Experimental control is the degree to which researchers have control over the recruitment, assignment to conditions and measurement of subjects and variables and includes obvious features (the ability to randomly assign respondents to treatment arms) as well as less obvious features (such as the construal of the treatments). See McDermott (2002, 32).

And yet, as a discipline we know relatively little about the consequences of using abstract versus concrete experimental designs. Certainly, increasing "color in the laboratory" *may* trigger "unknown (to the experimenter) impressions and memories of past experiences over which the experimenter has no control" (Friedman, Friedman and Sunder, 1994, 53), 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 experimental design: *situational hypotheticality*, *actor identity*, and *contextual detail*. Our theoretical framework helps clarify when and why experimental control and generalizability may be affected by design decisions, but we also show how this debate bears on *construct validity*—the degree to which the variables in question are ”measured in ways that correspond to the theoretical concepts under investigation" (McDermott, 2002, 334; see also Findley, Kikuta and Denly, 2021, 368). Questions of measurement in experimental design usually center on the wording of outcome measures, but are just as relevant to how treatments are designed and the amount of ”slippage" between the concept and its operationalization in the experiment, which, in turn affects ”the internal validity of a study by affecting the interpretation of the results, as well as external validity" (see discussion in Hartman, 2021). 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, for those cases where scholars do have leeway, we present a framework to guide design decisions regarding the appropriate level of abstraction and detail.

Second, like other recent work seeking to subject conventional wisdom about experimental design principles to empirical scrutiny (Jerit, Barabas and Clifford, 2013; Mullinix et al., 2015; Dafoe, Zhang and Caughey, 2018; Coppock, 2019; Mummolo and Peterson, 2019; Kertzer, 2020), we test our theoretical framework, replicating and extending three well-known survey experiments in political science, and manipulating their levels of abstraction in three different ways. We find some dimensions of abstraction matter more than others. We find no evidence that situational hypotheticality changes the results experimenters obtain, an important finding as our field more broadly becomes increasingly concerned about the use of deception alongside other ethical issues related to the rise of experimentation (see Morton and Tucker 2014; Desposato 2015). Whether with politicians in American politics experiments, or countries in International Relations experiments, we find relatively little evidence that varying the abstraction of actor identities changes experimental results, although cue-taking experiments that use real and highly salient cue-givers obtain stronger effects than those that use low salience or fake actors. The strongest effects we find relate to contextual detail: we show that adding contextual detail to experimental vignettes attenuates the size of treatment effects and that this can be explained by respondents' lowered ability to recall the treatment. This suggests that choosing the appropriate level of contextual detail in experimental work thus depends on how much statistical power the author expects, as well as the purpose of the study: if the purpose is to demonstrate that an effect exists, a sparser experimental design better enables researchers to identify it, 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.

1 *Abstraction and detail*

One of the many design choices political scientists face when designing 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). 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").

Experimenters engage in abstraction when designing their stimuli, with a stimuli's level of abstraction determined by the contextual background it includes, the complexity of information it provides, and its emphasis on superordinate or subordinate elements of a given scenario. These dimension closely relate to questions about the appropriate level of abstraction, that 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. We suggest that abstraction in experimental design can be conceptualized along at least three dimensions: situational hypotheticality, actor identity, and contextual detail.[2] We review each dimension in detail in the discussion below.

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.

[2] These three strike us as the most important dimensions to confront experimentalists designing their studies, but the list is not necessarily exhaustive nor mutually exclusive.

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. Especially in IR experiments, a widely used form of implicit hypotheticality involves setting a scenario 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.

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, since experimenters often use real world actors in hypothetical scenarios.

Moving up the ladder of abstraction, some experimenters describe hypothetical scenarios in artificial countries. 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 further, 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). Other experiments focus on hypothetical political candidates. Banerjee et al. (2014), for example, 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.

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 beyond the experimental treatment. 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 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. Similarly in comparative politics, Bassan-Nygate and Weiss (Forthcoming) randomize whether experts project that an Israeli unity government will form in the near future, but they do not include much contextual detail in their vignette.

"Contextual detail" is composed of at least three related dimensions. The first is simply the volume of information provided. 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).[3] 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. (2021) call "filler" or its opposite, what we term "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. (2021) also show that other attributes

[3] See Kreps and Roblin (2019) for an experimental evaluation of treatment formats.

(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 *Control, Generalizability, and Construct Validity*

Political scientists employ experiments that vary along multiple dimensions in their degree of abstraction and detail. However, there is little certainty about the consequences of this variation. To address the uncertainty around the implication of design choices relating to abstraction and detail, 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). However, this approach is less than ideal because adjusting levels of abstraction on multiple dimensions simultaneously provides limited insight regarding the specific dimension driving experimental results.

There are some circumstances where for logistical or ethical reasons, experimenters will be constrained in terms of how abstract or detailed their stimuli will be. In other cases, however, experimentalists have more of a choice when designing their studies. In such cases, they often expect abstraction and detail to be consequential (Bostyn, Sevenhant and Roets, 2018; FeldmanHall et al., 2012; McDonald, 2019), and in tension with one another—associating the former with experimental control, and the latter with generalizability. We diverge from this perspective, and explain why the possible tension between experimental control and generalizability is more complex, and how these considerations connect to construct validity.

In specifying the level of abstraction and which elements of a construct are primary and which are secondary, the act of abstraction is inherently a theoretical phenomenon. In fact, this is exactly why discussions of abstraction in design that center on experimental control and generalizability are incomplete without consideration of construct validity (whether our operationalizations “meaningfully capture the ideas contained in the concepts” Collier and Adcock, 2001, 529). As McDermott (2002) points out, threats to construct validity come from manipulations that affect other concepts simultaneously, exactly the concern that experimentalists have tended to frame as being about experimental control. Manipulations that trigger multiple things at once affect both control *and* construct validity, and construct validity is necessary for a treatment to be externally valid as well (Findley, Kikuta and Denly, 2021, 371). Taken together, it's clear that—while the framing experimentalists have often used to describe the tradeoffs involved in the design decisions we examine has been about control and generalizability—in fact, both aspects are intrinsically tied to construct validity as well.

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 and construct validity. For example, *if* using particular country names as treatments triggers feelings, beliefs or frames separate from what the experimenter was attempting to manipulate, it would introduce confounding reducing experimental control and raising concerns about construct validity.

Yet, we argue that such concerns are likely exaggerated, because they oversimplify the relationship between design choices and experimental control. 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.[4] In contrast, increasing 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."[5]

More generally, we argue that it is misleading to think that by turning from real to hypothetical actors, or from contextually sparse to rich vignettes, experimenters necessarily gain (or lose) control

[4] 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.

[5] 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."

over their study. Indeed, when presented with relatively pared down stimuli, participants may "fill in the blanks." And when exposed to additional detail in vignettes, respondents may exert diverging reactions. Thus, the level of control and the validity of the construct measured does not necessarily increase (or decrease) with higher (or lower) levels of abstraction.

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 and the validity of the measure to generalize to other contexts. Political scientists often suspect that like control, degrees of generalizability may be shaped by levels of abstraction in experimental design. According to this perspective, when framing an experiment as hypothetical or real, when selecting particular actors, and when calibrating levels of contextual detail, researchers can condition the degree to which their results generalize beyond a particular context.

We argue that generalizability concerns are also exaggerated, because they oversimplify the relationship between design choices and generalizability. For example, experimenters oftentimes 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, this is reflected 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). For that reason, researchers may lean towards abstraction, and choose an unnamed actor.

However, when considering other dimensions in experiments, abstraction may actually decrease, rather than increase generalizability. Indeed, 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 are assigned to extremely abstract vignettes where they 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 in a more detailed vignette, embedded in a larger amount of information, the treatment is likely to exert a (realistically) smaller effect. Accordingly, the expected consequences of abstraction and detail might have contradictory implications across different dimensions of our framework.

2.3 Expectations

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 detailed ones, it is not clear that the tradeoffs are actually so stark.

For situational hypotheticality, we argue that concerns about abstraction are overblown, and so we do not expect varying situational hypotheticality to alter experimental results. Although scholars operating out of an economic tradition often express concerns that respondents won't take scenarios seriously or offer meaningful answers when told a scenario is hypothetical, there is relatively little empirical basis for these concerns. This is relevant given that there are many contexts 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 political candidate has engaged in unethical behavior — e.g. Butler and Powell, 2014).

In contrast, we expect stronger moderating effects for contextual detail. We expect that increasing the amount of contextual detail in an experiment may decrease treatment dosage, and therefore reduce the magnitude of identified effects, but the effect should be larger for charged context than filler context. Consistent with Bansak et al. (2021), one can think of experimentalists as considering 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 likely to affect how respondents react to the treatment. Even with charged context, however, we expect it to be very unlikely that additional context would reverse the direction of treatment effects.

For actor identity, we argue that experimentalists deciding between real-world or fictional actors should keep three considerations in mind. First, experiments using real world actors should maintain *schema consistency* (Hashtroudi et al., 1984): the choice of actor should be seen as reasonable or plausible 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), the country used should "fit" with the rest of the scenario. Thus, we argue that experimental control decreases if the experiment features a country that respondents know 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. Second, in experiments where the treatment manipulates a feature of an actor itself, experimentalists should consider whether the real actor they use in their vignettes allows them to maintain *treatment consistency*: all levels of an attribute of the actor being manipulated need to be perceived as equally plausible by respondents. It is for this reason that researchers are limited in their ability to select real world actors when studying the effects of race or gender in candidate selection, or the effects of country-level characteristics on foreign policy preferences. Of course, respondents' prior knowledge varies, and so some respondents may not know that an actor is schema or treatment inconsistent, while others may immediately recognize such incongruities in an experiment.[6] Third, because respondents are likely to have stronger attitudes about more salient real world actors than less salient ones, any differences between real- and hypothetical actors should be lower for less salient actors than more salient ones.

3 *Research Design*

To provide guidance for experimentalists on how abstract their 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 experiments, each designed to address one of the dimensions of abstraction described above. As Appendix page 3 shows, our typology applies to any type of experiment where researchers provide information to respondents, but for purposes of tractability we focus here on survey experiments in particular. Our study selection criteria sought to replicate and extend 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 conducive to manipulating situational hypotheticality, actor identity, and contextual detail.

We focus on three experiments (depicted in Table 1), each of which features three levels of treatment: (1) the central treatments from the original 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 ei-

[6] We discuss and test the role of prior knowledge in Appendix pages 27-31.

ther explicitly hypothetical, implicitly hypothetical, or real. An additional summary of the structure of our survey instrument is depicted in Appendix pages 1-2 and the details of each replication and extension are contained in Appendix pages 3-15.

Our first study, the ELITE CUES experiment, extends Nicholson (2012), which compares support for immigration policy amongst respondents receiving an in-party (or out-party) politician endorsement. In our extension, we updated the relevant salient cue-givers (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 cue-givers (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 IN-GROUP FAVORITISM experiment, extends 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. Those respondents assigned to additional context receive either "filler" or "charged" context to evaluate the effects of different types of contextual detail.

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 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).[7] 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.

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

	Elite Cues	In-Group Favoritism	Nuclear Weapons
Treatments from original study	1. No endorsement 2. In-Party Cue 3. Out-Party Cue	1. U.S. gain 1,000 and other country gains 10 2. U.S. gain 10 and other country gains 1,000 3. U.S. gains 10 and other country loses 1,000	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. No mention of hypotheticality 2. Explicitly hypothetical 3. Real	Situation described as: 1. Implicitly hypothetical 2. Explicitly hypothetical	Situation described as: 1. Implicitly hypothetical 2. Explicitly hypothetical
Sample	Lucid	Dynata/SSI	Dynata/SSI
Sample Size	*n*= 4,039	*n*=4,491	*n*= 4,462
Original Study	Nicholson (2012)	Mutz and Kim (2017)	Press, Sagan and Valentino (2013)

Table 1: Summary of Treatments for 3 Studies

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 explicitly hypothetical, implicitly hypothetical, or real, in order to test whether manipulating hypotheticality moderates the experimental findings.[8] The structure of the studies are depicted in Table 1. 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.[9] In Appendix pages 16-18, we report results of power simulations demonstrating that we are well powered to identify our quantities of interest.

[8] In the In-Group favoritism and nuclear weapons experiments, respondents were assigned to one of two conditions describing a situation as either implicitly or explicitly hypothetical. In the elite cues experiment respondents were assigned to one of three conditions describing a situation as either explicitly hypothetical, real, or a pure control condition where no situational hypotheticality information was provided.

[9] More details about each platform are available in Appendix pages 2-3.

4 *Results*

4.1 Replication of Original Study Results

In Figure 1 we present the central treatment effects from the three studies under investigation along with comparable estimates from the original studies.[10] As expected, our Elite Cues study demonstrates that respondents are more likely to oppose an immigration policy endorsed by an out-party politician. 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. 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 show that our extensions replicate the main results of the original studies.[11] 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, implicitly hypothetical, 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 or implicitly hypothetical, while in our Elite cues experiment, respondents were either assigned to an explicit hypotheticality condition, a real condition, or a pure control condition where no information about hypotheticality was provided.

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.[12]

[10] 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.

[11] We use "replicate" here to refer to an effect of the same sign that does not significantly differ in magnitude from the original estimate. The significant original effects make them easier cases for abstraction to matter, since we can't attribute weak interaction effects to outcome measures that are hard to move.

[12] In our elite cues experiment, hypotheticality can take one of three values (explicitly hypothetical, real, or the pure control). In our main analysis, we compare the explicitly hypothetical condition with the real condition, which are most

Figure 1: Replication of ATEs from the three experiments

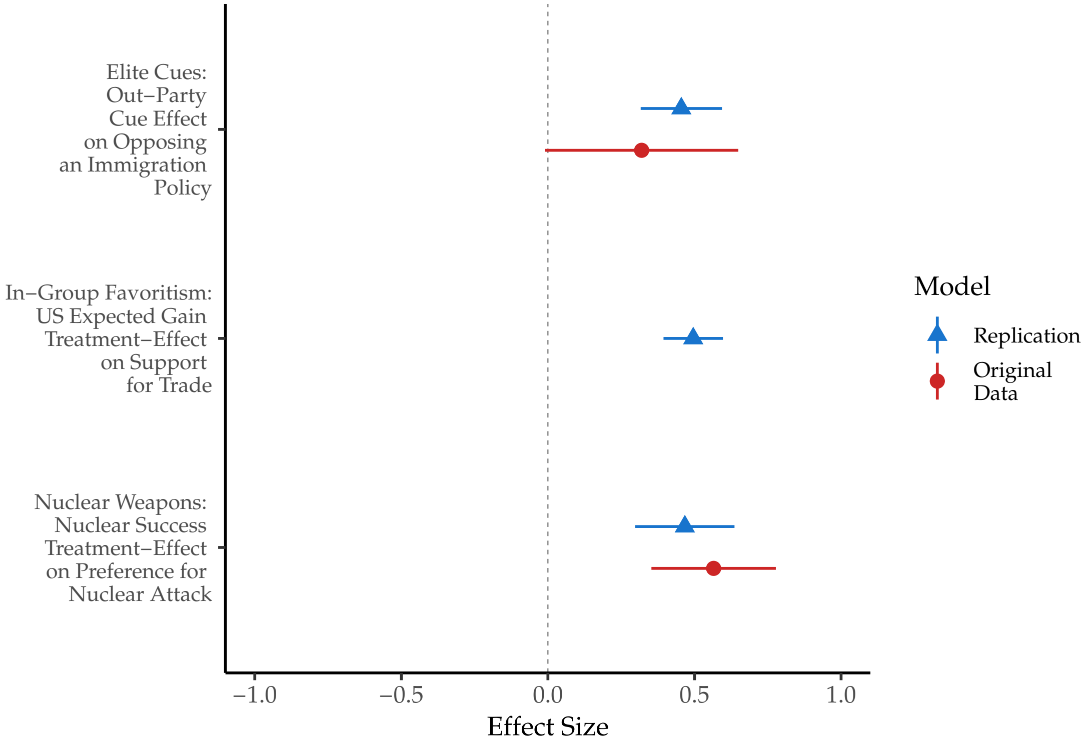


Figure 1 shows we successfully replicate the average treatment effects from the original studies. Point estimates and confidence interval are extracted from separate OLS models where original outcomes are predicted by treatments. All outcomes are standardized.

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

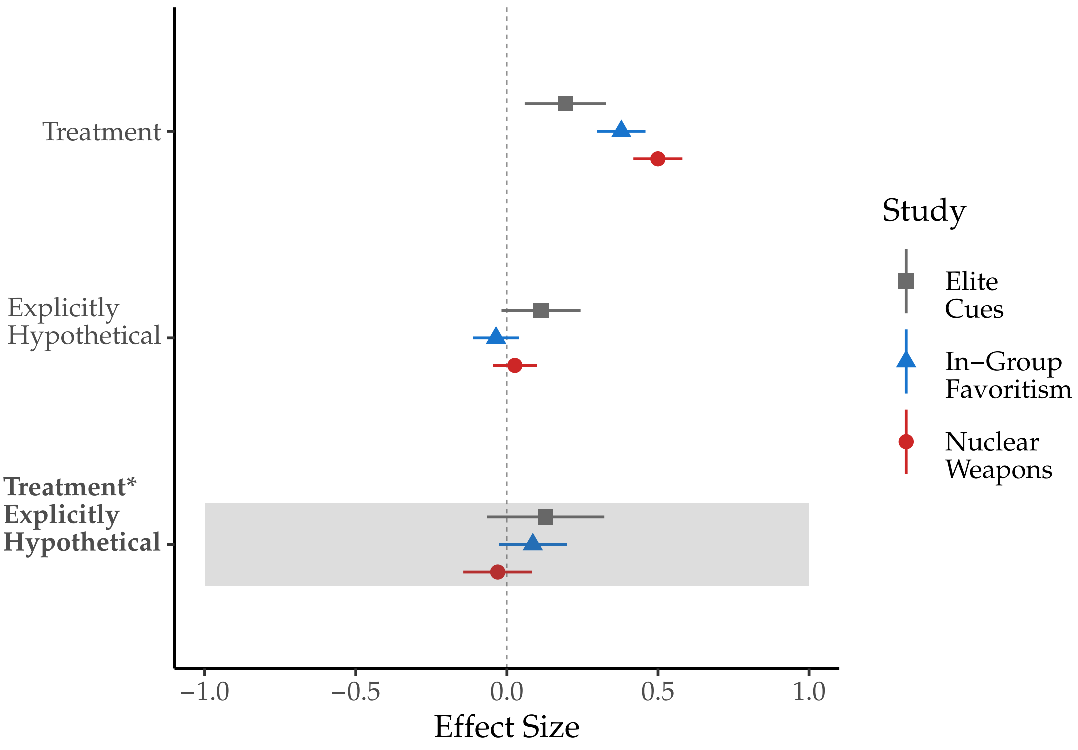


Figure 2 finds no evidence that situational hypotheticality significantly moderates our treatment effects in any of the three experiments. Point estimates and confidence intervals are extracted from three separate OLS models where original outcomes are predicted by original treatments interacted with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

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),

distinct, but comparing the explicitly hypothetical condition with the pure control yields similar results.

Figure 3: Moderating effects of actor identity condition

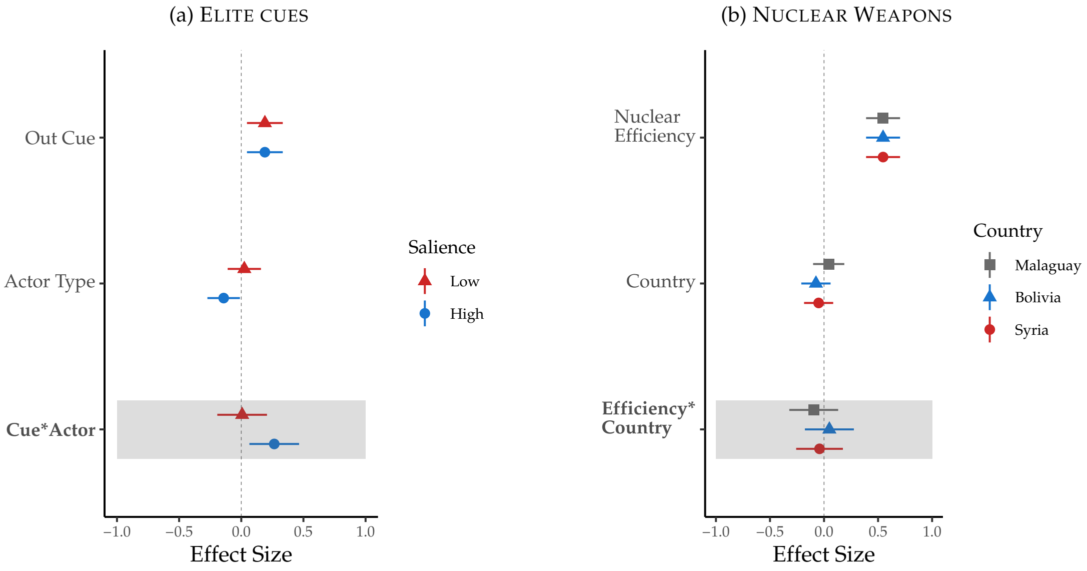


Figure 3 shows that manipulating country identity does not significantly moderate treatment effects in the NUCLEAR WEAPONS experiment, although we obtain slightly larger treatment effects in the ELITE CUES study when we use more salient cue-givers. Point estimates and confidence intervals are extracted 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.

or real and schema consistent (Syria). Similarly, in the ELITE CUES study 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 experiments 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.

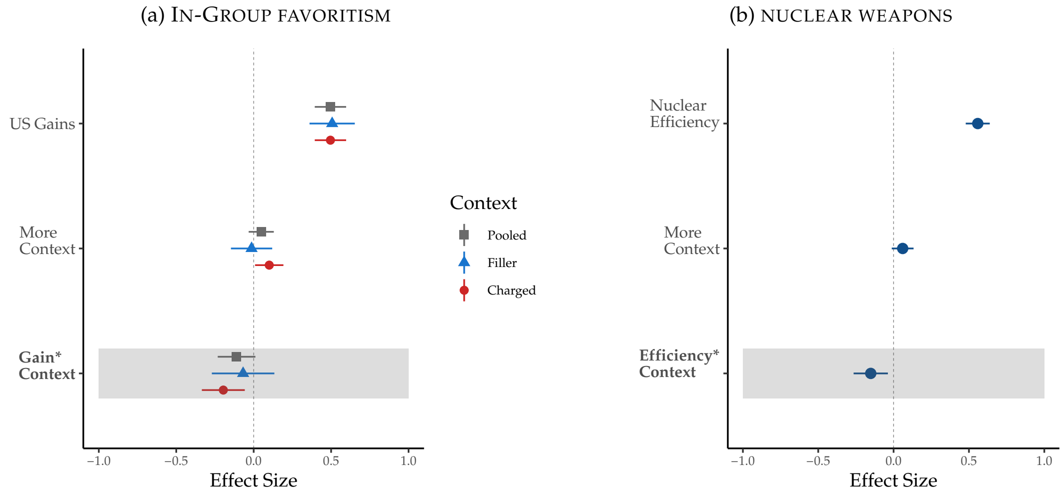
As demonstrated in Figure 3, with one important exception, we find that most actor identity conditions do not moderate the main treatment effects. Whether an actor is unnamed, fictional or real—and if real, schema-consistent or inconsistent—does not lead scholars to draw substantively different inferences 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 effect (when compared to baseline made-up actors).

There are two groups of potential mechanisms to explain the actor identity results from the ELITE CUE experiment. The first is a standard "online processing model" (Hill et al., 2013) in which respondents keep a running tally of evaluations that are updated when they come into contact with new information. McDonald (2019) proposes a version of this argument, arguing that hypothetical actors (compared to real actors) magnify treatment effects (by decreasing the role of prior knowledge or beliefs), and increase the cognitive burden on respondents, which would show up in increased response latency and lowered treatment recall (again, compared to real politicians). Yet, as we show in Appendix, pages 19-20, 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. A second potential mechanism that might be operative in this model is differential treatment recall: that respondents are better able to recall treatments from salient actors than non-salient ones. Yet, as Appendix pages 19-20 show, we find no evidence that treatment recall rates significantly vary with the actor identity treatment.

The second interpretation, which we believe is more consistent with our results, has to do with simple Bayesian models of persuasion, which focuses our attention on a different series of contrasts entirely. Bayesian models would first predict that when the DV is about measuring *attitudes towards a policy*, stronger respondent priors about the policy's endorser should lead to *more* updating (because respondents are likely to have stronger priors about the cue-giver). Our finding in the ELITE CUES study are an imperfect fit with the online processing model described above, but are consistent with this Bayesian model prediction. The results are also consistent with our expectation that, because respondents are likely to have stronger attitudes about more salient real world actors than less salient ones, any differences between real- and hypothetical actors should be smaller for less salient actors than more salient ones.[13] An additional prediction from this same Bayesian model—untested in our study— would be that when the DV involves measuring *attitudes about an actor*, stronger respondent priors should lead to *less* updating in response to information about the actor, consistent with Croco, Hanmer and McDonald (2021).

[13] See Appendix pages 27-35 for heterogeneous effects based on respondents' political knowledge and need for cognition.

Figure 4: Adding contextual detail attenuates treatment effects


(a) IN-GROUP FAVORITISM

(b) NUCLEAR WEAPONS

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*, we compare a baseline reduced context vignette with elaborate context conditions which are either filler (blue) or charged (red). We also consider a pooled model of both types of experimental context (gray). In panel *b*, a baseline reduced-context condition is compared with the original elaborate-context condition used in the original Nuclear Weapons experiment. All outcomes are standardized.

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 the original elaborate context vignette. In the IN-GROUP FAVORITISM experiment, respondents were either exposed to the original minimal context vignette (baseline), or an extended context vignette which included “filler” or “charged” additional context.

As demonstrated in Figure 4(b), 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(a) 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(a) this effect is driven by the "charged" context condition, which in and of itself has a statistically significant effect on the size (but not direction) of main treatment effects. In contrast, adding filler context does not significantly affect the magnitude of the treatment effect.

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. In Appendix pages 19-21, we further explore the positive effects of additional context on response latency, as well as the null effects of our actor identity treatment on correct recall and response latency.

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 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 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. Despite a recognition that these questions are, ultimately, subject to study and research like many other problems (e.g., Friedman,

Figure 5: Contextual Detail Effects on Treatment Recall Success

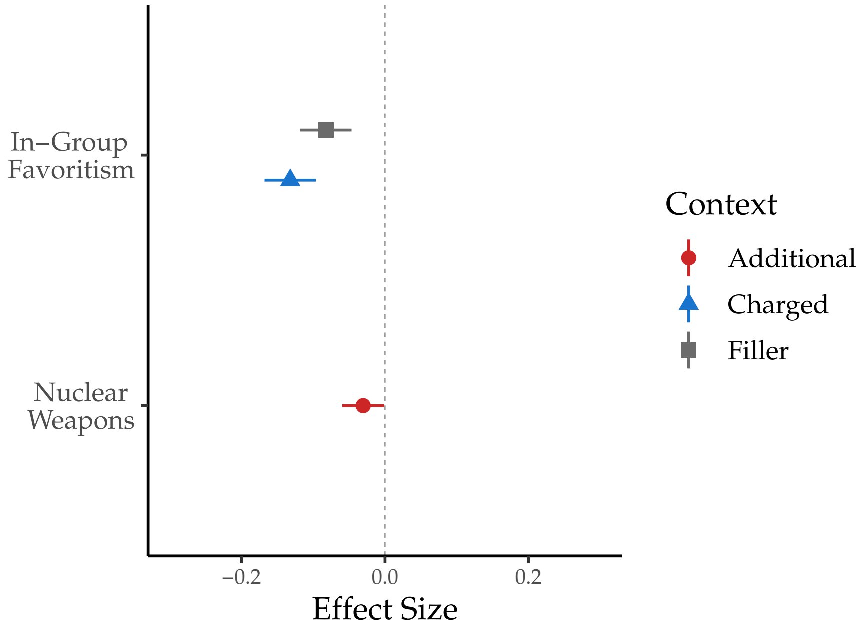


Figure 5 demonstrates how adding contextual detail negatively affects treatment recall. Point estimates and confidence intervals are extracted from three separate OLS models where a binary treatment recall success variable is predicted by the context condition to which respondents were assigned. The 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.

Friedman and Sunder, 1994), there is little in the way of theoretical frameworks or empirically-minded guidance for researchers who face these issues. In line with other recent work, we seek to subject these folk intuitions about experimental methods to empirical scrutiny (Mullinix et al., 2015; Coppock, 2019; Mummolo and Peterson, 2019; Kertzer, 2020)

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 elucidate when there are, and are not, important tradeoffs between experimental control and generalizability. 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. We also provide empirical leverage on the tricky question of how to appropriately operationalize the concepts we care about; empirical political scientists study "specific instances of units, treatments, observations, and settings" (Shadish, Cook and Campbell, 2002), but figuring out the implications of those specifics can now more appropriately be guided by theory and empirics in combination.

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 explicitly hypothetical, implicitly hypothetical, or real.

Our results suggest reasons for optimism. Situational hypotheticality does not make any substantial difference, failing to affect any of the main findings from the three 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 subtracting context from our NUCLEAR WEAPONS experiment. Our results are consistent across both studies: additional context leads to more conservative estimates of treatment effects, dampening 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 how much statistical power the author expects, as well as the purpose of the study: if the purpose is to demonstrate that an effect exists, a sparser experimental design better enables researchers to identify this effect, 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. Our results also suggest the utility of future work designed and powered to detect exactly how the potential causal mechanisms through which abstraction works—e.g., treatment recall or schema-consistency—interact with each other.[14]

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 attributes 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 important exception is that more salient politicians make more effective cue-givers than fictional actors do. Drawing out the implications of two potential explanations, we find suggestive evidence that this might be understandable within a Bayesian persuasion model, which would also be consistent with findings from research that manipulates the hypotheticality of the actor and measures outcomes related to attitudes about that actor (rather than the policy) (Croco, Hanmer and McDonald, 2021). Altogether, our framework and results clarifies where, when, and how researchers might have discretion in selecting particular levels of abstraction in their experimental stimuli.

[14] In Appendix pages 22-26, we find little evidence of interaction effects between different types of abstraction.

References

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. 2021. "Beyond the breaking point? Survey satisficing in conjoint experiments." *Political Science Research and Methods* 9(1):53–71.

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

Bassan-Nygate, Lotem and Chagai M Weiss. Forthcoming. "Party Competition and Cooperation Shape Affective Polarization: Evidence from Natural and Survey Experiments in Israel." *Comparative Political Studies* .

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.

Bostyn, Dries H., Sybren Sevenhant and Arne Roets. 2018. "Of Mice, Men, and Trolleys: Hypothetical Judgment Versus Real-Life Behavior in Trolley-Style Moral Dilemmas." *Psychological Science* 29(7):1084–1093.

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* 82(4):1596–1601.

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.

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.

Collier, David and Robert Adcock. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95(3):529–546.

Coppock, Alexander. 2019. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods* 7(3):613–628.

Croco, Sarah E, Michael J Hanmer and Jared A McDonald. 2021. "At What Cost? Reexamining Audience Costs in Realistic Settings." *The Journal of Politics* 83(1):000–000.

Dafoe, Allan, Baobao Zhang and Devin Caughey. 2018. "Information equivalence in survey experiments." *Political Analysis* 26(4):399–416.

Dawes, Christopher T, Peter John Loewen and James H Fowler. 2011. "Social preferences and political participation." *The Journal of Politics* 73(3):845–856.

Desposato, Scott. 2015. *Ethics and experiments: Problems and solutions for social scientists and policy professionals*. Routledge.

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.

FeldmanHall, Oriel, Dean Mobbs, Davy Evans, Lucy Hiscox, Lauren Navrady and Tim Dalgleish. 2012. "What we say and what we do: the relationship between real and hypothetical moral choices." *Cognition* 123(3):434–441.

Findley, Michael G, Kyosuke Kikuta and Michael Denly. 2021. "External Validity." *Annual Review of Political Science* 24:365–393.

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.

Hartman, Erin. 2021. "Generalizing Experimental Results." *Advances in Experimental Political Science* p. 385.

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.

Hill, Seth J, James Lo, Lynn Vavreck and John Zaller. 2013. "How quickly we forget: The duration of persuasion effects from mass communication." *Political Communication* 30(4):521–547.

Jerit, Jennifer, Jason Barabas and Scott Clifford. 2013. "Comparing contemporaneous laboratory and field experiments on media effects." *Public Opinion Quarterly* 77(1):256–282.

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* 45(3):576–594.

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.

McDermott, Rose. 2002. "Experimental methodology in political science." *Political Analysis* pp. 325–342.

McDonald, Jared. 2019. "Avoiding the Hypothetical: Why "Mirror Experiments" are an Essential Part of Survey Research." *International Journal of Public Opinion Research* 32(2):266–283.

Morton, Rebecca B and Joshua A Tucker. 2014. "Experiments, Journals, and Ethics." *Journal of Experimental Political Science* 1(2):99–103.

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.

Shadish, William, Thomas D Cook and Donald Thomas Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

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

A Survey Overview

The three experiments analyzed in our main text were embedded in two separate waves, 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 wave 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 A.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 condition in our first wave. In our second wave we randomized whether scenarios were described as explicitly hypothetical, real, or a pure control condition where 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 A.1: Overview of Study Protocol

A.1 Sample information

Our first survey, in which we embedded the Nuclear Weapons and In-Group Favoritism experiments, were implemented with Dynata (formerly known as Survey Sampling International (SSI)).[1] 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.[2] We present additional descriptive statistics for our Lucid sample in Table 2. Importantly, while the sampling strategies of both survey providers differ from one another, they both involve online panels. Future research should extend our findings on samples fielded using other sampling strategies, including on survey respondents that vary in their level of naïveté (e.g. Chandler, Mueller and Paolacci, 2014), as well as in other survey modes. One potential advantage of online survey experiments is that they afford researchers the opportunity to provide greater contextual detail than telephone-based survey experiments do, such that it is likely that contextual detail has greater attenuating effects in telephone-based survey experiments than it does online, for example.

Table 1: Descriptive Statistics - Study I (MK+PSV)

Statistic	N	Mean	St. Dev.	Min	Max
Age	4,289	51.040	17.052	18.000	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

[1] For other recent studies in political science employing this platform for experimental research, see e.g. Kam (2012); Malhotra, Margalit and Mo (2013); Kertzer and Brutger (2016); Brutger and Rathbun (2020).

[2] Recent investigations suggest that Lucid is a suitable platform for implementing survey experiments in the U.S. context (Coppock and McClellan, 2019), and have found that experiments fielded on Lucid before the COVID-19 pandemic replicated during the COVID-19 pandemic as well (Peyton, Huber and Coppock, 2021). For additional political science studies implemented with Lucid, see Tomz and Weeks (2020); Hill and Huber (2019); Orr and Huber (2020).

Table 2: Descriptive Statistics - Study II (Nicholson)

Statistic	N	Mean	St. Dev.	Min	Max
Age	4,025	45.236	17.262	18.000	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

B *Study Instrumentation*

B.1 Case selection

As Table 3 shows, the conceptual typology we employ in the piece for discussing different dimensions of abstraction is applicable to a wide range of experiments, which we illustrate below by coding a number of prominent experimental pieces in Table 3.

Table 3: Abstraction in experimental political science

		Type of abstraction		
		Situational	Actor	Contextual
Type of experiment	Example	Hypotheticality	Identity	Detail
Audit experiment	Butler and Broockman (2011)	Deception	N/A	Med
Conjoint experiment	Hainmueller and Hopkins (2015)	Implicit	Unnamed	Med
Econ-style lab experiment	Kanthak and Woon (2015)	Real	Unnamed	Low
Endorsement experiment	Lyall, Blair and Imai (2013)	Real	Real	Med
Framing experiment	Nelson, Clawson and Oxley (1997)	Deception	Real	High
Lab-in-the-field experiment	Habyarimana et al. (2007)	Real	Unnamed	Low
Scenario-based survey experiment	Tomz (2007)	Implicit	Unnamed	Med
War game	McDermott et al. (2007)	Simulation	Artificial	Med
Field Experiment	Lyall, Zhou and Imai (2020)	Real	Real	High

We believe our typology can be applied to any information provision experiment, where respondents are presented with information by the experimenter to see how it affects their behavior. In this article, we chose to investigate questions of abstraction in survey experiments in particular for two reasons. The first is their popularity. Survey experiments now constitute the most widely-used experimental method appearing in many top journals in political science: for example, the American Journal of Political Science published 205 experiments from 2011-2020, 121 of which were survey experiments, pointing to the relevance of our focus here. The second concerns questions of tractability. Although issues of abstraction and concreteness in experimental design apply to non-survey experiments as well (such as economics-style bargaining games (e.g. Tingley and Walter, 2011; Kanthak and Woon, 2015; Kertzer and Rathbun, 2015)), they also introduce additional considerations. Economics-style lab experiments, for example, embrace abstract designs in order to isolate the effect of incentives – a consideration we set aside in the manuscript itself.

The manuscript ultimately replicates and/or extends three different survey experiments, chosen through three selection criteria. First, we chose prominent studies that focused on core theoretical debates across a range of subfields of political science. The ELITE CUES experiment comes from American politics (Nicholson, 2012), though the construct being studied — the role of elite endorsements in support for policy preferences – features prominently in the study of political behavior regardless of subfield (e.g. Bullock, 2011; Druckman, Peterson and Slothuus, 2013; Guisinger and Saunders, 2017; Bisgaard and Slothuus, 2018; McDonald, Croco and Turitto, 2019). The IN-GROUP FAVORITISM experiment comes from international political economy (Mutz and Kim, 2017), but questions of in-group favoritism and intergroup relations also loom large in the study of domestic politics as well (e.g. Iyengar and Westwood, 2015; McClendon, 2018; Nugent, 2020). The NUCLEAR WEAPONS experiment comes from international security, but the underlying questions it tests about the strength of our commitment to moral principles is by no means exclusively the preserve of the security literature (e.g. Chu, 2019; Ryan, 2019; Jung, 2020).

Second, because our quantity of interest is the interaction between a dimension of abstraction and the study-level treatment, the selected studies needed to have relatively simple designs to afford us sufficient statistical power. Complex or high-dimensional factorial experiments, or conjoint experiments with multiple treatments of interest, are less useful for our purposes than experiments that had only one factor of interest, particularly if the factor had only two or three levels.

Third, the experiments selected needed to demonstrate a large and substantively meaningful effect. If we were replicating or extending studies whose original treatments barely moved their outcome variable, the absence of heterogeneous treatment effects by levels of abstraction would be less informative than when replicating studies whose treatment effects were substantively large.

Fourth, the experiments needed to be conducive to manipulating our three dimensions of abstraction: situational hypotheticality, actor identity, and contextual detail. We did so in each experiment in different ways. For example, in the NUCLEAR WEAPONS experiment, whose stimuli are relatively lengthy, we manipulated contextual detail by *cutting* context, whereas in the IN-GROUP FAVORITISM experiment, whose stimuli are relatively short, we were able to manipulate contextual detail by *adding* context. In the ELITE CUES experiment, we manipulate actor identity by manipulating the individual politician involved; in the NUCLEAR WEAPONS experiment we manipulate actor identity by manipulating the country involved. This also speaks to the value of selecting experiments from across multiple subfields of the discipline, since abstraction and concrete detail may manifest themselves in very different ways depending on the research question.

Finally, we note that the fact that our experiments differ from one another in a wide range of ways — they cover substantively different questions, from different quadrants of political science, were fielded on different survey platforms, at different points in time — and yet still converge on a common set of findings should add credibility to our findings, even though, as is the case with all social science research, additional studies are inevitably required, which we explore in additional work (Brutger et al., 2022).

B.2 ELITE CUES EXPERIMENT

The ELITE CUES experiment replicates and extends Nicholson's (2012) study of elite cues about immigration reform in the United States, to explore the effects of actor identity in experimental design.[3] 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" and used high-salience real actors: Barack Obama or John McCain. In our extension, 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.[4] 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 par-

[3] 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).

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

ticular 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 introduction and vignette presented in Figure B.2, whose features randomly varied across respondents:[5]

> There is much concern about immigration policy in American Politics. We are going to describe a situation / real situation / hypothetical situation.
> 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.
> What is your view of this immigration policy?" (5 point scale, ranging from strongly support to strongly oppose)

Figure B.2: Elite Cue Vignette

After administering our main outcome variable (shown at the bottom of Figure B.2) we asked respondents to complete a common thought listing task (See Figure B.3).

> 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.

Figure B.3: Thought Listing Exercise

Following the thought listing exercise in Figure B.3, we directly investigate respondents' attention to their main treatment condition, by employing a factual manipulation check (Kane and Barabas, 2019). To do so, we ask the question presented in Figure B.4:

[5] 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).

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.

- 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

Figure B.4: Elite Cue Treatment Manipulation Check

B.3 In-Group Favoritism experiment

The In-Group favoritism experiment replicates and extends 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. (2021), 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.[6] 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 study, 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.[7]

To replicate and extend 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:

[6] Crucially, the distinction between filler and charged context is less about whether the additional context is relevant to the scenario or not – in most circumstances, experimentalists are unlikely to be interested in adding totally irrelevant text, which would present somewhat jarringly to respondents — and more about whether to add additional contextual information that they expect to interact with their treatment of interest.

[7] Future work should consider exploring whether charged context with numeric context produces different results than charged context without out numeric context, particularly for individuals at varying levels of numeracy (Mérola and Hitt, 2016).

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.

Figure B.5: In-Group Favoritism Vignette

After reading the vignette described in Figure B.5, respondents were exposed to a two-stage outcome measure reported in Figure B.6:

Based on the questions reported in Figure B.6, we created our main DV, measuring support for

- Would you be likely to support this trade policy or oppose this trade policy? (Support / Oppose)
  - **If support**: Are you strongly supportive of this new trade policy or somewhat supportive of this new trade policy? (Strongly supportive / somewhat supportive)
  - **If oppose**: Are you strongly opposed of this new trade policy or somewhat opposed of this new trade policy? (Strongly opposed / somewhat opposed)

Figure B.6: In-Group Favoritism Outcomes

the described policy, on a four point scale ranging from strongly oppose to strongly support.

After collecting our main outcome variable we further ask respondents to engage in a thought listing task. The thought listing task is similar to the one reported in Figure B.3. Following the thought listing exercise detailed above, we directly investigate respondents' attention to their main treatment condition. To do so, we ask the following manipulation check reported in Figure B.7:

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

Figure B.7: In-Group Favoritism Manipulation Check

B.4 Nuclear Weapons experiment

The nuclear weapons experiment replicates and extends 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.[8]

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.[9] After the main outcome measure, we present respondents with a thought listing exercise and factual questions relating to the main treatment, as detailed in Appendix §2.3.

To replicate the main results in Press, Sagan and Valentino (2013), we present all subjects with the following text, as well as a summary table (see Table 4):

> 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.

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

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

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.

Table 4: Table Accompanying Nuclear Weapons Experiment

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 reported responses to three outcome question (see Figure B.8). Our main outcome of interest is approval of a nuclear weapon attack (the first item in Figure B.8).

- 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? (6 point approve disapprove scale)
- 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? (6 point approve disapprove scale)
- 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.

Figure B.8: Nuclear Weapons Outcome Questions

We further included a question from the original instrument, which is directed towards respondents who stated their preference for conventional attacks, and the reasons behind this selection. However, we do not analyze responses to this question in our paper. We also included a thought listing exercise relating to the nuclear weapons vignette, like the one depicted in Figure B.3. Lastly, we asked respondents a manipulation check question (see Figure B.9).

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

Figure B.9: Nuclear Weapons Manipulation Check

C 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 $\eta(\beta_{treatment} * \gamma_{design})$).[10]

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

We use our data, as well as simulation procedure in the `R` package `DeclareDesign` to address this concern (Blair et al., 2019). 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 C.10, we report the main results from the `DeclareDesign` diagnosis based on 1,000

[10] Lenth (2007) provides a useful discussion of power analysis, which discourages the use of retrospective analyses. However, Lenth (2007, E26) also notes that retrospective power analysis can be valuable when it uses "what we have learned (e.g., the error SD) and an effect size deemed of clinical importance" to determine the appropriate amount of data to identify effects of importance. We do so in our power analysis, which shows that we are well powered to identify interaction effects that would offset a significant proportion of the main treatment effects.

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.

Figure C.10: Power Calculations

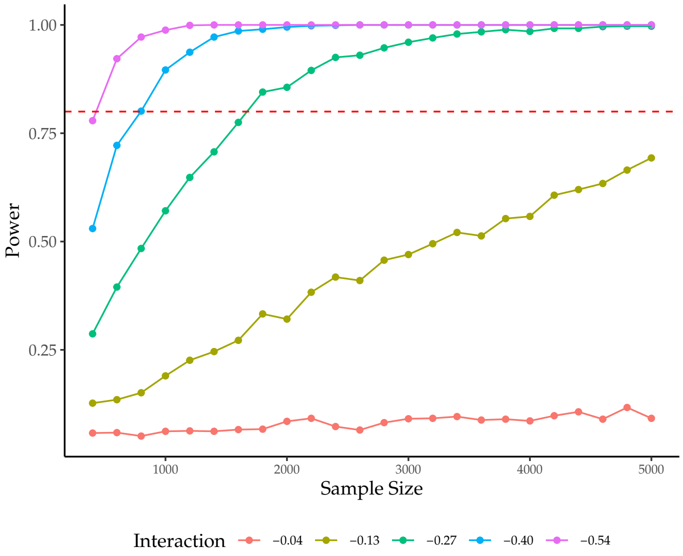


Figure C.10 demonstrates our power to detect different interaction effect sizes, conditional on sample size.

The results reported in Figure C.10 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 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).[11]

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.

D 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.[12]

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. 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 the 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:[13]

[11] We also acknowledge that surveys with less attentive respondents would reduce the value of the results and undermine the power to analyze interaction effects.

[12] 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).

[13] The model fit of a three-factor solution is good: RMSEA=0.055, TLI=0.963.

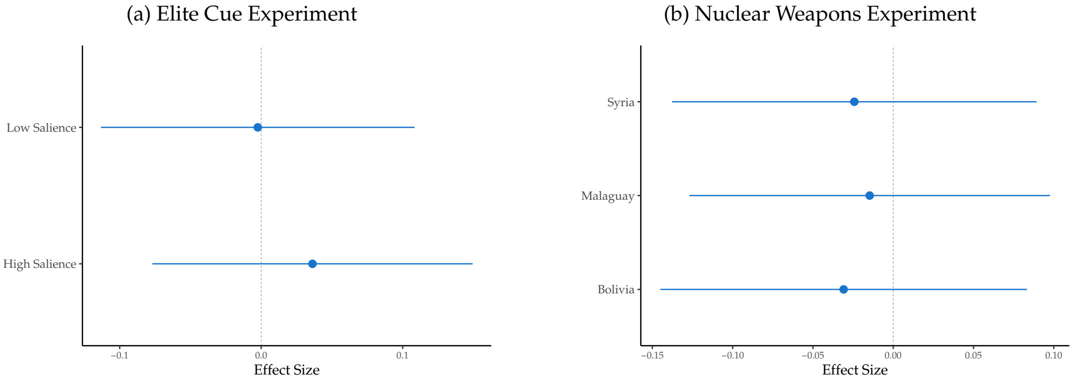
- **Countries outside the Middle East:** Ecuador, Bolivia, Vietnam, Ethiopia, Kyrgyzstan, Malaysia
- **Middle Eastern Adversaries:** Iran and Syria
- **Middle Eastern Others:** Egypt and Turkey

We therefore build 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.

E *Actor Identities and Cognitive Burden and Treatment Recall*

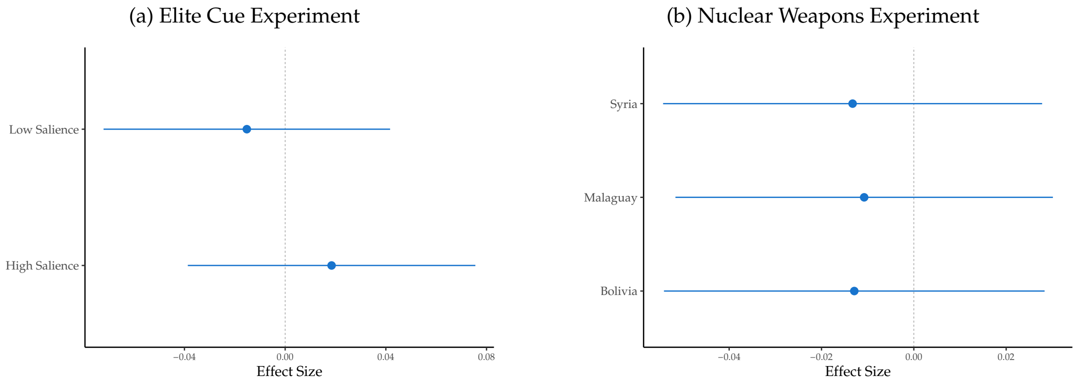
In this section we present results of additional analyses relating to the ELITE CUE and NUCLEAR WEAPONS experiments. Specifically, in Figure E.11, we consider how the salience (and type) of an elite cue-giver, influences cognitive burden during the experimental procedure (measured by response latency). In Figure E.12, we further consider the effects of actor identity on treatment recall. Generally, we do not find evidence that actor type impacts cognitive burden or treatment recall in both the ELITE CUE and NUCLEAR WEAPONS experiments.

Figure E.11: Actor Identity Effects on Response Time



Panel (a) of Figure E.11 demonstrates that moving from a hypothetical actor to a low or high salience actor does not impact the cognitive burden of respondents (measured by logged response latency). Similarly, Panel (b) of Figure E.11 demonstrates that switching from an unnamed to a made up or real world country 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 (correctly responding to the treatment recall question), is regressed over the actor identity treatment.

Figure E.12: Actor Identity Effects on Treatment Recall


(a) Elite Cue Experiment

(b) Nuclear Weapons Experiment

Panel (a) of Figure E.12 demonstrates that moving from a hypothetical actor to a low or high salience actor does not impact respondents' ability to correctly recall treatment. Similarly, panel (b) of Figure E.12 demonstrates that switching from an unnamed to a made up or real world country does not impact responses 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.

F *Additional Context and Response Time*

In this Section, we consider how adding more context into vignettes affects response time. We find, in Figure F.13, strong evidence that longer vignettes increase cognitive burden measured by response latency. Increased cognitive burden, can further explain why respondents assigned to longer vignettes are less likely to correctly recall treatment (Figure 5 in the main text).

Figure F.13: Additional Context Effects on Response Time (Nuclear Weapons and In-Group Favoritism Experiments)

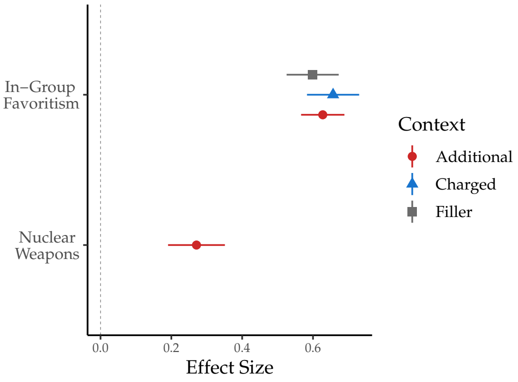


Figure F.13 demonstrates increasing context in experimental vignettes increases cognitive burden of respondents (measured by logged response latency). Point estimates and corresponding confidence intervals are extracted from separate OLS models where the dependent variable (response time for main outcome variable), is regressed over the actor identity treatment.

G *Situational Hypotheticality, Response Time, and Treatment Recall*

In this section we examine whether situational hypotheticality affects response time and treatment recall success. To do so, we run additional models where we regress a measure of response time or treatment recall, over a binary variable taking the value of 1 if an experiment is introduced as explicitly hypothetical. Results reported in Figure G.14, suggest that situational hypotheticality does not affect response time in all three experiments. Results reported in Figure G.15 suggest that situational hypthecticality has a null effect on treatment recall in the Elite Cues and Nuclear Weapons experiments. We do however, identify a small and marginally significant positive effect of situational hypotheticality on treatment recall in the In-Group Favoritism experiment. Given the magnitude of this effect, and the fact that situational hypotheticality does not moderate average treatment effects on our main outcomes, we suggest that varying levels of situational hypotheticality should not alter the substantive conclusions that researchers draw in their experimental studies.

Figure G.14: Situational Hypotheticality Effects on Response Time (ELITE CUES, IN-GROUP FAVORITISM, and NUCLEAR WEAPONS Experiments)

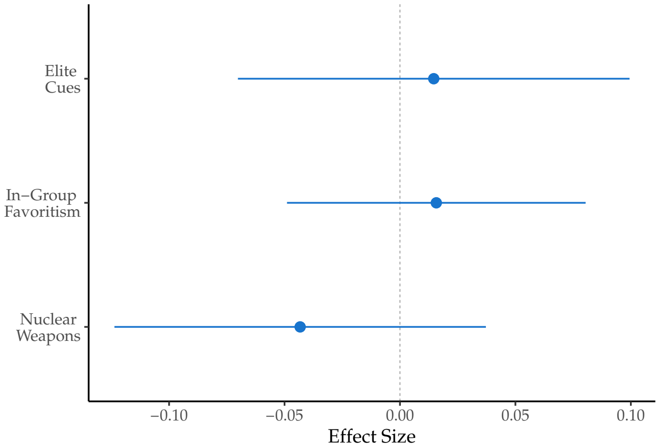


Figure G.14 demonstrates that introducing an experimental vignette as explicitly hypothetical does not affect the cognitive burden of respondents (measured by logged response latency). Point estimates and corresponding confidence intervals are extracted from separate OLS models where the dependent variable (response time for main outcome variable), is regressed over the situational hypotheticality treatment.

H Do Different Dimensions of Abstraction and Detail Interact?

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 H.16, we present models where we consider the moderating effects of country names 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 moderat-

Figure G.15: Situational Hypotheticality Effects on Treatment Recall (ELITE CUES, IN-GROUP FAVORITISM, and NUCLEAR WEAPONS Experiments)

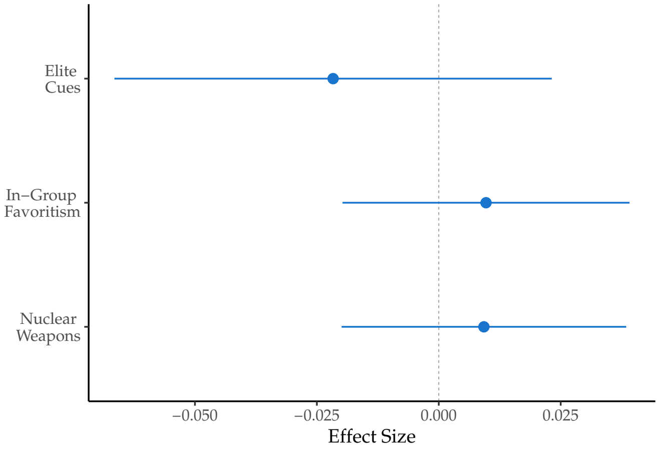


Figure G.15 demonstrates that introducing an experimental vignette as explicitly hypothetical does not affect treatment recall in the ELITE CUES and NUCLEAR WEAPONS experiments, but has a small and marginally significant positive effect on treatment recall in the IN-GROUP FAVORITISM experiment. Point estimates and corresponding confidence intervals are extracted from separate OLS models where the dependent variable (correct treatment recall), is regressed over the situational hypotheticality treatment.

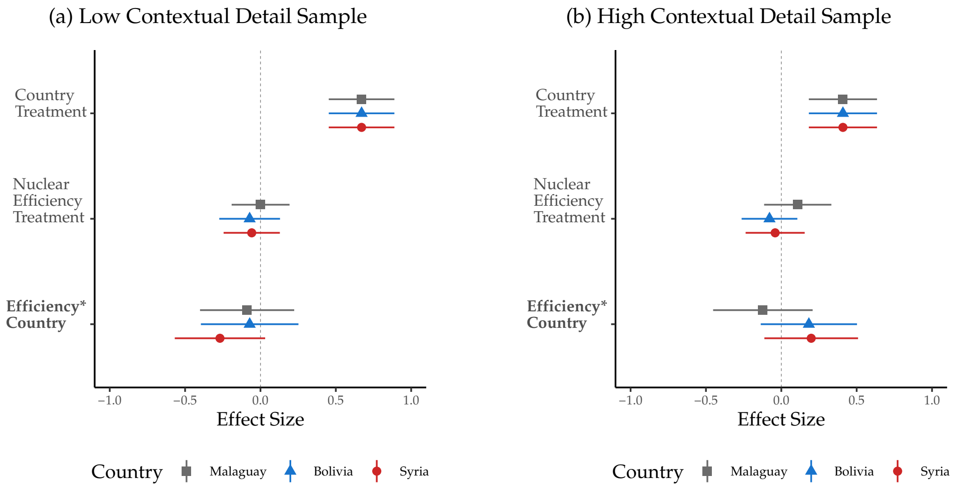
ing effect, which approaches conventional levels of statistical significance ($p < 0.08$) will not lead experimenters to draw substantively different inferences.

To further investigate the additive effect of abstraction and detail along different dimensions of our framework, we crated additive abstraction scores detailing the levels of abstraction and detail to which a subject was assigned (in any given vignette). This score is comprised of up to three dimensions: Situational hypothetically, actor identity and contextual detail, depending on the type of abstraction manipulated in any given study. Higher values denote more detailed and realistic experiments.

For example, if a respondent was assigned to a NUCLEAR WEAPONS vignette which was described as explicitly hypothetical, and the vignette included an un-named country and minimal context, than the respondents' corresponding abstraction score would be 0. Moving up in the ladder of detail on any one of our conceptual dimensions would increase this score. Thus, being assigned to a non-explicitly hypothetical vignette would increase the score by one point. Similarly, variation in our actor identity condition could increase the score by up to three points (because respondents were assigned to four conditions), and additional context can also increase the score by one point.

In Figure H.17, we test whether our abstraction scale moderates original ATEs. We find that overall levels of abstraction have a sharp null effect in our ELITE CUES and NUCLEAR WEAPONS experiments. In addition, the scale has a modest albeit statistically significant attenuating effect on the ATE of our IN-GROUP FAVORITISM experiment. Given the results reported in the main text, we expect this attenuation in the IN-GROUP FAVORITISM experiment to be driven, largely, by additional context which reduces the dosage of original treatments vis-a-vis background information.

Figure H.16: Moderating Effects of Country Name by Contextual Detail Subsamples


(a) Low Contextual Detail Sample

(b) High Contextual Detail Sample

Figure H.16 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.

Figure H.17: Moderating Effects of Abstraction Scale on all ATEs

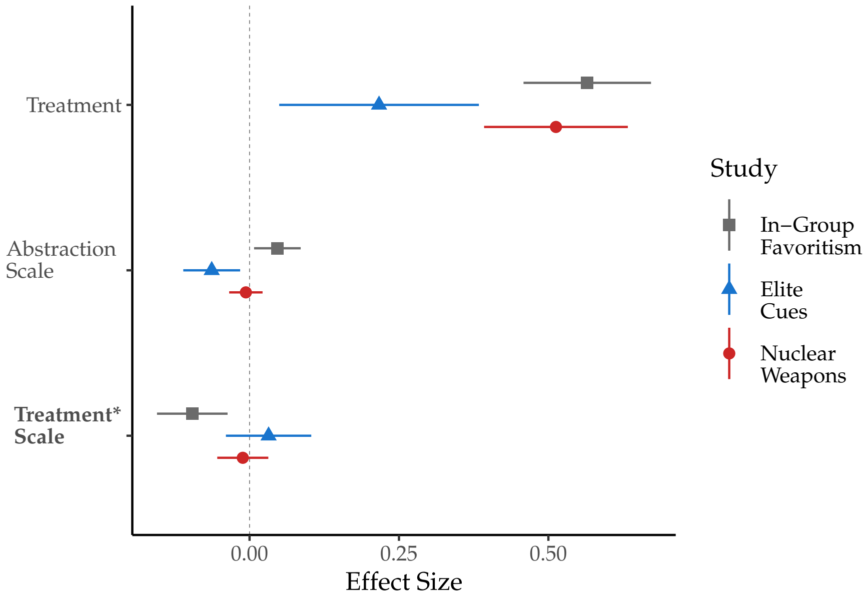


Figure H.17 demonstrates the limited moderating effects of our abstraction scale, on original ATEs. Point estimates and corresponding confidence intervals are extracted from separate OLS models where original outcomes are regressed over study treatments interacted with our abstraction scale.

I Individual Level Heterogeneity in Responses to Abstraction and Detail

In this section, we explore the extent to which different types of survey respondents react differently to levels of abstraction and detail in experimental design. To do so, we focus on two individual differences of theoretical relevance to the study of survey responses: political knowledge, and need for cognition. For ease of interpretation in the analysis below, we re-estimate our models from the main text on separate subsamples of respondents, mean-splitting by levels of political knowledge, and need for cognition, respectively. Although this facilitates ease of interpretation, it also reduces our sample size in each analysis, such that we encourage readers to take some caution when interpreting these additional results. In general, though, we find stronger results for political knowledge than we do for need for cognition. In particular, we find that contextual detail attenuates treatment effects for both high and low-knowledge respondents alike, and that high knowledge respondents are perhaps more sensitive to the use of high salience actors, but find few consistent patterns of differences between low and high-cognition respondents.

I.1 Political Knowledge

In our Nuclear Weapons and In-Group Favoritism experiments, we measured political knowledge with two multiple choice questions regarding: i) the identity of the United Kingdom's current prime-minister, and ii) the length of U.S. House of Representative terms for office. In the Elite Cues experiment, we added a third question regarding the identity of Israel's current prime-minister. Based on these questions, we split our sample in two based on whether respondents scored above the mean level of political knowledge.

In Figure I.18, we report results from models which consider the moderating effects of context in the In-Group Favoritism experiment, amongst two samples of respondents with high and low political knowledge. Given the small sample size, some care should be taken in interpreting the results, but the plot shows that the moderating effect of additional context is negatively signed for both high and low-knowledge respondents. The charged context has a slightly stronger negative effect, and attains statistical significance among high knowledge respondents. In Figure I.19, we conduct the same exercise for the Nuclear Weapons experiment. Again, the moderating effect of additional context is negatively signed for both high and low knowledge respondents. Here, the moderating effect attains statistical significance among low knowledge respondents, but the point estimates are similar in both instances.

Figure I.18: Moderating Effects of Context By Political Knowledge in In-Group Favoritism Experiment

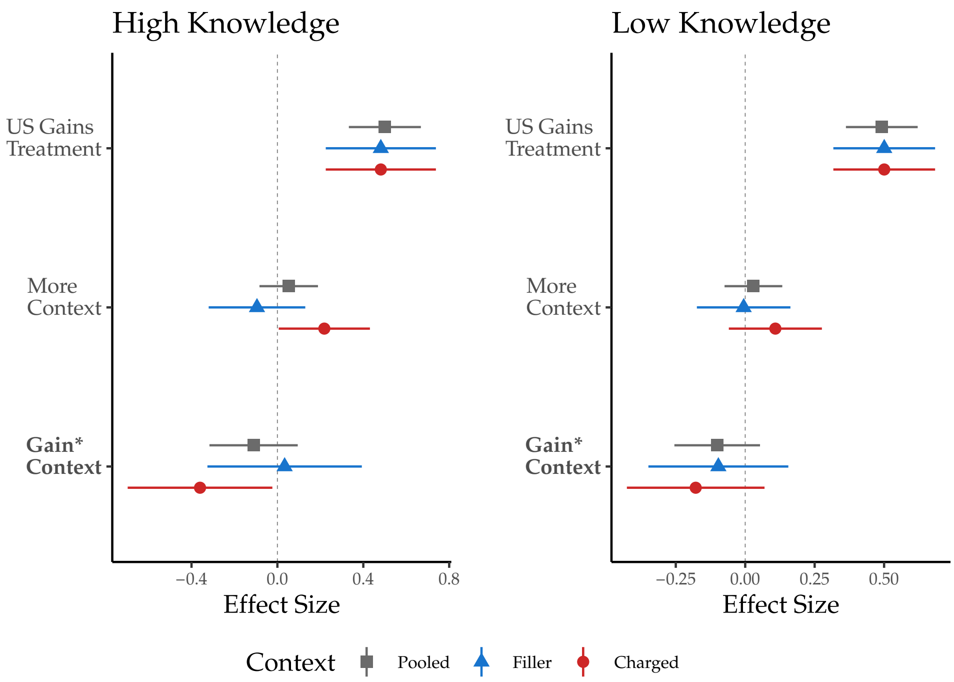


Figure I.18 demonstrates similar patterns for the moderating effect of context on original ATEs in the In-Group Favoritism experiment between high and low-knowledge respondents. Point estimates and corresponding confidence intervals are extracted from separate OLS models where original outcomes are regressed over study treatments interacted with a context indicator.

We therefore obtain relatively similar findings for the moderating effects of contextual detail between respondents high and low in political knowledge. In Figure I.20, we shift towards actor identity, examining the moderating effect of country name in the Nuclear Weapons experiment amongst our two subsamples. We find that for high political knowledge subjects, all country names (compared with an unnamed country) do not significantly moderate original average treatment effects. A somewhat similar pattern emerges when focusing on subjects with low political knowledge. Nonetheless, it appears that employing Bolivia as a country (rather than an unnamed country), has a small and unexpectedly positive moderating impact on the original average treatment effect for low knowledge respondents, though given the small sample size, some caution should once again be taken in interpreting the result.

Figure I.19: Moderating Effects of Context By Political Knowledge in Nuclear Weapons Experiment

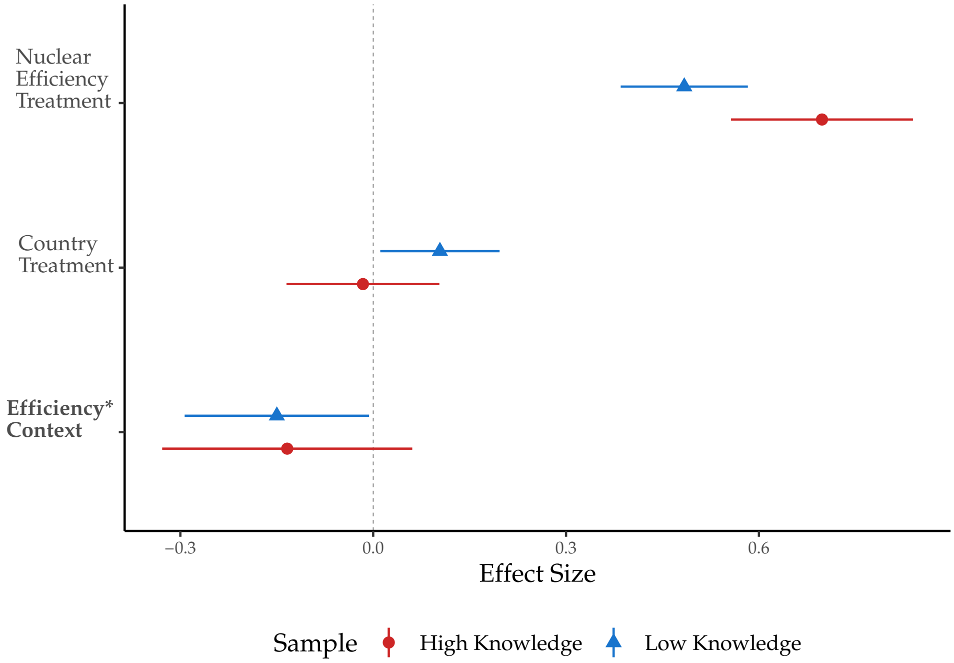


Figure I.19 demonstrates similar patterns for the moderating effect of context on original ATEs in the Nuclear Weapons experiment between high and low-knowledge respondents. Point estimates and corresponding confidence intervals are extracted from separate OLS models where original outcomes are regressed over study treatments interacted with a context indicator.

Figure I.20: Moderating Effects of Country By Political Knowledge in Nuclear Weapons Experiment

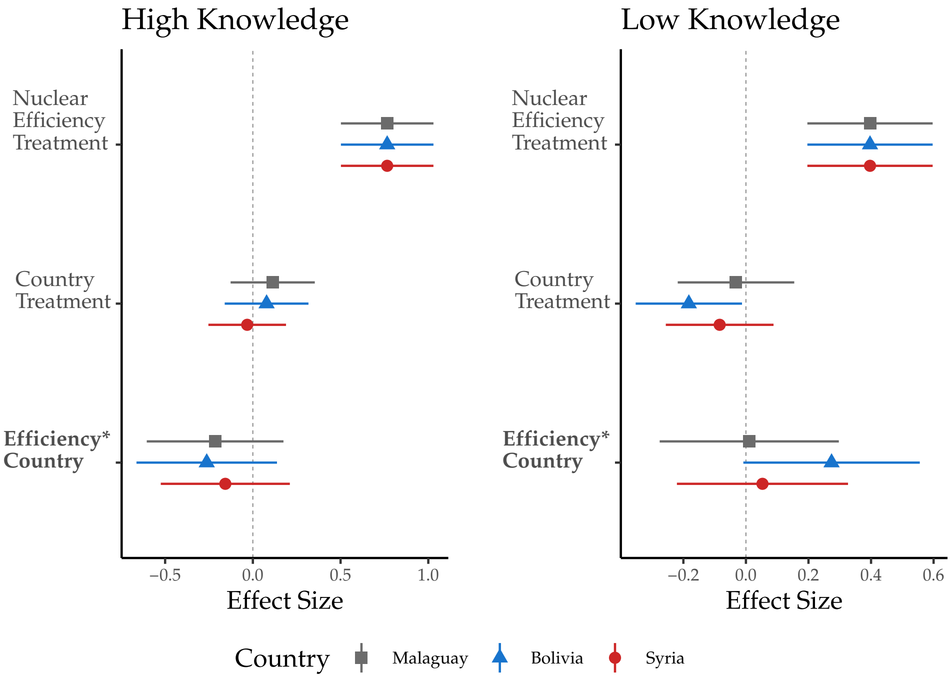


Figure I.20 demonstrates slight differences of the moderating effect of some country names on original ATEs, when focusing on two samples of subjects with low and high political knowledge in the Nuclear Weapons experiment. Point estimates and corresponding confidence intervals are extracted from separate OLS models where original outcomes are regressed over study treatments interacted with a context indicator.

In Figure I.21, we continue examining the moderating effects of actor identity, this time focusing on the moderating effect of high and low salience actors, amongst subjects with low and high levels of political knowledge. As a reminder, in our main analysis, we find that employing high salience actors has a positive moderating effect, increasing the size of the average treatment effects. When splitting our samples, we do not find much evidence that employing low or high salience actors (compared with made up actors) moderates effects amongst low knowledge subjects. However, for higher knowledge respondents, who presumably have stronger priors about real-world political figures, the employing high salience actors does have a positive and statistically significant moderating effect.

Figure I.21: Moderating Effects of Actor Identity By Political Knowledge in Elite Cue Experiment

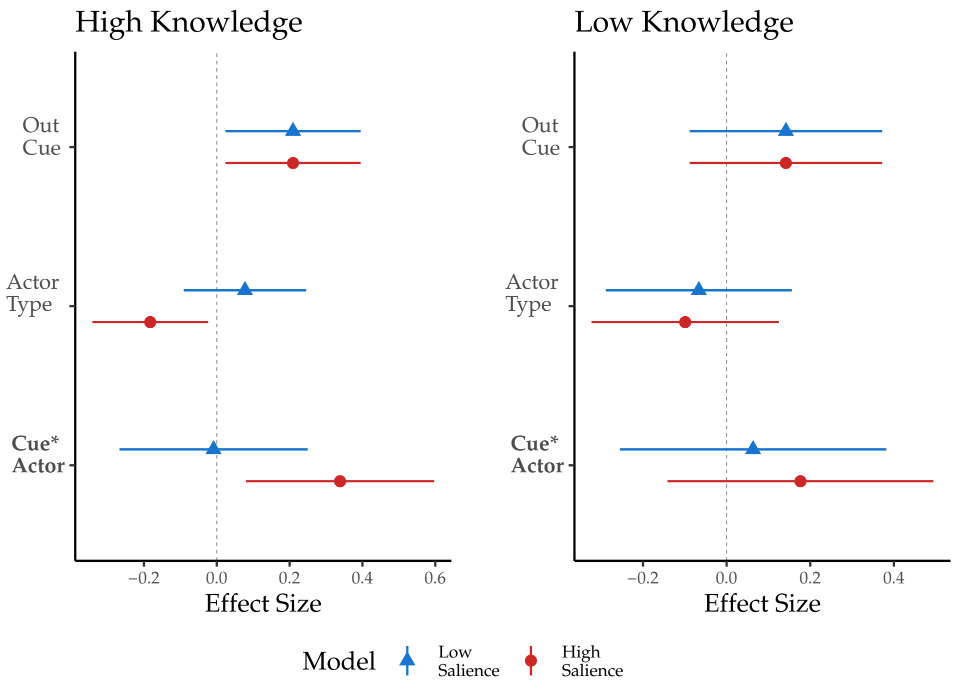


Figure I.21 suggests the persuasive effect of cues from high salient actors may be driven by higher knowledge respondents. Point estimates and corresponding confidence intervals are extracted from separate OLS models where original outcomes are regressed over study treatments interacted with a context indicator.

I.2 Need For Cognition

We next examine our results amongst two subsets of respondents with high and low levels of need for cognition. To do so, we utilized a shorter-form version of the need for cognition scale based on 14 commonly used questions (Cacioppo and Petty, 1982; Rathbun, Kertzer and Paradis, 2017). We mean-split this index to create two subsamples of respondents, based on whether they display high (above average) or low (below average) levels of need for cognition. Because we did not the need for cognition item in the dispositional battery that accompanied the *Elite Cue* experiment, our analysis below focuses onthe In-Group Favoritism and Nuclear Weapons Experiments.

In Figure I.22, we consider whether the moderating effects of context in the In-Group Favoritism experiment, vary between respondents with low and high levels of need for cognition. We find that varying levels of context does not affect respondents with high need for cognition. However, amongst respondents with low need for cognition, additional context, especially when charged, seems to attenuate average treatment effects. When turning to the Nuclear Weapons experiment in Figure I.23, we find that additional context negatively moderates average treatment effects, but this moderating effect is statistically significant only for respondents with higher levels of need for cognition. Given the mixed findings reported from the In-Group Favoritism and Nuclear Weapons experiments, we suggest that readers take these results with a grain of salt, and encourage future research to further examine the relationship between need for cognition and context.

Finally, in Figure I.24 we consider the extent to which the moderating effect of country name varies across our subsamples of respondents with low and high levels of need for cognition. We once again find little evidence for a consistent pattern. Indeed, it appears that country names are unlikely to moderate average treatment effects for respondents with low and high need for cognition.

Figure I.22: Moderating Effects of Context By Need For Cognition in In-Group Favoritism Experiment

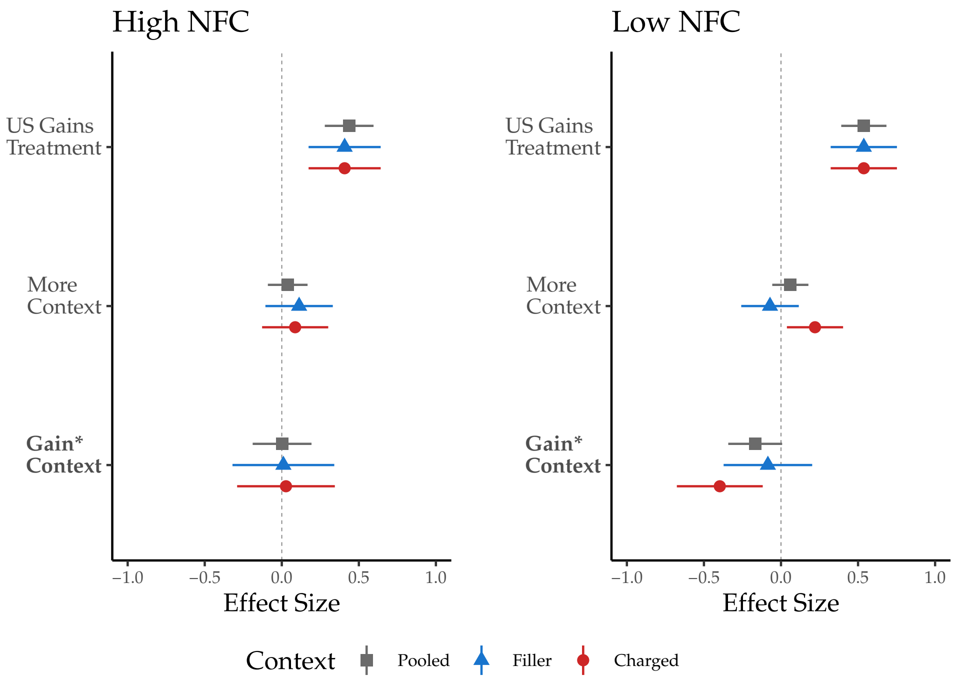


Figure I.22 demonstrates slight differences of the moderating effect of additional context on original ATEs, when focusing on two samples of subjects with low and high need for cognition in the IN-GROUP FAVORITISM experiment. Point estimates and corresponding confidence intervals are extracted from separate OLS models where original outcomes are regressed over study treatments interacted with a context indicator.

Figure I.23: Moderating Effects of Context By Need For Cognition in In-Group Nuclear Weapons Experiment

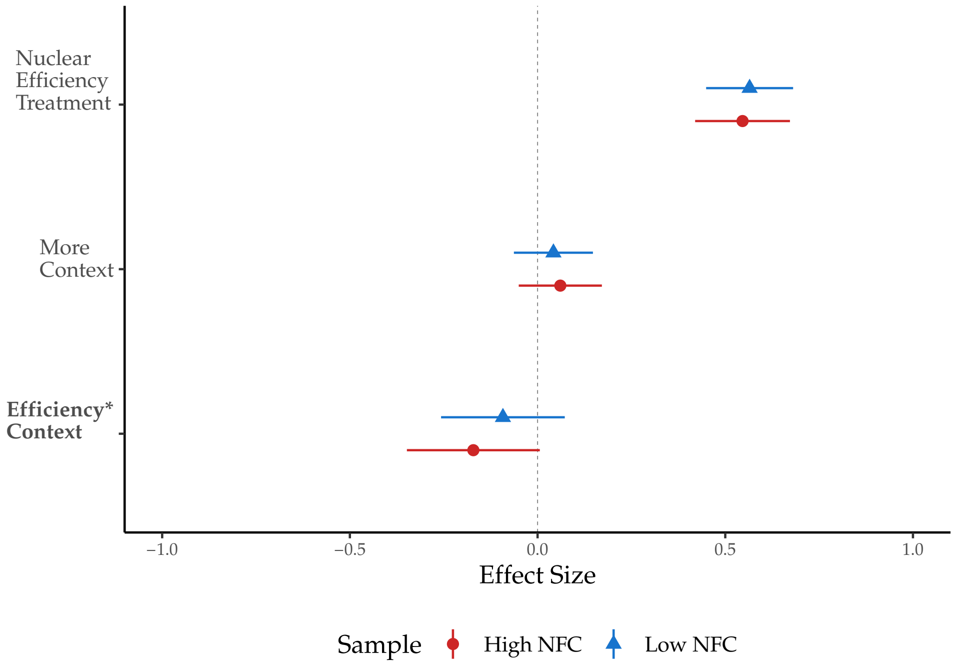


Figure I.23 demonstrates slight differences of the moderating effect of additional context on original ATEs, when focusing on two samples of subjects with low and high need for cognition in the NUCLEAR WEAPONS experiment. Point estimates and corresponding confidence intervals are extracted from separate OLS models where original outcomes are regressed over study treatments interacted with a context indicator.

Figure I.24: Moderating Effects of Country By Need For Cognition in In-Group Nuclear Weapons Experiment

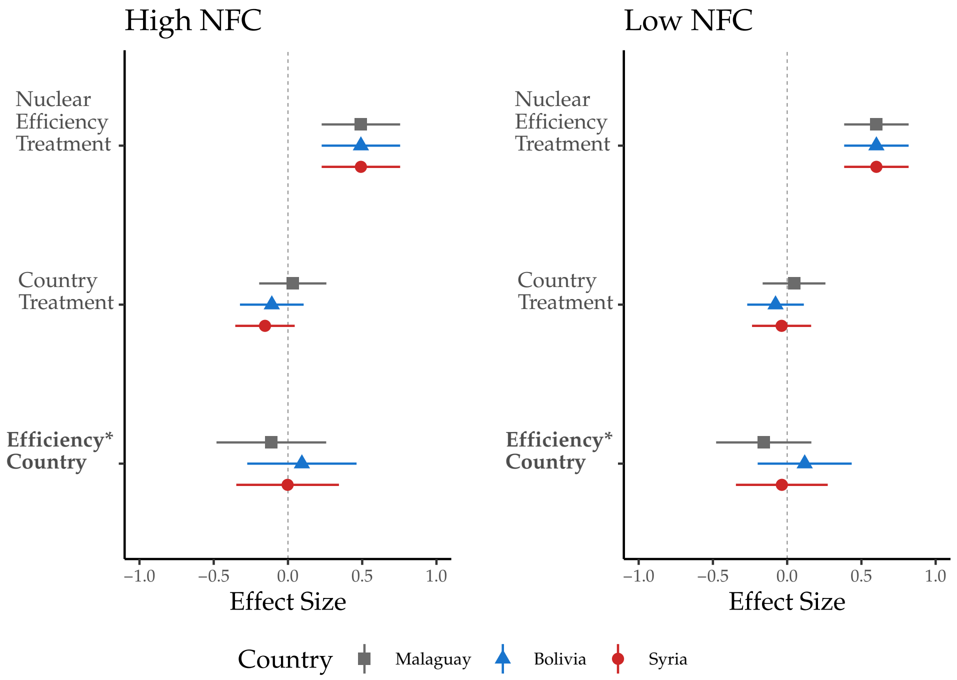


Figure I.24 demonstrates limited differences of the moderating effect of country on original ATEs, when focusing on two samples of subjects with low and high need for cognition in the NUCLEAR WEAPONS experiment. Point estimates and corresponding confidence intervals are extracted from separate OLS models where original outcomes are regressed over study treatments interacted with a context indicator. The high need for cognition sample includes respondents whose score on a need for cognition scale is above average, whereas low need for cognition sample includes respondents whose score on a need for cognition scale is below average.

Table 5: Replication of ATEs from Three Experiments

	Elite Cues		In-Group Favoritism	Nuclear Weapons	
	Model 1	Model 2	Model 3	Model 4	Model 5
Out-party Cue	0.25^{*} (0.06)				
Out-party Cue (Original)		0.32 (0.17)			
U.S. Gains			0.50^{*} (0.05)		
Nuclear Effectiveness				0.47^{*} (0.09)	
Nuclear Effectiveness (Original)					0.57^{*} (0.11)
Num. obs.	1151	240	1507	535	319

$^{*}p < 0.05$. In Replication Study we analyze a subset of data that resembles the abstraction level of the original study.

J Regression tables

To preserve space, in the main text we present our results graphically; the companion regression tables are presented below. First, in Table 5 we present results from Figure 1, where we replicate the results from the NUCLEAR WEAPONS, ELITE CUES, and IN-GROUP FAVORITISM experiments. Second, in Table 6, we present results from Figure 2, considering the moderating effect of situational hypotheticality on original average treatment effects of all studies. Third, in Table 7, we present results from Figure 3 of the main text, considering the moderating effects of actor identity in the NUCLEAR WEAPONS and ELITE CUES experiment. Fourth, in Table 8, we report results from Figure 4 in the main text, which considers how additional context attenuates average treatment effects in the NUCLEAR WEAPONS and IN-GROUP FAVORITISM experiments. Finally, in Table 9, we present results from Figure 5 of the main text, which show how additional context in experimental vignettes reduces success in treatment recall questions.

Table 6: No moderating effects of situational hypotheticality

	Elite Cues	In-Group Favoritism	Nuclear Weapons
Out-party Cue	0.19*		
	(0.07)		
U.S. Gains		0.38*	
		(0.04)	
Nuclear Effectiveness			0.50*
			(0.04)
Hypothetical	0.11	-0.04	0.03
	(0.07)	(0.04)	(0.04)
Cue*Hypothetical	0.13		
	(0.10)		
Gain*Hypothetical		0.09	
		(0.06)	
Effectiveness*Hypothetical			-0.03
			(0.06)
Num. obs.	1633	4491	4462

$^{*}p < 0.05$. In the elite cue experiment, we omit respondents who were assigned to a baseline condition, where scenarios weren't described as hypothetical or real.

Table 7: Moderating effects of actor identity condition

	Elite Cues		Nuclear Weapons		
	Model 1	Model 2	Model 3	Model 4	Model 5
Out-group Cue	0.19*	0.19*			
	(0.07)	(0.07)			
Low Salience	0.02	0.02			
	(0.07)	(0.07)			
High Salience		-0.14^{*}			
		(0.07)			
Cue*Low Salience	0.01	0.01			
	(0.10)	(0.10)			
Cue*High Salience		0.27*			
		(0.10)			
Nuclear Effectiveness			0.55*	0.55*	0.55*
			(0.08)	(0.08)	(0.08)
Malaguay			0.04		
			(0.07)		
Bolivia				-0.07	
				(0.07)	
Syria					-0.05
					(0.07)
Effective*Malaguay			-0.10		
			(0.12)		
Effective*Bolivia				0.05	
				(0.12)	
Effective*Syria					-0.04
					(0.11)
Num. obs.	1622	2398	1159	1122	1167

$^{*}p < 0.05$

Table 8: Adding contextual detail attenuates treatment effects

	In-Group Favoritism			Nuclear Weapons
	Model 1	Model 2	Model 3	Model 4
U.S. Gains	0.51*	0.50*	0.50*	
	(0.07)	(0.05)	(0.05)	
Filler	−0.01			
	(0.07)			
Charged		0.10*		
		(0.05)		
Pooled			0.05	
			(0.04)	
Gain*Filler	−0.07			
	(0.10)			
Gain*Charged		−0.20*		
		(0.07)		
Gain*pooled			−0.11	
			(0.06)	
Nuclear Effectiveness				0.56*
				(0.04)
Additional Context				0.06
				(0.04)
Effectiveness*Context				−0.15*
				(0.06)
Num. obs.	1511	2982	4491	4462

$^*p < 0.05$. In models 1-2, we compare charged and filler to a control condition. In model 3 we pool both conditions, and compare to control condition.

Table 9: Contextual Detail Effects on Treatment Recall Success

	In-Group Favoritism		Nuclear Weapons
	Model 1	Model 2	Model 3
Charged	−0.13*		
	(0.02)		
Filler		−0.08*	
		(0.02)	
Additional Context			−0.03*
			(0.01)
Num. obs.	2946	2971	4395

$^*p < 0.05$

References

Bansak, Kirk, Jens Hainmueller, Daniel J. Hopkins and Teppei Yamamoto. 2021. "Beyond the breaking point? Survey satisficing in conjoint experiments." *Political Science Research and Methods* 9(1):53–71.

Bisgaard, Martin and Rune Slothuus. 2018. "Partisan elites as culprits? How party cues shape partisan perceptual gaps." *American Journal of Political Science* 62(2):456–469.

Blair, Graeme, Jasper Cooper, Alexander Coppock and Macartan Humphreys. 2019. "Declaring and diagnosing research designs." *American Political Science Review* 113(3):838–859.

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.

Brutger, Ryan, Joshua D. Kertzer, Jonathan Renshon and Chagai M. Weiss. 2022. "Abstraction in Survey Experiments: Testing the Tradeoffs." Book manuscript.

Bullock, John G. 2011. "Elite Influence on Public Opinion in an Informed Electorate." *American Political Science Review* 105(3):496–515.

Butler, Daniel M. and David E. Broockman. 2011. "Do Politicians Racially Discriminate Against Constituents? A Field Experiment on State Legislators." *American Journal of Political Science* 55(3):436–477.

Cacioppo, John T and Richard E Petty. 1982. "The need for cognition." *Journal of personality and social psychology* 42(1):116.

Chandler, Jesse, Pam Mueller and Gabriele Paolacci. 2014. "Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers." *Behavior research methods* 46(1):112–130.

Chu, Jonathan A. 2019. "A clash of norms? How reciprocity and international humanitarian law affect American opinion on the treatment of POWs." *Journal of Conflict Resolution* 63(5):1140–1164.

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.

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.

Guisinger, Alexandra and Elizabeth N. Saunders. 2017. "Mapping the Boundaries of Elite Cues: How Elites Shape Mass Opinion Across International Issues." *International Studies Quarterly* 61(2):425–441.

Habyarimana, James, Macartan Humphreys, Daniel N Posner and Jeremy M Weinstein. 2007. "Why does ethnic diversity undermine public goods provision?" *American Political Science Review* 101(4):709–725.

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.

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.

Iyengar, Shanto and Sean J Westwood. 2015. "Fear and loathing across party lines: New evidence on group polarization." *American Journal of Political Science* 59(3):690–707.

Jung, Jae-Hee. 2020. "The Mobilizing Effect of Parties' Moral Rhetoric." *American Journal of Political Science* 64(2):341–355.

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

Kane, John V and Jason Barabas. 2019. "No harm in checking: Using factual manipulation checks to assess attentiveness in experiments." *American Journal of Political Science* 63(1):234–249.

Kanthak, Kristin and Jonathan Woon. 2015. "Women Don't Run? Election Aversion and Candidate Entry." *American Journal of Political Science* 59(3):595–612.

Kertzer, Joshua D. and Brian C. Rathbun. 2015. "Fair is Fair: Social Preferences and Reciprocity in International Politics." *World Politics* 67(4):613–655.

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

Lenth, Russell V. 2007. "Statistical power calculations." *Journal of animal science* 85(suppl_13):E24–E29.

Lyall, Jason, Graeme Blair and Kosuke Imai. 2013. "Explaining Support for Combatants during Wartime: A Survey Experiment in Afghanistan." *American Political Science Review* 107(4):679–705.

Lyall, Jason, Yang-Yang Zhou and Kosuke Imai. 2020. "Can Economic Assistance Shape Combatant Support in Wartime? Experimental Evidence from Afghanistan." *American Political Science Review* 114(1):126–143.

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.

McClendon, Gwyneth H. 2018. *Envy in politics*. Princeton, NJ: Princeton University Press.

McDermott, Rose, Dominic Johnson, Jonathan Cowden and Stephen Rosen. 2007. "Testosterone and Aggression in a Simulated Crisis Game." *Annals of the American Academy of Political and Social Science* 614(1):15–33.

McDonald, Jared, Sarah E Croco and Candace Turitto. 2019. "Teflon Don or Politics as Usual? An Examination of Foreign Policy Flip-Flops in the Age of Trump." *The Journal of Politics* 81(2):757–766.

Mérola, Vittorio and Matthew P Hitt. 2016. "Numeracy and the persuasive effect of policy information and party cues." *Public Opinion Quarterly* 80(2):554–562.

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

Nelson, Thomas E., Rosalee A. Clawson and Zoe M. Oxley. 1997. "Media Framing of a Civil Liberties Conflict and Its Effect on Tolerance." *American Political Science Review* 91(3):567–583.

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

Nugent, Elizabeth. 2020. "The Psychology of Repression and Polarization." *World Politics* 72(2):291–334.

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.

Peyton, Kyle, Gregory A Huber and Alexander Coppock. 2021. "The generalizability of online experiments conducted during the COVID-19 pandemic." *Journal of Experimental Political Science* .

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.

Rathbun, Brian C, Joshua D Kertzer and Mark Paradis. 2017. "Homo diplomaticus: Mixed-method evidence of variation in strategic rationality." *International Organization* pp. S33–S60.

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.

Ryan, Timothy J. 2019. "Actions versus consequences in political arguments: Insights from moral psychology." *The Journal of Politics* 81(2):426–440.

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

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

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

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