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


Word count: 8363 words in text + 2829 in bibliography + 686 in captions = 11841 words
Experimentalists in political science often face a question about how abstract or concrete their experimental stimuli should be. This question is typically thought of in terms of tradeoffs relating to experimental control and generalizability. At one end of the continuum, researchers in the economics tradition for example, tend to 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). This tradition is particularly popular in behavioral experiments in political science seeking to test the predictions of formal models (e.g., Dickson, 2009; Dawes, Loewen and Fowler, 2011; Tingley and Walter, 2011a,b; Kanthak and Woon, 2015; Kertzer and Rathbun, 2015; LeVeck and Narang, 2017; Quek, 2017; Hundley, 2020), but also arises in survey experimental work in political science as well (e.g. Renshon, 2015; Mutz and Kim, 2017; Tingley, 2017).

At the other end of the continuum, a different tradition originating in psychology has encouraged the use of rich and detailed vignette-based experiments. This approach, too, has found a foothold in political science (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; Bracic and Murdie, 2019; Tomz, Weeks and Yarhi-Milo, 2020). 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 doing so reduces experimental control (Camerer, 1997). Further, it has been our experience that this issue is com-

---

1 How political scientists do navigate these tradeoffs may have more to do with institutional path dependence rather than verifiable data or theoretical frameworks.
pounded in the publication process, where reviewers often have very strong priors about how to best navigate these tradeoffs.

And yet, as a discipline we know relatively little about the tradeoffs inherent in abstract versus concrete experimental designs. Certainly, increasing “color in the laboratory” may trigger “unknown (to the experimenter) impressions and memories of past experiences over which the experimenter has no control” (Friedman, Friedman and Sunder, 1994), but it is not obvious why sparse experiments would fare better in this respect. In fact, a review of the broader experimental literature suggests strong disagreement on which would be a bigger problem in terms of respondents “filling in the blanks”: rich, detailed experiments (e.g., Friedman, Friedman and Sunder, 1994) or abstract, sparse studies (e.g., Alekseev, Charness and Gneezy, 2017). And 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 different dimensions of abstraction implicated in experimental design: situational hypotheticality, actor identity, and contextual detail. We argue that there are certain types of questions where ethical or feasibility considerations mandate at least some form of hypotheticality or abstraction, while there are others where scholars have more leeway. Yet, in those cases where scholars do have leeway, we argue that the tradeoffs between abstraction and detail in experimental design are not as stark as political scientists often claim. Second, like other recent work seeking to subject conventional wisdom about experimental design principles to empirical scrutiny (Mullinix et al., 2015; White et al., 2018; Coppock, 2019; Lupton, 2019; Mummolo and Peterson, 2019; Kertzer, 2020), we test our theoretical framework, replicating three popular vignette-based survey experiments in political science, and manipulating their levels of abstraction in three different ways. We find no evidence that situational hypotheticality changes the results experimenters obtain, an important finding as Institutional Review Boards (IRBs) and our field more broadly become increasingly concerned about the use of deception. Whether with politicians in American politics experiments, or countries in International Relations experiments,
we generally find little evidence that varying the abstraction of actor identities changes experimental results, although cue-taking experiments that use real and highly salient cuegivers obtain stronger effects than those that use fake ones. And, we show that adding contextual detail to experimental vignettes attenuates the size of treatment effects because respondents are less likely to be able to recall the treatment, but doesn’t change the substantive conclusions reached. Altogether our results suggest that there are fewer tradeoffs between abstraction and detail in survey experiment design than political scientists often assume.

1 Abstraction and detail

One of the many design choices political scientists face when using experimental methods concerns the appropriate level of abstraction. There is a rich literature on abstraction in philosophy, psychology, and cognitive science, which often operationalizes abstraction in slightly different ways (e.g., Cantor and Mischel, 1979; Paivio, 1990; Semin and Fiedler, 1988; Colburn and Shute, 2007). For our purposes, we borrow from construal level theory in defining abstraction as a higher-level representation (Sartori, 1970, 1040-46; Trope and Liberman, 2003). It involves making “a distinction between primary, defining features, which are relatively stable and invariant, and secondary features, which may change with changes in context and hence are omitted from the higher-level representation” (Shapira et al., 2012, 231). As Table 1 shows, an abstract representation is sparse and decontextualized, reduced to the object’s most central elements (e.g., “A nuclear weapon”), whereas a concrete representation is contextualized and rich in specific detail, including subordinate considerations (e.g., “North Korea’s Hwasong-14 intercontinental ballistic missile”).

Table 1: Conceptualizing abstraction

<table>
<thead>
<tr>
<th>Abstract</th>
<th>Concrete</th>
</tr>
</thead>
<tbody>
<tr>
<td>High level of construal</td>
<td>Low level of construal</td>
</tr>
<tr>
<td>Decontextualized</td>
<td>Contextualized</td>
</tr>
<tr>
<td>Primary features</td>
<td>Secondary features</td>
</tr>
<tr>
<td>Sparse</td>
<td>Complex</td>
</tr>
<tr>
<td>Superordinate elements</td>
<td>Subordinate elements</td>
</tr>
</tbody>
</table>

Modified from Trope and Liberman (2003, 405).

Two points here are relevant for our purposes. First, in specifying which elements of a construct are primary and which are secondary, the act of abstraction is inherently a theoretical phenomenon,
rather than an empirical one. Although our interest in this article is in abstraction in experimental methods, similar questions also arise in formal modeling, and in quantitative and qualitative methods alike (Clarke and Primo, 2012).

Second, questions about the appropriate level of abstraction loom large in a variety of issues in experimental design: whether experiments should be “stylized” or “contextually rich” (Dickson, 2011; Kreps and Roblin, 2019), use real or hypothetical actors (McDonald, 2019; Nielson, Hyde and Kelley, 2019), and refer to imminent, future, or hypothetical situations. In this sense, experiments can be abstract or concrete along multiple dimensions at the same time. In the discussion below, we suggest that abstraction in experimental design can be conceptualized along at least three dimensions: situational hypotheticality, actor identity, and contextual detail. We classify a set of recent experiments in political science along these dimensions in Table 2, and review each dimension in detail in the discussion below.

Table 2: Abstraction in experimental political science

<table>
<thead>
<tr>
<th>Type of experiment</th>
<th>Type of experiment</th>
<th>Example</th>
<th>Situational Hypotheticality</th>
<th>Actor Identity</th>
<th>Contextual Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audit experiment</td>
<td>Butler and Broockman (2011)</td>
<td>Deception</td>
<td>N/A</td>
<td>Med</td>
<td></td>
</tr>
<tr>
<td>Econ-style lab experiment</td>
<td>Kanthak and Woon (2015)</td>
<td>Real</td>
<td>Unnamed</td>
<td>Low</td>
<td></td>
</tr>
<tr>
<td>Endorsement experiment</td>
<td>Lyall, Blair and Imai (2013)</td>
<td>Real</td>
<td>Real</td>
<td>Med</td>
<td></td>
</tr>
<tr>
<td>Framing experiment</td>
<td>Nelson, Clawson and Oxley (1997)</td>
<td>Deception</td>
<td>Real</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Lab-in-the-field experiment</td>
<td>Habyarimana et al. (2007)</td>
<td>Real</td>
<td>Unnamed</td>
<td>Low</td>
<td></td>
</tr>
<tr>
<td>War game</td>
<td>McDermott et al. (2007)</td>
<td>Simulation</td>
<td>Artificial</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Field Experiment</td>
<td>Lyall, Zhou and Imai (2020)</td>
<td>Real</td>
<td>Real</td>
<td>High</td>
<td></td>
</tr>
</tbody>
</table>

1.1 Situational Hypotheticality

The first type of abstraction in experimental design concerns whether a scenario is described as hypothetical or not.\(^2\) 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.,

\(^2\)Note that the relevant question here is whether the scenario is presented as hypothetical, rather than whether the details described in the scenario are in fact hypothetical; as we note below, experiments utilizing deception, for example, may present hypothetical scenarios as if they are real.
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.

There are a variety of techniques experimentalists often employ. One is to explicitly describe the scenario as hypothetical: respondents in Boettcher (2004, 344), for example, are asked to “en- vision a hypothetical presidency apart from the current administration.” Others implicitly invoke hypotheticality: respondents participating in conjoint experiments studying immigration preferences, for example (e.g., Hainmueller and Hopkins, 2015), are presumably not under the illusion that the immigrants they are being asked to choose between are real. Another widely used variant under the category of “implicit hypotheticality” is to describe a scenario as set in the future. Mattes and Weeks (2019), for example, tell respondents “We are going to describe a situation the U.S. could face in the future, in 2027.” This is often termed a prospective scenario, but ultimately the future setting is simply a mechanism to make the scenario implicitly hypothetical.

The rationale for these design choices are often not explicitly stated, but usually involve concerns that respondents will not take studies as seriously when scenarios are presented as explicitly hypothetical — the sense that researchers asking hypothetical questions will be rewarded with hypothetical answers (Converse and Presser, 1986, 23). Experimentalists operating out of an economics-style tradition tend to avoid both deception and situational hypotheticality in order to accentuate the effects of incentives (Morton and Williams, 2010). Yet, there is relatively little empirical work testing the conditions in which situational hypotheticality affects responses in political science experiments.

1.2 Actor identity

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

Moving up the ladder of abstraction, some experimenters describe hypothetical scenarios in artificial countries, in order to exert complete control over how much information participants bring to bear. For example, Brooks and Valentino (2011) describe a conflict between “Malaguay and Westria”, and Rubenzer and Redd (2010) describe a crisis in the state of “Gorendy.” Taking this approach a step forward, many experimentalists use unnamed countries, describing target states as “Country A” or “Country B” (Johns and Davies, 2012; Yarhi-Milo, Kertzer and Renshon, 2018), or simply referring to “A country” rather than providing a label (Tomz and Weeks, 2013).³

Concerns about actor identity and hypotheticality are not limited to the subfield of international relations. In comparative politics, Banerjee et al. (2014) describe hypothetical representatives (running for office in hypothetical districts) to study the concerns of voters in rural India. “Hypothetical candidate” experiments are also a long-running feature in the study of American politics (as in Rosenwasser et al., 1987; Colleau et al., 1990; Kam and Zechmeister, 2013) — and are particularly common in conjoint experiments — although the results here are mixed. In a meta-analysis of 111 studies of negative campaigning, Lau, Sigelman and Rovner (2007) find that experiments featuring hypothetical candidates don’t offer significantly different results from those featuring real ones. McDonald (2019), in contrast, argues that experiments on hypothetical candidates both increase cognitive burden and produce larger treatment effects than experiments on candidates about which respondents have strong priors.

As with the case of situational hypotheticality, the logic of using unnamed or hypothetical actors stems directly from the questions being tested. Political scientists turned to experimental methods to study the effects of candidate gender (Huddy and Terkildsen, 1993; Sanbonmatsu, 2002; 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; Wamble, 2020), or ethnicity

³In instructions to participants, researchers often note that this generality is “for scientific validity.”
Dunning and Harrison, 2010, though see Adida, 2015). In an IR context, it is hard to think of two real-world countries that are identical in all respects but one, such that IR scholars interested in manipulating the effects of regime type, military capabilities, or foreign policy interests usually do so with fictional or hypothetical countries (e.g., Rousseau and Garcia-Retamero, 2007).

1.3 CONTEXTUAL DETAIL

The third dimension of abstraction involves the amount of additional context provided in an experiment. Press, Sagan and Valentino (2013) present a lengthy newspaper article that provides participants with a large amount of context, as do experiments in American politics that generate fake campaign advertisements or news clips (Brader, Valentino and Suhay, 2008). In contrast, other experiments often present relatively little information. 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. Most experiments in the experimental economics tradition offer relatively little contextual detail at all, deliberately paring the design down to be void of explicit political content (e.g., Tingley and Walter, 2011b; Kanthak and Woon, 2015; Quek, 2017).

The argument usually offered in favor of contextual detail is that it increases realism and respondent engagement. Anecdotally, in some of our own research, when we include open-ended prompts at the ends of survey experiments soliciting feedback from participants, one of the most frequent comments we have received involved scenarios being “too vague”: participants wanted more information in order to help them make up their minds. Yet apart from Kreps and Roblin (2019) and Bansak et al. (2020), there has been little empirical work to adjudicate what the consequences of providing richer or sparser stimuli might be. Bansak et al. (2020) use a clever multi-stage conjoint design to first find “filler attributes” (information uncorrelated with the object of interest in the study) and then experimentally vary the amount of filler in the second stage, finding relatively stable treatment effects even with large numbers (up to 35) of filler items. Kreps and Roblin (2019) focus on treatment “formats,” in particular the difference between information presented in mock news stories versus short/long vignettes (with a slight conflation of the “length” versus “format” dimensions), finding that respondent attention (as a measure of satisficing) was unaffected by the presentational format.
This discussion suggests that what is often referred to as “contextual detail” is actually composed of at least three related dimensions. The first is simply the volume of information provided: more or less information can be provided in an experiment to supplement the treatments and add “realism.” We might, for example, provide a long or short biography of a candidate or background to an international dispute. 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 stories (Druckman and Nelson, 2003) and “ad-watch” style reports (Valentino, Neuner and Vandenbroek, 2018). The third is the content of the information itself, which is orthogonal to its volume. Any bit of information may be classified as either what Bansak et al. (2020) call “filler” or its opposite, what we 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. (2020) also show that other attributes (e.g., previous occupation and number of children) are associated with the object of interest and are thus ill-suited to be added simply to increase the “realism” of a vignette. But while they show that satisficing is less of a problem than we might expect once we introduce filler attributes, we are still largely in the dark with respect to understanding how the addition of charged (versus filler) content affects our interpretation of experimental results.

2 Navigating the tradeoffs

In sum, although political scientists tend to recognize that tradeoffs between abstract and concrete experiments exist, there is less certainty about how one should balance them. Often, for example, political scientists run both abstract and concrete versions of their experiment to test whether the results hold (e.g., Herrmann, Tetlock and Visser, 1999; Levine and Palfrey, 2007; Rousseau and Garcia-Retamero, 2007; Berinsky, 2009; Horowitz and Levendusky, 2011; LeVeck et al., 2014; Ren- shon, Dafoe and Huth, 2018; Nielson, Hyde and Kelley, 2019). However, doing so can be somewhat inefficient, as adjusting levels of abstraction on multiple dimensions simultaneously provides limited insight regarding the specific dimension driving experimental outcomes.4

There are some circumstances where for logistical or ethical reasons, experimenters will be constrained in terms of how abstract or concrete their stimuli will be. For example, researchers are

---

4Moreover, given the presence of a budget constraint, running two versions of the same experiment to preempt questions about abstraction can also lead to statistical power problems.
limited in their ability to select real world actors when studying the effects of race and gender in
candidate selection, or the effects of country-level characteristics on foreign policy preferences. Ad-
ditionally, there are experiments where some form of situational hypotheticality is required (often
at the demand of IRBs) to avoid the use of deception, and some contexts where the use of decept-
ion raises ethical challenges: for example, telling respondents that a real-world political candidate
is unethical (e.g., Butler and Powell, 2014). Experimentalists using subject pools that prohibit de-
ception, or seeking to publish in journals that do the same (Morton and Williams, 2010), will face
similar restrictions in choosing the optimal level of abstraction in experimental design.

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

2.1 Experimental Control

Experimenters seek to obtain “control” over the ways in which respondents construe the context-
tual features of vignettes, in order to ensure proper implementation of their experimental designs.\footnote{Similarly, experimenters seek to control the construal of treatments themselves. As Tomz and Weeks (2013) point out, if participants who are told that a country is democratic are also likely to assume that democracies are more likely to win the wars they fight, the regime type treatment becomes double-barreled, manipulating omitted details (see also Kertzer and Brutger, 2016; Dafoe, Zhang and Caughey, 2018).}

When experimental vignettes invoke different reactions amongst different types of respondents —
or perhaps invoke reactions the researcher never intended — experimenters can risk losing control
over their study, raising concerns regarding internal validity. By varying the information provided
along the three aforementioned levels of abstraction, experimenters can potentially shape the degree of control they obtain.

Yet we argue that there are less to these tradeoffs than meets the eye. First, the relationship between abstraction and control varies based upon the dimension under investigation. Increasing contextual detail is often thought to enhance experimental control, by fixing the type and degree of information that all subjects share regarding an issue area. For example, when implementing an endorsement experiment regarding a (fictional or real) immigration policy (Nicholson, 2012), researchers can provide detailed information regarding: i) who initiated the policy, ii) when it comes into effect, and iii) how it relates to previous policies. Presumably, this information can ensure an informational common denominator, and avoid a situation in which respondents with different background knowledge construe the experimental vignette in diverging ways.

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

More generally, it may be somewhat misleading to think that by turning from real to hypothetical actors, or from contextually sparse to rich vignettes, experimenters gain control over their study. Indeed, when presented with relatively pared down stimuli, participants often “fill in the blanks.” For example, scenarios in which “a country sent its military to take over a neighboring country” in which the US is considering sending troops to repel the invader (Herrmann, Tetlock and Visser, 1999; Tomz, 2007; Trager and Vavreck, 2011), may lead participants to think of the Gulf War.6

---

6Similarly, in their survey experiment on the effect of regime type on support for the use of force, Johns and Davies (2012, 1043) note that the vignette is “loosely based on recent debates about Iran but [makes] no explicit reference to that or any other real country,” acknowledging the possibility that at least politically sophisticated participants are likely thinking of Iran when they answer the questions.
over, different types of respondents may invoke different mental schema when faced with the same abstract scenario; depending on a respondent’s generation, “repel an invader” experiments may invoke Korea, Vietnam, Iraq, or none of the above. Likewise, it is possible that different respondents will exert diverging reactions to additional contextual detail, leading experimenters to lose, rather than gain control. Adopting an abstract design can thus both increase or decrease experimental control, such that the tradeoff here may not be as clean cut as experimentalists sometimes suggest.

Even if experimenters may have more leeway when choosing the appropriate level of abstraction for actor identity than is often claimed, this does not mean that all concrete actor identities are equally desirable. In particular, experimenters should attend to at least two considerations when choosing real world actors. The first is schema consistency (Hashtroudi et al., 1984): is the choice of actor reasonable given the scenario in which the actor is embedded? For example, in experimental scenarios in which a country is pursuing a nuclear weapons program (e.g., Tomz and Weeks, 2013), experimental control decreases if the experimenter chooses a country that already has nuclear weapons (e.g., Russia), or a country that respondents think is unlikely to pursue them (e.g., Canada). If a schema-inconsistent actor is chosen, the respondent is less likely to believe the scenario or accept the treatment, thus weakening the treatment effect. The second is treatment consistency: if the treatment manipulates an attribute of an actor, are all of the levels of the attribute being manipulated seen as plausible by respondents? In candidate selection experiments, for example, it would be difficult to manipulate the partisanship of politicians with well-known partisan identities, or to manipulate the policy stances of politicians on issues where they have already taken prominent positions. If respondents do not perceive the treatment as consistent with the identity of the actor, then the experimenter is likely to lose control since the respondent may not comply with the treatment, attenuating the treatment effect.

2.2 Generalizability

While experimental control is a fundamental aspect in designing vignettes, scholars may very well be concerned by other factors such as generalizability – the extent to which results from a given study speak to a broader set of real world scenarios. Like control, degrees of generalizability may be shaped by levels of abstraction in experimental design. Thus when framing an experiment as hypothetical or real, and when selecting particular actors, and levels of contextual detail, researchers may condition the degree to which their results generalize beyond a particular context.
Oftentimes, experimenters adopt unnamed actors in experimental vignettes in order to enhance generalizability. At least implicitly, the selection of an unnamed actor is motivated by the fact that a researcher’s quantity of interest is a main effect rather than a conditional effect. In other words – for example, the researcher is interested in the effect of past behavior on forming reputations for resolve in general, not the effect of past behavior on forming reputations for resolve for Iran specifically (Renshon, Dafoe and Huth, 2018).

Yet it is unclear that increased abstraction actually increases generalizability. First, when we generalize from these experiments to the problems in the real world that motivate us to conduct them in the first place, selecting unnamed actors may lead us to miss important sources of treatment heterogeneity, and may even make it harder to generalize results to any motivating real world cases. For example, because respondents are often “pre-treated” with partisan cues prior to participating in our studies (Gaines, Kuklinski and Quirk, 2007), experimenters might deliberately choose nonpartisan scenarios where these pretreatment effects are minimized, lest the effects of partisanship swamp or overwhelm the treatments of interest. Yet if many political phenomena have a partisan hue, the absence of partisan dynamics in the experiment actually makes it harder to generalize these results (McDonald, 2019).

Similarly, the degree of contextual detail provided by experimenters might shape the extent that findings from an experiment can generalize to real world scenarios. If participants in experiments only receive two pieces of information, one of which is the treatment being randomly assigned, the relative “dosage” of the treatment is likely to be unrealistically high, and may not hold in a more naturalistic setting (Barabas and Jerit, 2010). In contrast, if the treatment is presented to participants embedded in a larger amount of information (a full newspaper article, rather than just a few bullet points, for example), the treatment is likely to exert a (realistically) smaller effect. The same consideration holds with situational hypotheticality: rather than bolster generalizability, it may in fact decrease it, if when given strictly hypothetical scenarios, respondents provide strictly hypothetical answers (Converse and Presser, 1986).

In sum, then, although experimentalists frequently think about questions regarding experimental control and generalizability as two competing principles, the latter linked to abstract designs,  

---

7This is true both for contextual treatment heterogeneity in which treatment effects vary with features of the situation not being modeled in the study (e.g., do the treatment effects in a relatively abstract scenario reflect the treatment effect for an “average” country?), or in respondent-level treatment heterogeneity, in which the treatment effects in the real world vary with characteristics of respondents that would not manifest themselves in a highly stylized scenario.
and the former to concrete ones, it is not clear that the tradeoffs are actually as stark: adding contextual detail can increase control, but choosing real-world actors may lower it; respondents given pared down stimuli can still “fill in the blanks”, such that more abstract designs are not necessarily more generalizable than concrete ones. We seek to evaluate these conjectures empirically. Specifically, by experimentally manipulating the situational hypotheticality, actor identity, and contextual detail of a series of popular experiments, we aim to determine if and how different forms of abstraction shape the results of experiments. If introducing real actors or elaborate contextual detail systematically affects experimental control and generalizability, then one would expect to observe variation in outcomes across experiments varying in abstraction. If, however, the amount and type of detail across experiments only modestly shapes the results, then it would appear that the tradeoffs between these design choices are somewhat overstated.

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 survey experiments across two separate studies, each designed to address one of the dimensions of abstraction described earlier. We do so partly by building on three recent vignette-based survey experiments in political science: Nicholson’s (2012) study of elite cues in American immigration policy debates (which we refer to as the “Elite Cue” experiment), Press, Sagan and Valentino’s (2013) test of the nuclear taboo in public opinion towards the use of force (the “Nuclear Weapons” experiment), and Mutz and Kim’s (2017) exploration of the role of in-group favoritism in trade preferences (the “In-group Favoritism” experiment). The nuclear weapons and in-group favoritism experiments were fielded in a single survey, with the order of the experiments randomized, while the elite cues experiment was fielded in a separate survey.

As depicted in Table 3, our study is comprised of three levels of treatment: (1) the central treatments from the replicated studies, (2) contextual detail and actor identity treatments varying the amount of context or the names of the actors respondents are presented with, and (3) a situational hypotheticality treatment which describes experimental scenarios as either real, explicitly hypothetical, or implicitly hypothetical.8

Our selection of three different studies allows us to devote considerable attention to the ways

---

8An additional summary of the structure of our survey instrument is depicted in Appendix §1.
in which diverse types of abstraction and concreteness, relating to situational hypotheticality, actor identity, and contextual detail, affect the inferences drawn from experimental designs. The elite cues experiment replicates Nicholson’s (2012) study of elite cues about immigration reform in the United States, to explore the effects of actor identity in experimental design. The original study assigned respondents to read about an immigration proposal endorsed by one of two highly salient partisan elites or not endorsed by any actor. In our replication, we added actor identity treatments that vary whether the immigration reform endorsement is made by less salient partisan cuegivers, or by a fictional politician whose partisanship we manipulate. These treatments allow us to test how moving from more abstract (whether fictional, or low salience) actors to more concrete ones (high-salience and real) actors modifies experimental findings.

The ingroup favoritism experiment replicates portions of Mutz and Kim’s (2017) investigation of American trade preferences to study the effects of additional contextual detail. We build on their design by randomly assigning respondents to either the original short vignette, or a more elaborate vignette which provides further detail on the experimental scenario. Consistent with Bansak et al. (2020), we provide two types of additional context. The first is “filler” context, with peripheral information that increases the volume of text respondents are presented with, but is not expected to interact with the treatment. The second is “charged” context that similarly increases the length of the stimulus, but which is more relevant to the treatment. In so doing, we test how additional information that is either likely or unlikely to interact with the study’s main treatment moderates the original findings.

The nuclear weapons experiment replicates Press, Sagan and Valentino’s (2013) examination of norms against the use of nuclear weapons in public opinion, to study the effects of both actor identity and contextual detail in tandem. First, we manipulate the identity of the country in which the potential attack would take place, choosing either an unnamed country, a fictional country, or a country that is either schema-consistent with the setup of the experiment, or schema-inconsistent. 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 Press, Sagan and Valentino (2013) in its un-named country format. We then presented the study’s main

---

9While 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).
outcome questions, and asked respondents to rate the likelihood that each of eleven different countries would be the actor in each scenario. Throughout the discussion below, we refer to countries rated as the most applicable for an experimental scenario as schema-consistent, and countries rated the least applicable as schema inconsistent, thereby ensuring our selection of specific country names depicted in Table 3 is empirically grounded.

Lastly, throughout all of the studies we introduce a situational hypotheticality treatment (randomized at the subject-, not the study level) which refers to the depicted scenarios as either real, explicitly or implicitly hypothetical in order to test whether manipulating hypotheticality moderates the experimental findings. The details of the three individual studies are below, and depicted in Table 3. 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.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments from original study</td>
<td>1. No Endorsement 2. In-Party Cue 3. Out-Party Cue</td>
<td>1. US gains 1000 and other country gains 10 2. US gains 10 and other country gains 1000 3. US gains 10 and other country loses 1000</td>
</tr>
<tr>
<td>Actor identity and contextual detail treatments</td>
<td>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)</td>
<td>1. No Additional Context (original) 2. Filler Context 3. Charged Context</td>
</tr>
<tr>
<td>Situational hypotheticality treatment</td>
<td>Situation described as: 1. Implicitly hypothetical 2. Explicitly hypothetical 3. Real</td>
<td>Situation described as: 1. Implicitly hypothetical 2. Explicitly hypothetical</td>
</tr>
<tr>
<td>Factorial design (total cells in parentheses)</td>
<td>( (1 \times 3) + (2 \times 3 \times 3) = 21 )</td>
<td>( 3 \times 3 \times 2 = 18 )</td>
</tr>
</tbody>
</table>

Table 3: Summary of Treatments for 3 Studies

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

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

More details about each platform are available in Appendix §1.
3.1 Elite Cues experiment: manipulating actor identity

In our extended replication of Nicholson’s (2012) study we focus on a common dilemma for experimental political scientists: whether to name specific actors — whether politicians, countries or organizations — in a vignette or leave them unspecified, and if the actors are named, whether they should be highly recognizable, relatively low-salience, or made up from scratch. 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 replication, we updated the relevant salient cuegivers (Joe Biden or Donald Trump), while also adding additional actor identity treatments that vary whether the immigration reform endorsement is made by less salient partisan cuegivers (Senator Tom Carper of Delaware or Senator Mike Rounds of South Dakota), or by a fictional politician (Stephen Smith) whose partisanship we manipulate.13

Our procedure for the Nicholson replication is detailed in Appendix §2.1. Like Nicholson we present respondents with an experimental scenario in which information is presented about an immigration reform policy “in the news.” Depending on which treatment arm they were assigned to, respondents also read that the proposal was backed by either Donald Trump, Joe Biden, Mike Rounds, Tom Carper, Stephen Smith or did not receive any additional information about endorsements. In each condition respondents were told whether the endorser was a Republican or Democrat and for the fictional endorser — Stephen Smith — the partisan affiliation was randomized. Respondents then indicated their support for the immigration reform policy. Following the main outcome variable, respondents were asked to think about the situation again then asked to complete a thought listing exercise and a factual manipulation check (whether the policy was endorsed by a member of a particular party or not endorsed by anyone). These latter questions enable us to determine how actor identities affect respondents comprehension and recall of the general experimental scenario as well as the treatment.

3.2 Ingroup Favoritism experiment: manipulating contextual detail

Mutz and Kim (2017) use a survey experiment to examine the impact of in-group favoritism on American attitudes towards international trade, by randomizing the expected US gains and losses

---

13Additionally, we update the substantive context of the experiment to focus on protection for “Dreamers” in the U.S.
from a trade agreement, in comparison to another country. In replicating their basic framework, we
focus on a common decision experimentalist grapple with when designing instruments: how much
contextual detail should vignettes include? Should they be stripped-down, bare-bones vignettes
that highlight salient factors, or context-rich stories that in their level of detail more closely resemble
the volume of information we might encounter in our daily lives?

In particular, when implementing our study, we consider how providing respondents with in-
creased 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 treat-
ment. Apart from our contextual detail treatment, we follow a simplified version of the procedure

A detailed description of our replication protocol is provided in Appendix §2.2. In a similar
fashion to our replication of Nicholson, 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.

3.3 Nuclear Weapons experiment: manipulating both contextual detail and actor
identity

In our final study, we replicate Press, Sagan and Valentino’s (2013) experiment on “atomic aver-
sion.” They 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. They
do so by randomizing the relative probability of success for conventional attacks relative to nu-
clear attacks.\footnote{The original study includes three treatment conditions where conventional attack have a 90%, 70% or 45% success probability in relation to nuclear attacks which are held constant at 90% probability of success. The original study also randomizes whether the scenario is described as prospective or retrospective. For the sake of simplicity we only consider the 90% and 45% success probability conditions, and we described scenarios as prospective.} 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).
As detailed in Appendix §2.3, apart from our two additional treatment arms and the simplification of original treatments, we follow a similar procedure to Press, Sagan and Valentino. After the main outcome measure, we present respondents with a thought listing exercise and factual questions relating to the main treatment.

4 Results

4.1 Replication of Original Study Results

In Figure 1 we present our initial replication of the three studies under investigation. In the lower portion of the figure, we plot the main treatment on our replication of Mutz and Kim’s study of trade and in-group preferences. More specifically, the treatment in this study is exposure to a vignette describing a trade deal in which the U.S. gains 1000 jobs, in relation to a trading partner who loses 10 jobs. Respondents in the comparison conditions were either told that (i) the U.S. will gain 10 jobs, in relation to a trading partner who will gain 1000 jobs, or that (ii) the U.S. will gain 10 jobs, in relation to a trading partner who will lose 1000 jobs. Our treatment essentially exposes respondents to a trade deal that highly benefits (treatment) or doesn’t really benefit (control) the U.S., and as demonstrated in Figure 1, respondents in the treatment condition are much more likely to support trade deals when the U.S is expected to gain more.\footnote{\textsuperscript{15}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.}

In the middle region of Figure 1, we present our replication of Nicholson’s immigration-policy endorsement experiment. Specifically, we consider how exposing partisans to out-party cues supporting a given immigration reform affects support for the named reform. As evident from the study’s original data (depicted in red), as well as our replication (depicted in blue), out-party endorsement cues have a negative causal effect on support for immigration policy.\footnote{\textsuperscript{16}For ease of interpretation, in these analyses we compare respondents receiving in-party cues with respondents receiving out-party cues, omitting all respondents in the no-cue condition.} Notably effect sizes are comparable in the original study and our replication, and standard errors are smaller in the replication, likely as a result of our substantially larger sample size.

Lastly, in the upper portion, we present our replication of the Press, Sagan, and Valentino study. Doing so, we test whether exposing respondents to a vignette in which a nuclear attack is more effective than a conventional attack, alters support rates for nuclear attacks. As evident, we replicate Press, Sagan, and Valentino’s initial results, demonstrating that respondents are more likely
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 corresponding confidence interval are extracted from separate OLS models where original outcomes are predicted by treatments. When possible we benchmark our replication (Blue) to original studies (Red). In doing so we consider respondents exposed to the original format of the experiment, omitting respondents exposed to new variants of the experiment where we introduce diverging elements of abstraction ore detail. All outcomes are standardized.
to support a nuclear attack when it is described as more effective than a conventional attack. Our replication of average treatment effects (depicted in blue) is substantively similar to replications using the original data (depicted in red).

Taken together, the results in Figure 1 demonstrate our initial success in replicating our studies of interest. Now we turn to consider how our additional treatments moderate the main results depicted above. We start by considering our hypotheticality treatment. Then, we turn to consider the moderating effects of actor identity and contextual detail.

4.2 Situational Hypotheticality Effects

Does describing an experimental scenario as explicitly hypothetical, prospective, or real moderate identified effects in experimental designs? To answer this question, we administered our situational hypotheticality treatment which assigned respondents to introductions describing each experimental vignette in a particular way. Specifically, in our initial replication of Press, Sagan and Valentino and Mutz and Kim we described experimental vignettes as either explicitly hypothetical or prospective, and thus implicitly hypothetical. When considering Nicholson’s experiment in our second study, we introduced experimental vignettes as either hypothetical, real, or without addressing hypothetically at all. Throughout all our studies, subjects were randomly assigned to one hypotheticality condition at the beginning of the survey instrument. So any given subject would consistently see the same hypotheticality treatment at the start of each of our replicated experiment.

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

As evident in Figure 2, framing an experimental vignette as explicitly hypothetical does not seem to change the main findings from experimental studies. Indeed, in all models, our situational hypotheticality treatment, and its interaction with original treatments are statistically and substan-

---

17Comparisons of the explicit and implicit hypothetical conditions yield similar results.
tively insignificant. We construe 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 corresponding 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 replication of Press, Sagan and Valentino’s nuclear weapons experiment we randomized the target country as: unnamed (our baseline condition), fictional (Malaguy), real and schema inconsistent (Bolivia), or real and schema consistent (Syria). Similarly, in the Nicholson elite cues replication we randomized whether an out-party endorsement was by a made-up politician (Stephen Smith [D or R], our pooled baseline condition), a low salience politician (Senators Mike Rounds [R] or Tom Carper [D]), or a high salience politician (Donald Trump
Figure 3: Moderating effects of actor identity condition

Figure 3 shows that manipulating the country identity does not significantly moderate treatment effects in the nuclear weapons experiment, although we obtain slightly larger treatment effects in the elite cue study when we use more salient cue-givers. Point estimates and corresponding 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 condition with a fake country name condition (gray), schema inconsistent country name condition (blue), and schema consistent country name condition (red). All outcomes are standardized.

We interact this actor identity treatment with each study’s original treatment, and present results for both our elite cues and nuclear weapons replications in Figure 3 (Panel A and B respectively). In these figures, our main quantity of interest is the interaction between the original treatment and our additional actor identity treatment.

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

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

4.4 Contextual Detail Effects

Lastly, we consider the moderating effects of contextual detail in Figure 4. We administered two versions of our context treatments. In the nuclear weapons experiment, respondents were either exposed to a reduced context vignette (baseline) or an original elaborate context vignette. In the in-group favoritism experiment, respondents were either exposed to an original minimal context vignette (baseline), or an extended context vignette which included “filler” or “charged” additional context. By focusing on the interaction effect of original treatments with contextual treatments, both panels of Figure 4 lend insight into the consequences of providing additional context in experimental vignettes.

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

\[18\] In this sense, our findings offer helpful scope conditions for other experimental work arguing that using unnamed or hypothetical actors artificially inflates the size of treatment effects (McDonald, 2019): if the dependent variable involves measuring attitudes about an actor, a simple Bayesian framework would predict that the stronger the respondents’ priors (as would be the case if the actor is salient), the less they should update in response to new information about the actor. However, if the dependent variable involves measuring attitudes about a policy, that same Bayesian framework would predict that the stronger the respondents’ priors about the policy’s endorser (as would be the case if the actor is salient), the more they should update in response to information about the cuegiver.
Figure 4 shows that adding contextual detail to studies weakens the treatment effects. Point estimates and corresponding confidence intervals are extracted from three separate OLS models where original outcomes are predicted by original treatments interacted with study level context. In panel a, a baseline reduced-context condition is compared with the original elaborate-context condition used by Press, Sagan, and Valentino. In panel b, we compare a baseline reduced context vignette with elaborate context conditions which are either non-innocuous (blue) or innocuous (red). We also consider a pooled model of both types of experimental context (gray). All outcomes are standardized.
effects.

Panel (b) in Figure 4 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. Results from panel (b) 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 ($p < 0.08$). Clearly, as evident in panel (b) of Figure 4 this effect is driven by the “charged” context condition, which in and of itself has a statistically significant impact on the size (but not direction) of main treatment effects.

To better understand why adding contextual detail to experimental vignettes dampens original 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 Press, Sagan, and Valentino, and expected consequences of trade in Mutz and Kim) 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. This provides insight as to the reason why increased context may dampen main treatment effects.

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 experimental designs: 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 strong 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. Specifically, the strong
Figure 5 demonstrates how adding Contextual Detail Negatively Affects Treatment Recall. Point estimates and corresponding 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 Press, Sagan and Valentino model (left) compares recall rates of respondents assigned to a baseline reduced context conditions, with respondents assigned to extended context condition. Mutz and Kim models, compare respondents assigned to a minimal baseline condition, with respondents assigned to innocuous (blue) and non-innocuous conditions. All outcomes are standardized.
preference from those in the economics tradition is to use sparse, abstract designs, while those in the psychology tradition will prefer more rich context and detail and the use of real or made-up actor names. One thing nearly everyone agrees upon, however, is that — whatever our researcher chooses — they will face a trade-off between experimental control and generalizability. However, despite a recognition that these questions are, ultimately, subject to study and research like many other problems (e.g., Friedman, Friedman and Sunder, 1994), there is little in the way of theoretical frameworks or empirically-minded guidance for researchers who face these issues.

Our contribution here 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 outlines three dimensions of abstraction—situational hypothetically, actor identity and contextual detail—and helps us to classify prototypical experiments along these three dimensions. Most importantly, our framework and theoretical discussion of the implications of each of these three dimensions of abstraction for internal and external validity help to clarify a key point: the oft-remarked upon tradeoff between experimental control and generalizability is not nearly as stark as it is often assumed and in some cases is not a direct trade-off at all. Abstraction may in some cases enhance, rather than decrease, experimental control, which, in any case, experimentalists have less of than they realize in many cases.

Empirically, we test our theoretical framework through a replication and extension of three well-known vignette-based survey experiments in political science: Nicholson’s (2012) exploration of elite cues, Press, Sagan and Valentino’s (2013) test of the nuclear taboo, and Mutz and Kim’s (2017) study of the role of in-group favoritism in trade preferences. To each of these, we add our layers of experimental manipulations to test the implications of abstraction in experimental design. To Nicholson (2012), we manipulate the actor identity of politician presented in the vignette; to Mutz and Kim’s (2017) relatively sparse experiment we add two types of context (“filler” context and “charged” context likely to interact with the treatment) and to Press, Sagan and Valentino (2013) we add manipulations on the levels of both context and actor identity. In addition, for all three experiments, we manipulate the degree of situational hypothetically, presenting scenarios as either real, implicitly hypothetical, explicitly hypothetical, or without any mention of hypotheticality.

Our empirical results suggest reasons for optimism. Framing a study as (implicitly or explicitly) hypothetical, or real does not make any substantial difference, failing to affect any of the main
findings from the three replicated studies. This suggests that the difficult ethical decisions about whether or not to use deception 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 to Mutz and Kim (2017) and subtracting context from Press, Sagan and Valentino (2013) to render the scenario more abstract. Our results are consistent across both studies: in the vignette experiments we replicate here, we find that additional context leads to more conservative estimates of treatment effects, as long as the study has sufficient statistical power it is unlikely to affect the likelihood of finding an effect in the first place, and that context dampens treatments effects by hindering respondents’ ability to successfully recall the main treatment. Choosing the appropriate level of contextual detail in experimental work thus depends on the purpose of the study: if the purpose of a study is to demonstrate that an effect exists, a sparser experimental design better enables researchers to identify the mechanism, but if the purpose is instead to understand how important an effect might be relative to other considerations, or whether respondents in a more naturalistic setting would be likely to receive the treatment (Barabas and Jerit, 2010), a more contextually-rich design may be beneficial.

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

In line with other recent work seeking to subject widely held assumptions about experimental methods to empirical scrutiny (Mullinix et al., 2015; Coppock, 2019; Mummolo and Peterson,
we find limited empirical support to substantiate commonly held concerns regarding the consequences of design choices for the substantive interpretation of experiments in political science. Our conceptual framework clarifies where, when, and how researchers might have discretion in selecting particular levels of abstraction in their experimental stimuli. However, somewhat ironically, our evidence suggests that in cases where researchers have discretion over design choices relating to abstraction, their choices bear limited empirical consequences. Our null findings do not imply that levels of abstraction never moderate average treatment effects from experiments, but rather suggest that even if levels of abstraction moderate average treatment effects, they do so in a subtle manner which does not impact the substantive interpretation of any given experiment.
References


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.


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


Kam, Cindy D and Elizabeth J Zechmeister. 2013. “Name recognition and candidate support.”


Abstraction and Detail in Experimental Design:
Supplementary appendix

Contents

1 Survey Overview
   Figure 1.1: Overview of Study Protocol .......................... 2
   1.1 Sample information ........................................ 3

2 Study Instrumentation
   2.1 Elite Cues experiment ...................................... 5
   2.2 In-Group Favoritism experiment .............................. 6
   2.3 Nuclear Weapons experiment ................................. 9

3 Pretest Procedure .................................................. 14

4 Actor Identities and Cognitive Burden and Treatment Recall
   Figure 4.2: Actor Identity Effects on Response Times (Elite Cue Experiment) ........ 16
   Figure 4.3: Actor Identity Effects on Treatment Recall (Elite Cue Experiment) ........ 17

5 Moderating Effects of Country Name Conditional on Contextual Detail
   Figure 5.4: Country Moderating Effect by Subsample ............... 18
1 Survey Overview

The three experiments analyzed in our main text were embedded in two separate surveys, implemented in Spring 2019, and Spring 2020. Specifically, our Nuclear Weapons and In-group Favoritism experiments were fielded in Spring 2019, followed by a second study in Spring 2020 in which we fielded the Elite Cue experiment. The implementation process of these findings followed a simple and common procedure further detailed in Figure 1.1.

1. **Informed consent and screening**: Respondents are asked to consent to the study, and are screened out if they are located outside of the US or are using a mobile device to answer the survey.

2. **Assignment to situational hypotheticality treatment**: Respondents are assigned to either an explicitly or implicitly hypothetical framing in our first wave. In our second study we randomized whether scenarios were described as real, or hypothetical or whether no description of hypotheticality was mentioned in the introduction the experiment. This treatment varies across respondents but remains constant across all studies for a given respondent. To strengthen this treatment, the emphasis on hypotheticality recurs in follow up questions that mention the experimental scenario.

3. **Assignment to order of experiments**: In both studies we randomized the order of studies to avoid ordering effects.

4. **Assignment to original study-level treatments**: Respondents are randomly assigned to the original conditions of studies. Unlike the assignment of the hypotheticality treatment, this assignment is independent across all studies.

5. **Assignment to contextual detail/actor identity treatments**: Respondents are randomly assigned versions of the original studies that vary in their amount of contextual detail, and in the identities of the actors in the scenarios. Unlike the situational hypotheticality treatment, this assignment is independent across all studies.

6. **Pre-Treatment Covariate Collection**: Respondents answered a battery of pre-treatment covariates, which we will employ in future analyses.

7. **Experiment completion**: Respondents participate in experiments and respond to our main outcome measures detailed below. Outcomes include original survey items as well as additional questions which investigate respondents’ attention to the general vignette context and treatment.

8. **Additional Demographic and individual difference batteries**: Respondents respond to covariate batteries relating to: Foreign policy attitudes, cooperative internationalism, need for cognition, cognitive reflection (Thomson and Oppenheimer, 2016), political knowledge (Clifford and Jerit, 2016), and demographics.

Figure 1.1: Overview of Study Protocol
1.1 SAMPLE INFORMATION

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

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

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>4,311</td>
<td>50.807</td>
<td>17.322</td>
<td>0.000</td>
<td>99.000</td>
</tr>
<tr>
<td>Male</td>
<td>4,330</td>
<td>0.469</td>
<td>0.499</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Female</td>
<td>4,330</td>
<td>0.525</td>
<td>0.499</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Education</td>
<td>4,317</td>
<td>3.645</td>
<td>1.650</td>
<td>1.000</td>
<td>8.000</td>
</tr>
<tr>
<td>White</td>
<td>4,320</td>
<td>0.797</td>
<td>0.403</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Black</td>
<td>4,320</td>
<td>0.082</td>
<td>0.274</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Hispanic</td>
<td>4,320</td>
<td>0.043</td>
<td>0.203</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Asian</td>
<td>4,320</td>
<td>0.050</td>
<td>0.218</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Democrat</td>
<td>4,330</td>
<td>0.343</td>
<td>0.475</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Republican</td>
<td>4,330</td>
<td>0.305</td>
<td>0.461</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Independent</td>
<td>4,330</td>
<td>0.274</td>
<td>0.446</td>
<td>0.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Table 2: Descriptive Statistics - Study II (Nicholson)

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

2.1 Elite Cues experiment

To replicate the main results presented in Nicholson (2012), all subjects read the following introduction, followed by a vignette whose features randomly varied across respondents.¹

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

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

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

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

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

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

We’ve deliberately provided more boxes below than we think most people will

¹Note 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).
need, just so you have plenty of room.

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

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

possible responses include:

• Endorsed by a member of the Democratic party
• Endorsed by a member of the Republican party
• Endorsed by an independent candidate
• Not endorsed by anyone

2.2 IN-GROUP FAVORITISM EXPERIMENT

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

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

Here is the hypothetical situation / situation:

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

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

*Possible combinations are: 1,000 - 10 - gain, 10 - 1,000 - gain, 10 - 1000 - lose.

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

After collecting our main outcome variable we further ask respondents:

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

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

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

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

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

Possible responses include:

- The trading partner will benefit more than the US
- The US will benefit more than the trading partner
- Both countries will benefit equally
2.3 Nuclear Weapons experiment

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

There is much concern these days about the spread of nuclear weapons. We are going to describe a hypothetical situation the United States could face in the future. Some parts of the description may strike you as important; other parts may seem unimportant. Please read the details very carefully. After describing the situation, we will ask your opinion about a policy option.

Joint Chiefs Report Concludes Nuclear and Conventional Options for Destroying Al Qaeda Nuke Lab Equally Effective / Joint Chiefs Say U.S. Nuclear Options Offers Dramatically Increased Chances of Destroying Nuke Lab Expected Civilian Casualties, Physical Destruction Equivalent for Both Options / Chiefs Conclude Nuclear Option Has 90% Chance of Success, Conventional Only 45%

The Associated Press

A report from General Martin Dempsey, Chairman of the Joint Chiefs of Staff, 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. 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 has refused to allow international inspectors access to the facility.

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. 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 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 / Malaguayan / 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>
<thead>
<tr>
<th>Target: Al Qaeda Nuclear Weapons</th>
<th>U.S Nuclear Strike</th>
<th>U.S Conventional Strike</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability of Success</td>
<td>90%</td>
<td>90% / 45%</td>
</tr>
<tr>
<td>Estimated Syrian / Malaguayan / Foreign / Ecuadorian Civilian Deaths</td>
<td>1,000</td>
<td>1,000</td>
</tr>
</tbody>
</table>

IF U.S. STRIKE FAILS 50,000 - 70,000 US. CIVILIAN FATALITIES

Chart from Joint Chief’s report describing nuclear and conventional options for strike on Al Qaeda nuclear lab

After reading the scenario, respondents are asked:

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

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

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

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

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

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

Lastly, we implement a similar set of post-treatment questions, to determine how country names and context impact respondents ability to recall the main treatment. These question include a recall survey item and a factual question regarding the treatment which are detailed below:
When you think about the scenario / hypothetical scenario you just read, what features of the scenario / hypothetical scenario come to mind? Please list these thoughts or considerations below.

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

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

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

- Nuclear attacks will be more successful than conventional attacks
- Conventional attacks will be more successful than nuclear attacks
- Conventional and nuclear attacks have similar probabilities of success
3 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. We also considered the schema consistency of countries for another replication of Tomz and Weeks (2013) democratic peace experiment, which we discuss in other work.

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

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

- Egypt
- Iran
- Ecuador
- Bolivia
- Sudan
- Vietnam
- Turkey
- Ethiopia

2For 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).
• Kyrgyzstan
• Malaysia
• Syria

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

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

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

4 **Actor Identities and Cognitive Burden and Treatment Recall**

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

---

3The model fit of a three-factor solution is good. For Tomz and Weeks: RMSEA=0.047, TLI=0.976; for Press, Sagan and Valentino: RMSEA=0.055, TLI=0.963.
Figure 4.2 demonstrates that switching from a hypothetical actor to a low or high salience actor does not impact the cognitive burden of respondents (measured by response latency). Point estimates and corresponding confidence intervals are extracted from separate OLS models where the dependent variable—response time for the primary outcome measure—is regressed over a indicator taking the value of one for respondents assigned to high or low salience (rather than made up) condition. Sample Size for model comparing un-named and High-Salience actors is $n = 2428$. Sample Size for model comparing un-named and Low-Salience actors is $n = 2435$. 

Figure 4.2: Actor Identity Effects on Response Times (Elite Cue Experiment)
Figure 4.3 demonstrates that switching from a hypothetical actor to a low or high salience actor does not impact respondents’ ability to correctly recall treatment. Point estimates and corresponding confidence intervals are extracted from separate OLS models where the dependent variable (correctly responding to the treatment recall question), is regressed over the actor identity treatment.
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 Press Sagan and Valentino replication, in which we randomized both actor identity and contextual detail.

In figure 5.4, 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 moderating effect, which approaches conventional levels of statistical significance ($p < 0.08$) will not lead experimenters to draw substantively different inferences.

Figure 5.4 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 each panel, point estimates and corresponding confidence intervals are extracted from three separate OLS models where original outcomes are predicted by original treatments interacted with country names. In all models across both panels un-named countries are the reference category.
References


