Abstraction and Detail in Experimental Design

Ryan Brutger University of California, Berkeley
Joshua D. Kertzer Harvard University
Jonathan Renshon University of Wisconsin–Madison
Dustin Tingley Harvard University
Chagai M. Weiss University of Wisconsin–Madison

Abstract: *Political scientists designing experiments often face the question of how abstract or detailed their experimental stimuli should be. Typically, this question is framed in terms of trade-offs relating to experimental control and generalizability: the more context introduced into studies, the less control, and the more difficulty generalizing the results. Yet, we have reason to question this trade-off, and there is relatively little systematic evidence to rely on when calibrating the degree of abstraction in studies. We make two contributions. First, we provide a theoretical framework that identifies and considers the consequences of three dimensions of abstraction in experimental design: situational hypotheticality, actor identity, and contextual detail. Second, we replicate and extend three survey experiments, varying these levels of abstraction. We find no evidence that situational hypotheticality substantively changes results in any of our studies, but do find that increased contextual detail dampens treatment effects, and that the salience of actor identities moderates results in our endorsement experiment.*

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

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

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

Ryan Brutger is Assistant Professor, Travers Department of Political Science, University of California, Berkeley, 778 Social Sciences Building, Berkeley, CA 94720 (brutger@berkeley.edu). Joshua D. Kertzer is Professor, Department of Government, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138 (jkertzer@gov.harvard.edu). Jonathan Renshon is Associate Professor and Glenn B. and Cleone Orr Hawkins Chair, Department of Political Science, University of Wisconsin–Madison, 1050 Bascom Mall, Madison, WI 53706. (renshon@wisc.edu). Dustin Tingley is Professor, Department of Government, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138 (dtingley@gov.harvard.edu). Chagai M. Weiss is PhD Candidate, Department of Political Science, University of Wisconsin–Madison, 1050 Bascom Mall, Madison, WI 53706 and Predoctoral Fellow, Middle East Initiative, Harvard University, 1 Brattle Street, 02138. (cmweiss3@wisc.edu).

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

American Journal of Political Science, Vol. 00, No. 00, April 2022, Pp. 1–16

 DOI: 10.1111/ajps.12710

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 they 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), whereas others worry that detail reduces experimental control (Camerer 1997).[1] Decisions regarding abstraction and detail are particularly important for the design of survey experiments because of their emphasis on vignettes (Gaines, Kuklinski, and Quirk 2007), but they also arise in almost any experiment where researchers present respondents with information, whether in the lab (Renshon 2015) or in the field (Karpowitz, Monson, and Preece 2017).

And yet, as a discipline we know relatively little about the consequences of using abstract versus concrete experimental designs. Certainly, increasing "color in the laboratory" *may* trigger "unknown (to the experimenter) impressions and memories of past experiences over which the experimenter has no control" (Friedman, Friedman, and Sunder 1994, 53), but it is not obvious why sparse experiments would fare better in this respect. In fact, a review of the broader experimental literature suggests strong disagreement on which would be a bigger problem in terms of respondents "filling in the blanks": rich, detailed experiments (e.g., Friedman, Friedman, and Sunder 1994) or abstract, sparse studies (e.g., Alekseev, Charness, and Gneezy 2017). Although 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, 53–54), there is surprisingly little systematic work that does so, forcing experimentalists in political science to rely on hunches and intuitions rather than systematic evidence and theoretical guidance.

In this article, we seek to make both a theoretical and an empirical contribution. First, we offer an overarching conceptual framework outlining three dimensions of abstraction implicated in experimental design: *situational hypotheticality*, *actor identity*, and *contextual detail*. Our theoretical framework helps clarify when and why experimental control and generalizability may be affected by design decisions, but we also show how this debate bears on *construct validity*–the degree to which the variables in question are "measured in ways that correspond to the theoretical concepts under investigation" (McDermott 2002, 334; see also Findley, Kikuta, and Denly 2021, 368). Questions of measurement in experimental design usually center on the wording of outcome measures, but they are just as relevant to how treatments are designed and the amount of "slippage" between the concept and its operationalization in the experiment, which in turn affects "the internal validity of a study by affecting the interpretation of the results, as well as external validity" (see discussion in Hartman 2021). 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. For those latter cases where scholars do have leeway, we present a framework to guide design decisions regarding the appropriate level of abstraction and detail.

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

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

depends on how much statistical power the author expects, as well as the purpose of the study. If the purpose is to demonstrate that an effect exists, a sparser experimental design better enables researchers to identify it, but if the purpose is instead to understand how important an effect might be relative to other considerations, or whether respondents in a more naturalistic setting would be likely to receive the treatment (Barabas and Jerit 2010), a more contextually rich design may be beneficial.

Abstraction and Detail

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

Experimenters engage in abstraction when designing their stimuli, with a stimulus's level of abstraction determined by the contextual background it includes, the complexity of information it provides, and its emphasis on superordinate or subordinate elements of a given scenario. These dimensions closely relate to questions about the appropriate level of abstraction that loom large in a variety of issues in experimental design: whether experiments should be "stylized" or "contextually rich" (Dickson 2011; Kreps and Roblin 2019), use real or hypothetical actors (McDonald 2019; Nielson, Hyde, and Kelley 2019), and refer to imminent, future, or hypothetical situations. In this sense, experiments can be abstract or concrete along multiple dimensions at the same time. We suggest that abstraction in experimental design can be conceptualized along at least three dimensions: situational hypotheticality, actor identity, and contextual detail.[2] We review each dimension in detail in the discussion below.

Situational Hypotheticality

The first type of abstraction in experimental design concerns whether a scenario is described as hypothetical or not. The rationale for using hypothetical scenarios in survey experiments is simple: In experiments' 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 a control condition. Some experiments intervene by giving respondents in the treatment condition information about the world that they might not otherwise have (e.g., Butler and Nickerson 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., Arceneaux 2012; Brader, Valentino, and Suhay 2008) or find another way to justify—whether to respondents or to institutional review boards (IRBs)—why the scenario being described to respondents deviates from the one they are in.

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

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,

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

as in Boettcher and Cobb's (2006) study of how casualty frames shape support for the war in Iraq, or Evers, Fisher, and Schaaf (2019), who experimentally investigate audience costs using Donald Trump and Barack Obama. In this sense, the artificiality of the actors in an experiment is distinct from the hypotheticality of the situations in which actors are embedded since experimenters often use real-world actors in hypothetical scenarios.

Moving up the ladder of abstraction, some experimenters describe hypothetical scenarios in artificial countries. For example, Brooks and Valentino (2011) describe a conflict between "Malaguay and Westria," and Rubenzer and Redd (2010) describe a crisis in the state of "Gorendy." Taking this approach a step further, many experimentalists use unnamed countries, describing target states as "Country A" or "Country B" (Johns and Davies 2012; Yarhi-Milo, Kertzer, and Renshon 2018) or simply referring to "a country" rather than providing a label (Tomz and Weeks 2013). Other experiments focus on hypothetical political candidates. Banerjee et al. (2014), for example, describe hypothetical representatives (running for office in hypothetical districts) to study the concerns of voters in rural India. Hypothetical candidate experiments are also a long-running feature in the study of American politics (as in Colleau et al. 1990; Kam and Zechmeister 2013)—and are particularly common in conjoint experiments.

As with the case of situational hypotheticality, the logic of using unnamed or hypothetical actors stems directly from the questions being tested. Political scientists turned to experimental methods to study the effects of candidate gender (Brooks and Valentino 2011), for example, precisely because it is difficult to find two real-world candidates identical to one another on all dimensions other than their gender. The same is true in studies of race in politics (Burge, Wamble, and Cuomo 2020) or ethnicity (Dunning and Harrison 2010). In an IR context, it is hard to think of two real-world countries that are identical in all respects but one; accordingly, 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).

Contextual Detail

The third dimension of abstraction involves the amount of additional context provided in an experiment beyond the experimental treatment. Press, Sagan, and Valentino (2013) present a lengthy newspaper article that provides participants with a large amount of context, as do experiments in American politics that generate fake campaign advertisements or news clips (Brader, Valentino, and Suhay 2008). In contrast, other experiments often present relatively little information (Kanthak and Woon 2015; Tingley and Walter 2011). This decision is not limited to economics-style bargaining games: Trager and Vavreck (2011), for example, manipulate the president's strategy in a foreign policy crisis as well as information about the U.S. domestic political environment, but as with most audience cost experiments, they say relatively little about the context of the intervention itself. Similarly in comparative politics, Bassan-Nygate and Weiss (2022) randomize whether experts project that an Israeli unity government will form in the near future, but they do not include much contextual detail in their vignette.

Contextual detail is composed of at least three related dimensions. The first is simply the volume of information provided. The second concerns *how* the information is presented, and here there have been examples of any number of treatment formats in experiments, from bullet-pointed vignettes (Tomz 2007) to mock news reports (Druckman and Nelson 2003; Valentino, Neuner, and Vandenbroek 2018).[3] The third is the content of the information itself, which is orthogonal to its volume. Any bit of information may be classified as either what Bansak et al. (2021) call "filler" or its opposite, what we term "charged" content, which may interact with the treatment in some way and affect the results of a study through a mechanism other than simple respondent satisficing. If a president's "favorite highway" is filler, then Bansak et al. (2021) also show that other attributes (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 (vs. filler) content affects our interpretation of experimental results.

Control, Generalizability, and Construct Validity

Political scientists employ experiments that vary along multiple dimensions in their degree of abstraction and detail. However, there is little certainty about the consequences of this variation. To address the uncertainty

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

around the implication of design choices relating to abstraction and detail, one method has been to run both abstract and concrete versions of an experiment to test whether the results hold (e.g., Berinsky 2009; Herrmann, Tetlock, and Visser 1999; Nielson, Hyde, and Kelley 2019; Renshon, Dafoe, and Huth 2018; Rousseau and Garcia-Retamero 2007). However, this approach is less than ideal because adjusting levels of abstraction on multiple dimensions simultaneously provides limited insight regarding the specific dimension driving experimental results.

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

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

Experimental Control

Experimenters seek to obtain control over the ways in which respondents construe the contextual features of vignettes, in order to ensure proper implementation of their experimental designs. When experimental vignettes provoke different reactions among different types of respondents— perhaps reactions the researcher never intended—experimenters can risk losing control over their study, raising concerns regarding internal validity and construct validity. For example, if using particular country names as treatments triggers feelings, beliefs, or frames separate from what the experimenter was attempting to manipulate, it would introduce confounding, reducing experimental control and raising concerns about construct validity.

Yet, we argue that such concerns are likely exaggerated because they oversimplify the relationship between design choices and experimental control. First, the relationship between abstraction and control varies based upon the dimension under investigation. Increasing contextual detail is often thought to enhance experimental control by fixing the type and degree of information that all subjects share regarding an issue area.[4] In contrast, increasing detail in terms of actor identity is usually argued to reduce experimental control. In an international relations context, Herrmann, Tetlock, and Visser (1999, 556) note that "the use of real countries [adds] a degree of realism . . . but it also sacrifice[s] a degree of experimental control. Affective reactions to the various countries may differ, and [characteristics of the countries] may not be perceived uniformly by all participants."[5]

More generally, we argue that it is misleading to think that by turning from real to hypothetical actors, or from contextually sparse to rich vignettes, experimenters necessarily gain (or lose) control over their study. Indeed, when presented with relatively pared down stimuli, participants may "fill in the blanks." And when exposed to additional detail in vignettes, respondents may exert diverging reactions. Thus, the level of control or the validity of the construct measured does not necessarily increase (or decrease) with higher (or lower) levels of abstraction.

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

[4]For example, when implementing an endorsement experiment regarding a (fictional or real) immigration policy (Nicholson 2012), researchers can provide detailed information regarding who initiated the policy, when it comes into effect, and how it relates to previous policies.

[5]In American politics, Reeves and Rogowski (2018, 428) write that "the use of hypothetical candidates comes at the cost of reducing the real-world attributes of the experiment, but this cost is offset by removing respondents from their feelings about any actual politician, which could serve as confounders."

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

We argue that generalizability concerns are also exaggerated because they oversimplify the relationship between design choices and generalizability. For example, experimenters oftentimes adopt unnamed actors in experimental vignettes in order to enhance generalizability. At least implicitly, the selection of an unnamed actor is motivated by the assumption that a researcher's quantity of interest is a main, rather than a conditional, effect. For example, this is reflected when a researcher is interested in the effect of past behavior on forming reputations for resolve in general, not the effect of past behavior on forming reputations for resolve for Iran specifically (Renshon, Dafoe, and Huth 2018). For that reason, researchers may lean toward abstraction and choose an unnamed actor.

However, when considering other dimensions in experiments, abstraction may actually decrease, rather than increase, generalizability. Indeed, the degree of contextual detail provided by experimenters might shape the extent that findings from an experiment can generalize to real-world scenarios. If participants in experiments are assigned to extremely abstract vignettes where they only receive two pieces of information, one of which is the treatment being randomly assigned, the relative "dosage" of the treatment is likely to be unrealistically high and may not hold in a more naturalistic setting (Barabas and Jerit 2010). In contrast, if the treatment is presented to participants in a more detailed vignette, embedded in a larger amount of information, the treatment is likely to exert a (realistically) smaller effect. Accordingly, the expected consequences of abstraction and detail might have contradictory implications across different dimensions of our framework.

Expectations

In sum, although experimentalists frequently think about questions regarding experimental control and generalizability as two competing principles, with the latter linked to abstract designs and the former to detailed ones, it is not clear that the trade-offs are actually so stark.

For situational hypotheticality, we argue that concerns about abstraction are overblown, and so we do not expect varying situational hypotheticality to alter experimental results. Although scholars operating out of an economic tradition often express concerns that respondents will not take scenarios seriously or offer meaningful answers when told a scenario is hypothetical, there is relatively little empirical basis for these concerns. This is relevant given that there are many contexts where some form of situational hypotheticality is required (often at the demand of IRBs) to avoid the use of deception, and some contexts where the use of deception raises ethical challenges (e.g., telling respondents that a political candidate has engaged in unethical behavior; Butler and Powell 2014).

In contrast, we expect stronger moderating effects for contextual detail. We expect that increasing the amount of contextual detail in an experiment may decrease treatment dosage, and therefore reduce the magnitude of identified effects, but the effect should be larger for charged context than filler context. Consistent with Bansak et al. (2021), one can think of experimentalists as considering two types of additional context: "filler" context—peripheral information that increases the volume of text but is not expected to interact with the treatment—and "charged" context that similarly increases the length of the stimulus, but which is more likely to affect how respondents react to the treatment. Even with charged context, however, we expect it to be very unlikely that additional context would reverse the direction of treatment effects.

For actor identity, we argue that experimentalists deciding between real-world or fictional actors should keep three considerations in mind. First, experiments using real-world actors should maintain *schema consistency* (Hashtroudi et al. 1984): The choice of actor should be seen as reasonable or plausible given the scenario in which the actor is embedded. For example, in experimental scenarios in which a country is pursuing a nuclear weapons program (e.g., Tomz and Weeks 2013), the country used should "fit" with the rest of the scenario. Thus, we argue that experimental control decreases if the experiment features a country that respondents know already has nuclear weapons (e.g., Russia), or a country that respondents think is unlikely to pursue them (e.g., Canada). If a schema-inconsistent actor is chosen, the respondent is less likely to believe the scenario or accept the treatment, thus weakening the treatment effect. Second, in experiments where the treatment manipulates a feature of an actor itself, experimentalists should consider whether the real actor they use in their vignettes allows them to maintain *treatment consistency*: All levels

of an attribute of the actor being manipulated need to be perceived as equally plausible by respondents. It is for this reason that researchers are limited in their ability to select real-world actors when studying the effects of race or gender in candidate selection, or the effects of country-level characteristics on foreign policy preferences. Of course, respondents' prior knowledge varies, so some respondents may not know that an actor is schema or treatment inconsistent, whereas others may immediately recognize such incongruities in an experiment.[6] Third, because respondents are likely to have stronger attitudes about more salient real-world actors than less salient ones, any differences between real and hypothetical actors should be lower for less salient actors than more salient ones.

Research Design

To provide guidance for experimentalists on how abstract their experiment ought to be as well as how scholars should balance the potential trade-offs associated with differing levels of abstraction, we fielded a series of experiments, each designed to address one of the dimensions of abstraction described above. As the Appendix (p. 3) shows, our typology applies to any type of experiment where researchers provide information to respondents, but for purposes of tractability we focus here on survey experiments in particular. Our study selection criteria sought to replicate and extend studies that (a) focused on core theoretical debates in political science, (b) had simple designs (so that we would be sufficiently powered to detect moderation effects), (c) uncovered a large and substantively meaningful effect, and (d) were conducive to manipulating situational hypotheticality, actor identity, and contextual detail.

We focus on three experiments (depicted in Table 1), each of which features three levels of treatment: (1) the central treatments from the original studies, (2) the 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 that describes experimental scenarios as either explicitly hypothetical, implicitly hypothetical, or real. In the Appendix, we further summarize the structure of our survey experiments (pp. 1–2), along with the details of each replication and extension (pp. 3–15).

[6]We discuss and test the role of prior knowledge in the Appendix (pp. 27–31).

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

Our second study, the In-Group Favoritism experiment, extends Mutz and Kim (2017), which tests how manipulating the expected relative gains in a trade deal shapes public support. We use this study to explore the effects of additional contextual detail, randomly assigning respondents to either the original short vignette or a more elaborate vignette that provides additional detail. Those respondents assigned to additional context receive either "filler" or "charged" context to evaluate the effects of different types of contextual detail.

Our final study, the Nuclear Weapons experiment, replicates Press, Sagan, and Valentino (2013), which tests how manipulating the relative effectiveness of nuclear weapons affects public support for nuclear attacks. We use this study to explore the effects of both contextual detail and actor identity, adding two additional treatment arms. First, we manipulate the vignette's context to include either elaborate context (as in the original study) or reduced context. Second, we manipulate the identity of the actor in the dispute: (1) Syria (as in the original study), (2) an unnamed country ("a foreign country"), (3) a fictitious country name ("Malaguay"), or (4) a real and schema-inconsistent country (Bolivia).[7] Following the main outcome variable for all three experiments, respondents were asked to complete a thought-listing exercise and a factual manipulation check. These questions enable us to investigate *why* decisions about how abstract the stimuli are might moderate (or fail to moderate) treatment effects.

Throughout all of the studies, we introduce a situational hypotheticality treatment (randomized at the

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

TABLE 1 Summary of Treatments for the Three Studies

	Elite Cues	In-Group Favoritism	Nuclear Weapons
Treatments from original study	1. No endorsement 2. In-party cue 3. Out-party cue	1. U.S. gains 1,000 and other country gains 10 2. U.S. gains 10 and other country gains 1,000 3. U.S. gains 10 and other country loses 1,000	1. 45% success for conventional attack 2. 90% success for conventional attack
Actor identity and contextual detail treatments	If assigned to cue: 1. Real + High salience (Donald Trump/Joe Biden) 2. Real + Low salience (Mike Rounds/Tom Carper) 3. Fictional (Stephen Smith/Stephen Smith)	1. No additional context (original) 2. Filler context 3. Charged context	1. Extended context (original) 2. Reduced context 1. Unnamed (foreign country) 2. Made up (Malaguay) 3. Real + Schema consistent (Syria) 4. Real + Schema inconsistent (Bolivia)
Situational hypotheticality treatment	1. No mention of hypotheticality 2. Explicitly hypothetical 3. Real	1. Implicitly hypothetical 2. Explicitly hypothetical	1. Implicitly hypothetical 2. Explicitly hypothetical
Sample	Lucid	Dynata	Dynata
Sample size	4,039	4,491	4,462
Original study	Nicholson (2012)	Mutz and Kim (2017)	Press, Sagan, and Valentino (2013)

subject level, not the study level) that refers to the depicted scenarios as either explicitly hypothetical, implicitly hypothetical, or real to test whether manipulating hypotheticality moderates the experimental findings.[8] The structure of the studies is depicted in Table 1. The In-Group Favoritism and Nuclear Weapons experiments were fielded on a sample of $N = 4{,}686$ respondents through Dynata in spring 2019. The Elite Cues experiment was fielded on a sample of $N = 4{,}070$ respondents through Lucid's "Theorem" respondent pool in spring 2020.[9] In the Appendix (pp. 16–18), we report results of power simulations demonstrating that we are well powered to identify our quantities of interest.

Results

Replication of Original Study Results

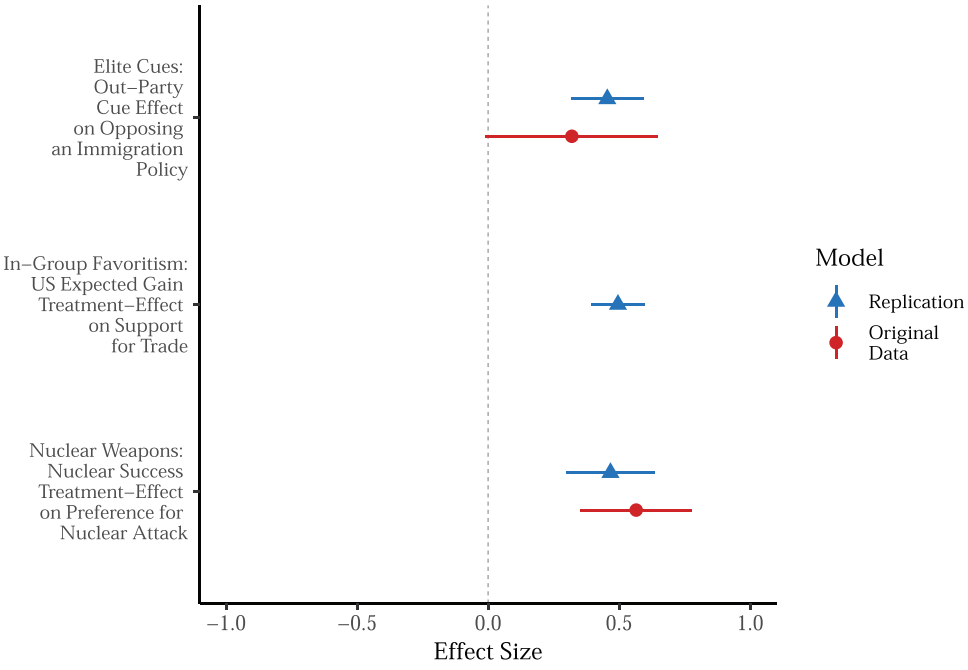
In Figure 1, we present the central treatment effects from the three studies under investigation along with comparable estimates from the original studies.[10] As expected, our Elite Cues study demonstrates that respondents are more likely to oppose an immigration policy endorsed by an out-party politician. Our In-Group Favoritism study shows that respondents are more likely to support trade deals in which the United States gains more than a rival country. Finally, our Nuclear Weapons study suggests

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

[9]More details about each platform are available in the Appendix (pp. 2–3).

[10]We do not include the original data estimate for Mutz and Kim (2017) because the original study included a more complex design with the potential for each country to gain or lose 1, 10, 100, and 1,000 jobs, in contrast to our simplified version.

FIGURE 1 Replication of Average Treatment Effects from the Three Experiments



Note: The figure shows we successfully replicate the average treatment effects from the original studies. Point estimates and confidence intervals are extracted from separate OLS models where original outcomes are predicted by treatments. All outcomes are standardized. For the Elite Cues experiment, we replicate the effect from the original study, even though our estimand slightly differs to be consistent with the analyses that follow

that respondents are more likely to support the use of nuclear weapons when they are described as more effective than conventional weapons. Taken together, the results in Figure 1 show that our extensions replicate the main results of the original studies.[11] More important is how our additional treatments moderate the main results depicted above.

Situational Hypotheticality Effects

Does describing an experimental scenario as explicitly hypothetical, implicitly hypothetical, or real affect the results obtained in experimental designs? To answer this question, we administered our situational hypotheticality treatment, which assigned respondents to introductions describing each experimental vignette as follows: In the Nuclear Weapons and In-Group Favoritism studies, we described experimental vignettes as either explicitly or implicitly hypothetical, while in our Elite Cues experiment, respondents were assigned to either an explicit hypotheticality condition, a real condition, or a pure control condition where no information about hypotheticality was provided.

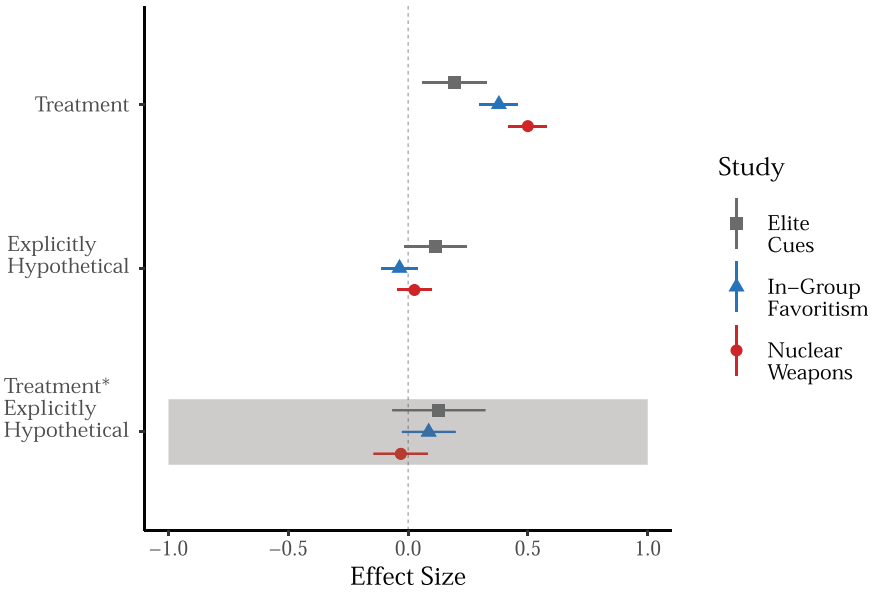
To examine the effect of this design choice, we use standard ordinary least squares (OLS) models in which we interact the original treatment from a given study (e.g., in the Elite Cues experiment, whether an out-party politician is the endorser of the immigration reform policy) with our hypotheticality treatment. Figure 2 presents results in which our main quantity of interest is the interaction effect, representing the moderating effect of our hypotheticality treatment on the original treatments.[12]

As evident in Figure 2, framing an experimental vignette as explicitly hypothetical does not change the main findings from experimental studies. In all models, our situational hypotheticality treatment and its interaction

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

[12]In our Elite Cues experiment, hypotheticality can take one of three values (i.e., explicitly hypothetical, real, or pure control). In our main analysis, we compare the explicitly hypothetical condition with the real condition, which are most distinct, but comparing the explicitly hypothetical condition with the pure control yields similar results.

FIGURE 2 No Moderating Effects of Situational Hypotheticality



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

with original treatments are statistically and substantively insignificant. We interpret these results as evidence for the limited empirical consequences of design choices relating to situational hypotheticality.

Actor Identity Effects

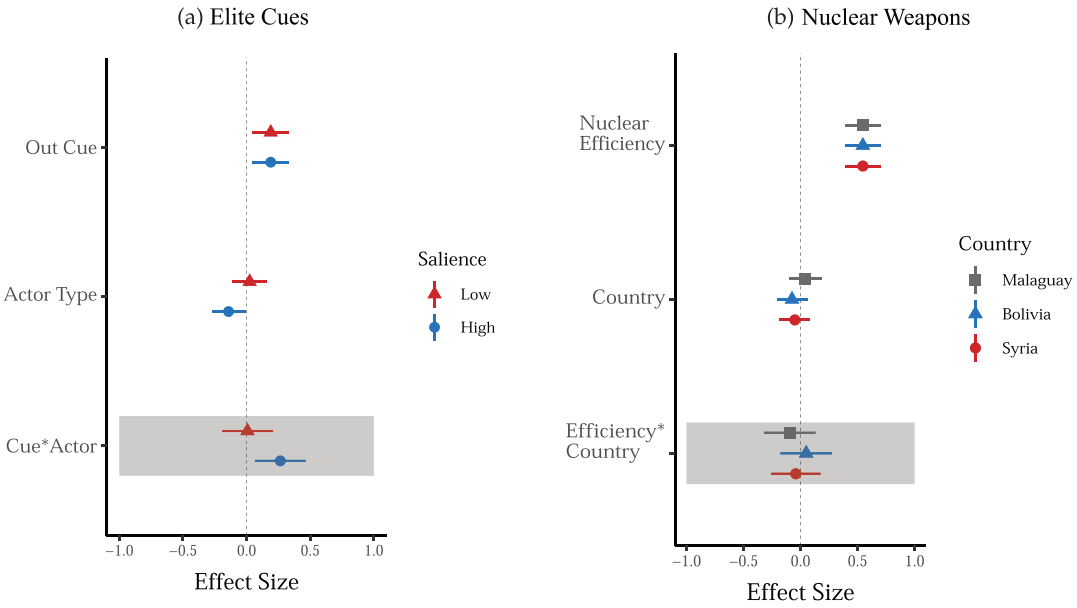
We now turn to an analysis of how actor identities of different levels of abstraction affect findings from experimental vignettes. In our Nuclear Weapons study, we randomized the target country as unnamed (our baseline condition), fictional (Malaguay), real and schema inconsistent (Bolivia), or real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

We interact this actor identity treatment with each study's original treatment and present results for both our Elite Cues and Nuclear Weapons experiments in Figure 3 (Panels a and b, respectively). In these figures, our main quantity of interest is the interaction between the original treatment and our additional actor identity treatment.

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

There are two groups of potential mechanisms to explain the actor identity results from the Elite Cues experiment. The first is a standard "online processing model" (Hill et al. 2013) in which respondents keep a running tally of evaluations that are updated when they come into contact with new information. McDonald (2019) proposes a version of this argument, contending that hypothetical actors (compared to real actors) magnify treatment effects (by decreasing the role of prior knowledge or beliefs) and increase the cognitive burden on respondents, which would show up in increased response latency and lowered treatment recall (again, compared to real politicians). Yet, as we show in the Appendix (pp. 19–20), there is no significant effect of the actor identity treatment on response latency in our study, so it does not appear that moving from a hypothetical to a low- or high-salience actor alters cognitive burden among

FIGURE 3 Moderating Effects of Actor Identity Condition



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

our respondents. A second potential mechanism that might be operative in this model is differential treatment recall: Respondents are better able to recall treatments from salient actors than nonsalient ones. Yet, as indicated in the Appendix (pp. 19–20), we find no evidence that treatment recall rates significantly vary with the actor identity treatment.

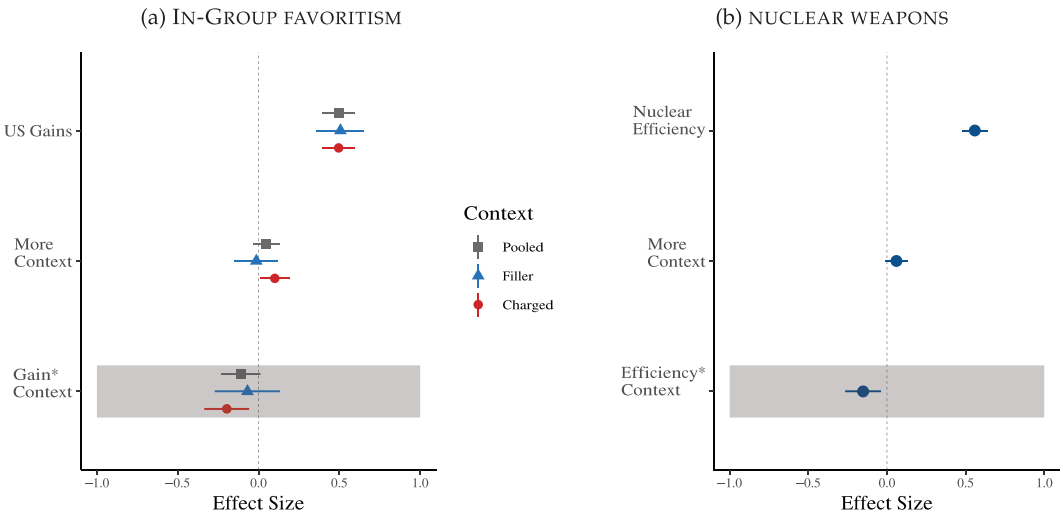
The second interpretation, which we believe is more consistent with our results, has to do with simple Bayesian models of persuasion, which focuses our attention on a different series of contrasts entirely. Bayesian models would first predict that when the dependent variable is about measuring attitudes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

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 exposed to either a reduced-context vignette (baseline) or the original elaborate context vignette. In the In-Group Favoritism experiment, respondents were exposed to either the original minimal context vignette (baseline) or an extended-context vignette that included filler or charged additional context.

[13]See the Appendix (pp. 27–35) for heterogeneous effects based on respondents' political knowledge and need for cognition.

FIGURE 4 Adding Contextual Detail Attenuates Treatment Effects



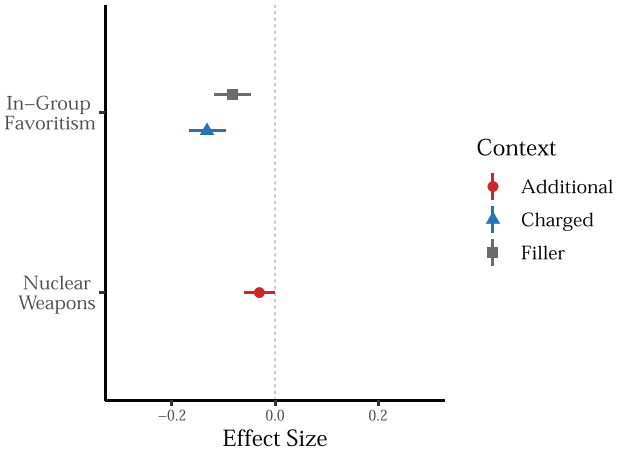
Note: The figure shows that adding contextual detail to studies weakens the treatment effects. Point estimates and confidence intervals are extracted from three separate OLS models where original outcomes are predicted by original treatments interacted with study-level context. In panel (a), we compare a baseline reduced-context vignette with elaborate context conditions that are either filler (triangles) or charged (circles). We also consider a pooled model of both types of experimental context (squares). In panel (b), a baseline reduced-context condition is compared with the original elaborate context condition used in the original Nuclear Weapons experiment. All outcomes are standardized

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

Figure 4(a) provides us with further insight into the moderating effects of contextual detail on main treatments. In this panel, we consider the general effect of adding contextual detail to experimental vignettes (gray, pooled model), as well as the particular effects of adding either filler or charged context. These results further suggest that adding contextual detail to experimental vignettes will dampen treatment effects. Indeed, the moderating effect of extended contextual detail (in relation to a baseline minimal context condition), when pooling together both filler and charged context conditions, approaches statistical significance ($p < .08$). As evident in Figure 4(a), this effect is driven by the charged context condition, which in and of itself has a statistically significant effect on the size (but not direction) of the main treatment effects. In contrast, adding filler context does not significantly affect the magnitude of the treatment effect.

To better understand why adding contextual detail to experimental vignettes dampens treatment effects, we consider the effects of our contextual detail treatment on treatment recall success. To do so, we regress respondents' recall success of the original study-level treatments (i.e., nuclear attack effectiveness in the Nuclear Weapons study and expected consequences of trade in the In-Group Favoritism study) on respondents' contextual detail condition. Figure 5 demonstrates that increased context in experimental design hinders respondents' ability to successfully recall the treatment condition to which they were assigned. In the Appendix (pp. 19–21), we further explore the positive effects of additional context on response latency, as well as the null effects of our actor identity treatment on correct recall and response latency.

FIGURE 5 Contextual Detail Effects on Treatment Recall Success



Note: The figure demonstrates how adding contextual detail negatively affects treatment recall. Point estimates and confidence intervals are extracted from three separate OLS models where a binary treatment recall success variable is predicted by the context condition to which respondents were assigned. The Nuclear Weapons model compares recall rates of respondents assigned to a baseline reduced-context condition with respondents assigned to extended-context condition. In-Group Favoritism models compare respondents assigned to a minimal baseline condition with respondents assigned to filler or charged conditions. All outcomes are standardized

Concluding Thoughts

We began this article by calling attention to a significant problem faced by political scientists who seek to test their theories using experiments: In most cases, they have a wide degree of latitude in how to design the experimental stimuli and must make choices about whether to use real actor names or make them up (or leave them unnamed), whether to add rich, contextual detail (and if so, how much, and what kind), how to present the information in the experiment (whether as explicitly hypothetical, implicitly hypothetical, or real), whether to use deception, and so on. In confronting the issues raised by these "design degrees of freedom," scholars have no shortage of folk wisdom to fall back on from their peers, mentors, and textbooks, but the conventional wisdom on which they can rely is either nonexistent or contradictory. Despite a recognition that these questions are, ultimately, subject to study and research like many other problems (e.g., Friedman, Friedman, and Sunder 1994), there is little in the way of theoretical frameworks or empirically minded guidance for researchers who face these issues. In line with other recent work, we seek to subject these folk intuitions about experimental methods to empirical scrutiny (Coppock 2019; Kertzer 2020; Mullinix et al. 2015; Mummolo and Peterson 2019).

Our contribution is twofold. First, we provided a conceptual framework that helps to make sense of the many choices that experimentalists face in terms of the degree of abstraction or concreteness of their designs. In particular, our framework draws from construal level theory to outline three dimensions of abstraction—situational hypotheticality, actor identity, and contextual detail. Most importantly, our framework and theoretical discussion of the implications of each of these three dimensions of abstraction for internal and external validity help to elucidate when there are, and are not, important trade-offs between experimental control and generalizability. Abstraction may in some cases enhance, rather than decrease, experimental control, which, in any case, experimentalists have less of than they realize. We also provide empirical leverage on the tricky question of how to appropriately operationalize the concepts we care about; empirical political scientists study "specific instances of units, treatments, observations, and settings" (Shadish, Cook, and Campbell 2002), but figuring out the implications of those specifics can now more appropriately be guided by theory and empirics in combination.

Empirically, we test our theoretical framework through a replication and extension of three well-known vignette-based survey experiments in political science (Mutz and Kim 2017; Nicholson 2012; Press, Sagan, and Valentino 2013). To each of these, we add our layers of experimental manipulations to test the implications of abstraction in experimental design. In our Elite Cues study, we manipulate the actor identity of the politician presented in the vignette; to the In-Group Favoritism study's relatively sparse design we add two types of context ("filler" and "charged") and to the Nuclear Weapons experiment we add manipulations of both context and actor identity. In addition, for all three experiments, we manipulate the degree of situational hypotheticality, presenting scenarios as either explicitly hypothetical, implicitly hypothetical, or real.

Our results suggest reasons for optimism. Situational hypotheticality does not make any substantial difference, failing to affect any of the main findings from the three studies. This suggests that the difficult ethical decisions about whether to use deception in order to increase respondent engagement may in many cases be unnecessary, adding empirical weight to an important normative debate in the field. We examined contextual detail in two ways: adding two types of context in our In-Group Favoritism study and subtracting context from our Nuclear Weapons experiment. Our results are consistent across

both studies: Additional context leads to more conservative estimates of treatment effects, dampening treatment effects by hindering respondents' ability to successfully recall the main treatment. Choosing the appropriate level of contextual detail in experimental work thus depends on how much statistical power the author expects, as well as the purpose of the study. If the purpose is to demonstrate that an effect exists, a sparser experimental design better enables researchers to identify this effect, but if the purpose is instead to understand how important an effect might be relative to other considerations, or whether respondents in a more naturalistic setting would be likely to receive the treatment (Barabas and Jerit 2010), a more contextually rich design may be beneficial.

Our results also suggest the utility of future work designed and powered to detect exactly how the potential causal mechanisms through which abstraction works, such as treatment recall or schema consistency, interact with each other.[14] We also investigated the effects of varying the level of abstraction of the actors in the experiments. We manipulated actor identity in the Nuclear Weapons experiment by exposing respondents to conditions in which the country was either unnamed, fictional, or real and either consistent with the main attributes of the scenario or not. In the Elite Cues experiment, actor identity was manipulated using made-up, low-salience, or high-salience cue-givers. Across both experiments, which considered different types of actors (i.e., countries or politicians), most actor-related design choices did not matter, in that the interaction between the actor identity treatment and the main treatment was not statistically significant. The important exception is that more salient politicians make more effective cue-givers than fictional actors do. Drawing out the implications of two potential explanations, we find suggestive evidence that this might be understandable within a Bayesian persuasion model, which would also be consistent with findings from research that manipulates the hypotheticality of the actor and measures outcomes related to attitudes about that actor (rather than the policy) (Croco, Hanmer, and McDonald 2021). Altogether, our framework and results clarify where, when, and how researchers might have discretion in selecting particular levels of abstraction in their experimental stimuli.

References

Aguinis, Herman, and Kyle J. Bradley. 2014. "Best Practice Recommendations for Designing and Implementing Experimental Vignette Methodology Studies." *Organizational Research Methods* 17(4): 351–71.

Alekseev, Aleksandr, Gary Charness, and Uri Gneezy. 2017. "Experimental Methods: When and Why Contextual Instructions Are Important." *Journal of Economic Behavior & Organization* 134: 48–59.

Alexander, Cheryl S., and Henry Jay Becker. 1978. "The Use of Vignettes in Survey Research." *Public Opinion Quarterly* 42(1): 93–104.

Arceneaux, Kevin. 2012. "Cognitive Biases and the Strength of Political Arguments." *American Journal of Political Science* 56(2): 271–85.

Banerjee, Abhijit, Donald P. Green, Jeffery McManus, and Rohini Pande. 2014. "Are Poor Voters Indifferent to Whether Elected Leaders Are Criminal or Corrupt? A Vignette Experiment in Rural India." *Political Communication* 31(3): 391–407.

Bansak, Kirk, Jens Hainmueller, Daniel J. Hopkins, and Teppei Yamamoto. 2021. "Beyond the Breaking Point? Survey Satisficing in Conjoint Experiments." *Political Science Research and Methods* 9(1): 53–71.

Barabas, Jason, and Jennifer Jerit. 2010. "Are Survey Experiments Externally Valid?" *American Political Science Review* 104(2): 226–42.

Bassan-Nygate, Lotem, and Chagai M. Weiss. 2022. "Party Competition and Cooperation Shape Affective Polarization: Evidence from Natural and Survey Experiments in Israel." *Comparative Political Studies* 55(2): 287–318.

Berinsky, Adam J. 2009. *Time of War: Understanding American Public Opinion from World War II to Iraq.* Chicago: University of Chicago Press.

Boettcher, William A., III. 2004. "The Prospects for Prospect Theory: An Empirical Evaluation of International Relations Applications of Framing and Loss Aversion." *Political Psychology* 25(3): 331–62.

Boettcher, William A., III, and Michael D. Cobb. 2006. "Echoes of Vietnam? Casualty Framing and Public Perceptions of Success and Failure in Iraq." *Journal of Conflict Resolution* 50(6): 831–54.

Bostyn, Dries H., Sybren Sevenhant, and Arne Roets. 2018. "Of Mice, Men, and Trolleys: Hypothetical Judgment versus Real-Life Behavior in Trolley-Style Moral Dilemmas." *Psychological Science* 29(7): 1084–93.

Brader, Ted, Nicholas A. Valentino, and Elizabeth Suhay. 2008. "What Triggers Public Opposition to Immigration? Anxiety, Group Cues, and Immigration." *American Journal of Political Science* 52(4): 959–78.

Brooks, Deborah Jordan, and Benjamin A. Valentino. 2011. "A War of One's Own: Understanding the Gender Gap in Support for War." *Public Opinion Quarterly* 75(2): 270–86.

Burge, Camille, Julian J. Wamble, and Rachel Cuomo. 2020. "A Certain Type of Descriptive Representative? Understanding How the Skin Tone and Gender of Candidates Influences Black Politics." *Journal of Politics* 82(4): 1596–1601.

Butler, Daniel M., and David W. Nickerson. 2011. "Can Learning Constituency Opinion Affect how Legislators Vote? Results From a Field Experiment." *Quarterly Journal of Political Science* 6(1): 55–83.

[14] In the Appendix (pp. 22–26), we find little evidence of interaction effects between different types of abstraction.

Butler, Daniel M., and Eleanor Neff Powell. 2014. "Understanding the Party Brand: Experimental Evidence on the Role of Valence." *Journal of Politics* 76(2): 492–505.

Camerer, Colin. 1997. "Rules for Experimenting in Psychology and Economics, and Why They Differ." In Wulf Albers, Werner Güth, Peter Hammerstein, Benny Moldovanu, Eric van Damme. *Understanding Strategic Interaction.* Heidelberg: Springer, 313–27.

Colburn, Timothy, and Gary Shute. 2007. "Abstraction in Computer Science." *Minds and Machines* 17(2): 169–84.

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

Collier, David, and Robert Adcock. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95(3): 529–46.

Coppock, Alexander. 2019. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods* 7(3): 613–28.

Croco, Sarah E., Michael J. Hanmer, and Jared A. McDonald. 2021. "At What Cost? Reexamining Audience Costs in Realistic Settings." *Journal of Politics* 83(1): 8–22.

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

Dawes, Christopher T., Peter John Loewen, and James H. Fowler. 2011. "Social Preferences and Political Participation." *Journal of Politics* 73(3): 845–56.

Desposato, Scott. 2015. *Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals.* New York: Routledge.

Dickson, Eric S. 2009. "Do Participants and Observers Assess Intentions Differently during Bargaining and Conflict?" *American Journal of Political Science* 53(4): 910–30.

Dickson, Eric S. 2011. "Economics vs. Psychology Experiments: Stylization, Incentives, and Deception." In *Handbook of Experimental Political Science*, ed. James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia. pp. 58–70. New York: Cambridge University Press.

Druckman, James N., and Kjersten R. Nelson. 2003. "Framing and Deliberation: How Citizens' Conversations Limit Elite Influence." *American Journal of Political Science* 47(4): 729–45.

Druckman, James N., Erik Peterson, and Rune Slothuus. 2013. "How Elite Partisan Polarization Affects Public Opinion Formation." *American Political Science Review* 107(1): 57–79.

Dunning, Thad, and Lauren Harrison. 2010. "Cross-Cutting Cleavages and Ethnic Voting: An Experimental Study of Cousinage in Mali." *American Political Science Review* 104(1): 21–39.

Evers, Miles M., Aleksandr Fisher, and Steven D. Schaaf. 2019. "Is There a Trump Effect? An Experiment on Political Polarization and Audience Costs." *Perspectives on Politics* 17(2): 433–52.

FeldmanHall, Oriel, Dean Mobbs, Davy Evans, Lucy Hiscox, Lauren Navrady, and Tim Dalgleish. 2012. "What We Say and What We Do: The Relationship between Real and Hypothetical Moral Choices." *Cognition* 123(3): 434–41.

Findley, Michael G., Kyosuke Kikuta, and Michael Denly. 2021. "External Validity." *Annual Review of Political Science* 24: 365–93.

Friedman, Sunder, Daniel Friedman, and Shyam Sunder. 1994. *Experimental Methods: A Primer for Economists.* New York: Cambridge University Press.

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

Hainmueller, Jens, and Daniel J. Hopkins. 2015. "The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes toward Immigrants." *American Journal of Political Science* 59(3): 529–48.

Hartman, Erin. 2021. "Generalizing Experimental Results." In *Advances in Experimental Political Science*, ed. James N. Druckman and Donald P. Green. Cambridge: Cambridge University Press, 385–410.

Hashtroudi, Shahin, Sharon A. Mutter, Elizabeth A. Cole, and Susan K. Green. 1984. "Schema-Consistent and Schema-Inconsistent Information: Processing Demands." *Personality and Social Psychology Bulletin* 10(2): 269–78.

Herrmann, Richard K., Philip E. Tetlock, and Penny S. Visser. 1999. "Mass Public Decisions to Go to War: A Cognitive-Interactionist Framework." *American Political Science Review* 93(3): 553–73.

Hill, Seth J., James Lo, Lynn Vavreck, and John Zaller. 2013. "How Quickly We Forget: The Duration of Persuasion Effects from Mass Communication." *Political Communication* 30(4): 521–47.

Jerit, Jennifer, Jason Barabas, and Scott Clifford. 2013. "Comparing Contemporaneous Laboratory and Field Experiments on Media Effects." *Public Opinion Quarterly* 77(1): 256–82.

Johns, Robert, and Graeme A. M. Davies. 2012. "Democratic Peace or Clash of Civilizations? Target States and Support for War in Britain and the United States." *Journal of Politics* 74(4): 1038–52.

Kam, Cindy D., and Elizabeth J. Zechmeister. 2013. "Name Recognition and Candidate Support." *American Journal of Political Science* 57(4): 971–86.

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

Karpowitz, Christopher F., J. Quin Monson, and Jessica Robinson Preece. 2017. "How to Elect More Women: Gender and Candidate Success in a Field Experiment." *American Journal of Political Science* 61(4): 927–43.

Kertzer, Joshua D. 2022. "Re-assessing Elite-Public Gaps in Political Behavior." *American Journal of Political Science.* https://doi.org/10.1111/ajps.12583

Kreps, Sarah, and Stephen Roblin. 2019. "Treatment Format and External Validity in International Relations Experiments." *International Interactions* 45(3): 576–94.

LeVeck, Brad L., and Neil Narang. 2017. "The Democratic Peace and the Wisdom of Crowds." *International Studies Quarterly* 61(4): 867–80.

Mattes, Michaela, and Jessica L. P. Weeks. 2019. "Hawks, Doves, and Peace: An Experimental Approach." *American Journal of Political Science* 63(1): 53–66.

McDermott, Rose. 2002. "Experimental Methodology in Political Science." *Political Analysis* 10(4): 325–42.

McDonald, Jared. 2019. "Avoiding the Hypothetical: Why 'Mirror Experiments' Are an Essential Part of Survey Research." *International Journal of Public Opinion Research* 32(2): 266–83.

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

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

Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman, and Jeremy Freese. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2(2): 109–38.

Mummolo, Jonathan, and Erik Peterson. 2019. "Demand Effects in Survey Experiments: An Empirical Assessment." *American Political Science Review* 113(2): 517–29.

Mutz, Diana C., and Eunji Kim. 2017. "The Impact of In-Group Favoritism on Trade Preferences." *International Organization* 71(4): 827–50.

Nicholson, Stephen P. 2012. "Polarizing Cues." *American Journal of Political Science* 56(1): 52–66.

Nielson, Daniel L., Susan D. Hyde, and Judith Kelley. 2019. "The Elusive Sources of Legitimacy Beliefs: Civil Society Views of International Election Observers." *Review of International Organizations* 14(4): 685–715.

Paivio, Allan. 1990. *Mental Representations: A Dual Coding Approach*. New York: Oxford University Press.

Press, Daryl G., Scott D. Sagan, and Benjamin A. Valentino. 2013. "Atomic Aversion: Experimental Evidence on Taboos, Traditions, and the Non-Use of Nuclear Weapons." *American Political Science Review* 107(1): 188–206.

Raffler, Pia. 2019. "Does Political Oversight of the Bureaucracy Increase Accountability? Field Experimental Evidence from an Electoral Autocracy." Working paper.

Reeves, Andrew, and Jon C. Rogowski. 2018. "The Public Cost of Unilateral Action." *American Journal of Political Science* 62(2): 424–40.

Reiley, David. 2015. "The Lab and the Field: Empirical and Experimental Economics." In *Handbook of Experimental Economic Methodology*, ed. Guillaume R. Fréchette and Andrew Schotter. New York: Oxford University Press, 410–12.

Renshon, Jonathan. 2015. "Losing Face and Sinking Costs: Experimental Evidence on the Judgment of Political and Military Leaders." *International Organization* 69(3): 659–95.

Renshon, Jonathan, Allan Dafoe, and Paul Huth. 2018. "Leader Influence and Reputation Formation in World Politics." *American Journal of Political Science* 62(2): 325–39.

Rousseau, David L., and Rocio Garcia-Retamero. 2007. "Identity, Power, and Threat Perception: A Cross-National Experimental Study." *Journal of Conflict Resolution* 51(5): 744–71.

Rubenzer, Trevor, and Steven B. Redd. 2010. "Ethnic Minority Groups and U.S. Foreign Policy: Examining Congressional Decision Making and Economic Sanctions." *International Studies Quarterly* 54(3): 755–77.

Sartori, Giovanni. 1970. "Concept Misformation in Comparative Politics." *American Political Science Review* 64(4): 1033–53.

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

Shapira, Oren, Nira Liberman, Yaacov Trope, and SoYon Rim. 2012. "Levels of Mental Construal." In *SAGE Handbook of Social Cognition*, ed. Susan T. Fiske and C. Neil Macrae. London: Sage, 229–50.

Steiner, Peter M., Christiane Atzmüller, and Dan Su. 2016. "Designing Valid and Reliable Vignette Experiments for Survey Research: A Case Study on the Fair Gender Income Gap." *Journal of Methods and Measurement in the Social Sciences* 7(2): 52–94.

Teele, Dawn Langan, Joshua Kalla, and Frances Rosenbluth. 2018. "The Ties That Double Bind: Social Roles and Women's Underrepresentation in Politics." *American Political Science Review* 112(3): 525–41.

Tingley, Dustin H., and Barbara F. Walter. 2011. "The Effect of Repeated Play on Reputation Building: An Experimental Approach." *International Organization* 65(2): 343–65.

Tomz, Michael. 2007. "Domestic Audience Costs in International Relations: An Experimental Approach." *International Organization* 61(4): 821–40.

Tomz, Michael R., and Jessica L. P. Weeks. 2013. "Public Opinion and the Democratic Peace." *American Political Science Review* 107(4): 849–65.

Trager, Robert F., and Lynn Vavreck. 2011. "The Political Costs of Crisis Bargaining: Presidential Rhetoric and the Role of Party." *American Journal of Political Science* 55(3): 526–45.

Trope, Yaacov, and Nira Liberman. 2003. "Temporal Construal." *Psychological Review* 110(3): 403–21.

Valentino, Nicholas A., Fabian G. Neuner, and L. Matthew Vandenbroek. 2018. "The Changing Norms of Racial Political Rhetoric and the End of Racial Priming." *Journal of Politics* 80(3): 757–71.

Yarhi-Milo, Keren, Joshua D. Kertzer, and Jonathan Renshon. 2018. "Tying Hands, Sinking Costs, and Leader Attributes." *Journal of Conflict Resolution* 62(10): 2150–79.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix A: Survey Overview