The Generalizability of IR Experiments Beyond the U.S.[*]

Lotem Bassan-Nygate[†] Jonathan Renshon[‡] Jessica L.P. Weeks[§] Chagai M. Weiss [¶]

May 19, 2023

Theories of international relations (IR) typically make predictions intended to hold across many countries. Nonetheless, existing experimental evidence testing the micro-foundations of IR theories relies overwhelmingly on studies fielded in the U.S. We argue that the nature of what constitutes a *theory of IR* makes it particularly important to know whether theories hold across countries. To examine the generalizability of IR experimental findings beyond the U.S., we implemented a pre-registered and harmonized multi-site replication study, fielding four prominent IR experiments in seven countries: Brazil, Germany, India, Israel, Japan, Nigeria, and the U.S. We find that all four experiments replicate in nearly all of the countries, a pattern likely due to treatment effect homogeneity. Our study not only reveals that findings from the U.S. are similar to findings from a wide range of countries, but offers important implications for the design and interpretation of future experimental research in IR.

[*] We thank Priyadarshi Amar, Ashika Amar, Lauren Barden-Hair, Saloni Boghale, Ryan Brutger, Alexander Coppock, Jamie Druckman, Kristian Vrede Skaaning Frederiksen, Erin Hartman, Josh Kertzer, Masanori Kikuchi, Oliver Lang, Ned Littlefield, Mitsuru Mukaigawara, Anton Peez-Zvetina, Victória Santos, Lauren Sukin, Dustin Tingley, Mike Tomz, Geoff Wallace and participants in the UW-Madison IR colloquium, Brown University's Watson Institute Security Seminar, the University College London Comparative Political Economy & Behavior Workshop, and UC Berkeley MIRTH for helpful comments and suggestions. Studies described were approved by the IRB at UW-Madison (#2022-0748) and pre-registered with As.predicted: https://aspredicted.org/blind.php?x=6QK_X59

[†] Department of Political Science, UW-Madison, ✉: lbassan@wisc.edu, 🌐: lotembassanygate.com

[‡] Department of Political Science, UW-Madison, ✉: renshon@wisc.edu, 🌐: jonathanrenshon.com

[§] Department of Political Science, UW-Madison, ✉: jweeks@wisc.edu, 🌐: jessicalpweeks.com

[¶] Conflict and Polarization Lab, Stanford University ✉: cmweiss@stanford.edu, 🌐: chagaiweiss.com

In recent years, the field of international relations (IR) has often turned to experiments to test the individual-level "micro-foundations" of important IR theories (Hyde, 2015; Kertzer, 2017). Given the advantages of experiments in terms of causal identification (McDermott, 2011*b*), this approach has provided valuable evidence about theories of international conflict (Tomz and Weeks, 2013), trade (Mutz and Kim, 2017), nationalism (Powers, 2022), and immigration (Hainmueller and Hiscox, 2010), among others. Over time, a cottage industry has sprung up to further improve the internal validity of experimental research, shoring up what is already one of the method's key strengths.[1]

A new wave of research in political science has focused on issues of external validity and generalizability, questioning whether and how scholars can extrapolate from a single study to different contexts, populations, and measurement strategies (Egami and Hartman, 2022). Recent work has sought to provide theoretical foundations for these concepts (Humphreys and Scacco, 2020; Egami and Hartman, 2022; Findley, Kikuta and Denly, 2021; Slough and Tyson, 2021) and engaged in empirical exercises designed to probe questions such as whether experimental findings hold across various country contexts (Dunning, Grossman, Humphreys, Hyde, McIntosh and Nellis, 2019; Coppock and Green, 2015). For example, scholars of Comparative Politics have engaged in multi-site replications (or "meta-keta" studies) (Dunning, Grossman, Humphreys, Hyde, McIntosh and Nellis, 2019; Dunning, Grossman, Humphreys, Hyde, McIntosh, Nellis, Adida, Arias, Bicalho, Boas et al., 2019), and recent research in American Politics has combined large-scale replication projects with meta-analyses (Coppock, Hill and Vavreck, 2020; Blair, Coppock and Moor, 2020; Schwarz and Coppock, 2022).

IR, however, has lagged behind these important endeavors. To the extent that scholars have examined the "generalizability" of existing IR experiments, they have tended to focus on extending findings from a single study to one or several additional contexts (Tomz and Weeks, 2013; Renshon, Yarhi-Milo and Kertzer, 2023; Suong, Desposato and Gartzke, 2020), oftentimes introducing design changes across countries and providing limited motivation for case selection. In that sense, even large multi-site experiments in IR are often designed or analyzed such that they cannot provide direct evidence on the extent to which existing findings generalize to other countries.

[1] See, for example, Keele, McConnaughy and White (2012); Clifford and Jerit (2015); Dafoe, Zhang and Caughey (2018); Offer-Westort, Coppock and Green (2021); Blair, Coppock and Moor (2020); Clifford, Sheagley and Piston (2021); Mutz (2021); Brutger, Kertzer, Renshon, Tingley and Weiss (2022); Brutger, Kertzer, Renshon and Weiss (2022).

Assessing the generalizability of IR findings is crucial for remedying the mismatch between the scope of IR theory and the breadth of its underlying evidence. The nature of what constitutes a "theory of international relations" makes it particularly important to evaluate support for IR theories across a range of country contexts, but the vast majority of microfoundational experimental evidence has come from the United States, a country that is unusually powerful, conflict-prone, and wealthy, and whose citizens are particularly "WEIRD" (Western, educated, industrialized, rich and democratic; Henrich, Heine and Norenzayan, 2010*b*). Thus we do not currently know whether IR theories are truly *international*, rather than merely theories equipped to explain the foreign policy preferences of Americans. Moreover, assessing experimental results in other countries allows us to examine whether findings from *non*-U.S. contexts yield generalizable results, potentially reducing barriers to entry for scholars based outside the U.S.

To explore these issues, we implemented a pre-registered and harmonized multi-site replication study, with both outcomes and treatments congruent across all replication sites (Slough and Tyson, 2021). We fielded four prominent IR experiments—about audience costs (Kertzer and Brutger, 2016; Tomz, 2007), democratic peace (Tomz and Weeks, 2013), international law (Wallace, 2013), and reciprocity in international trade (Chilton, Milner and Tingley, 2020)—in a set of seven carefully-selected countries: the U.S., India, Germany, Japan, Brazil, Nigeria, and Israel. We employed a "purposive variation" site selection strategy (Egami and Hartman, 2022), choosing a broad range of countries within the original theories' scope conditions. Committing to a harmonized pre-registered design helps us sidestep two central challenges of cumulative learning: publication bias (i.e., selective reporting of positive results) and study comparability (a prerequisite for generating externally valid knowledge, Slough and Tyson, 2021).

Our study makes three central contributions. First, we provide rigorous evidence bolstering the generalizability of prominent experiments testing the micro-foundations of important IR theories. Our meta-analysis yields striking levels of cross-country similarity in average treatment effects, suggesting that the results from our four experiments in seven countries were substantively similar to results from the original U.S.-based experiments. Moreover, sign generalization tests (Egami and Hartman, 2022) suggest that an overwhelming majority of country-experiment combinations (93%) yield results in the theoretically expected direction. Though we cannot say whether we would find such consistent results across the universe of IR theories or countries of interest, our theoretical focus on prominent experiments from different substantive domains and our empirical focus on a set of purposively varied countries suggests that a broad range of IR experiments generalize to a diverse array of countries.

Second, our findings emphasize that the U.S. is not an outlier in terms of experimental evidence on the micro-foundations of general IR theories – and nor are any of the countries we studied. American citizens may very well be "WEIRD", and different from other populations in terms of foreign policy dispositions (as our data demonstrate). However, our main results suggest that these dispositional differences may not dramatically shape experimental findings. Only a single country-experiment pair yielded a decisive null result. Moreover, the only situation in which a study yielded effects in the *opposite* direction from the original U.S.-based study was a case in which the original paper's theory anticipated heterogeneous treatment effects (as a function of individual and potentially other contextual differences). Thus, while our results in this case were different from the original study, they still conform to the study's theoretical expectations.

Third, our study provides insights for scholars seeking to probe the generalizability of future experimental research. We provide suggestive evidence that our consistent pattern of results is due to low treatment effect heterogeneity (Coppock, 2019) in the studies we replicated (and with the moderators suggested by those theories). In line with recent studies in American politics (Coppock, 2022*a*), we found that respondents with different covariate profiles reacted to our experimental stimuli in very similar ways. To the extent that we found heterogeneity, it tended to shape the magnitude rather than the direction of effects. This exercise, we argue, has important implications for future experimental research in both international relations and other subfields. It suggests that researchers should theoretically and empirically interrogate the extent to which their effects are heterogeneous, theorizing ex-ante about individual-, situational- and country-level variables that may moderate treatment effects, and incorporating measures of these moderators at the design stage. Doing so allows scholars to investigate whether treatment effects are heterogeneous within their sample, in which case they should consider how samples in other contexts might differ along relevant covariates and be cautious about making strong claims about generalizability.

Together, these results suggest a somewhat surprising conclusion that despite the U.S.-centric base of experimental IR research, the field does not appear to be in an evidentiary crisis. Given evidence of low treatment effect heterogeneity for the four "general" IR theories we studied, the fact that past research has focused disproportionately on U.S.-based subjects does not appear to have produced dramatically skewed findings. Just as importantly, our findings suggest that scholars testing such theories can learn much by fielding single-country studies in non-U.S. contexts, particularly if they measure potential moderators and demonstrate evidence in support of low treatment effect heterogeneity. This conclusion has important practical and normative implications for future scholarship by reducing "barriers to entry" for non-U.S. based scholars conducting experimental research, and for correcting the impression that the U.S. should be the "normal" site for experimental research.

At the same time, our study demonstrates the value of harmonized multi-site replication studies in the context of experimental research testing the micro-foundations of IR theories. Our empirical exercise provides reassuring insights regarding the breadth of generalizability in IR experimental research. However, our approach also allows us to identify an important context in which one of our original experiments does not replicate, as well as instances where researchers should be more cautious with regard to generalizability (i.e., theories that predict heterogeneous responses to treatment). In that sense, we view pre-registered harmonized multi-site replication studies as an important component in the IR research cycle in which researchers establish the generalizability and scope of single-country findings.

1 Defining External Validity and Generalizability

Scholars often refer to a dichotomy between *internal* and *external* validity. Internal validity refers to confidence that a given finding results from a particular experimental manipulation (McDermott, 2011*a*, 28). Specific findings–in our case, results from empirical tests of IR theories—may or may not be internally valid, and that quality is specific to a particular study (McDermott, 2011*a*, 28; see also Shadish, Cook and Campbell, 2002). In contrast, external validity—"the extent to which a given result is generalizable to alternative contexts, populations, and measurement strategies—is not a property of individual experiments" (Renshon, 2015, 667). Rather, we learn about external validity "as replications across time and populations seek to delineate the extent to which…conclusions can generalize" (McDermott, 2011*a*, 28).

Critics have lamented that political science has "fallen down an internal validity rabbit hole" (Findley, Kikuta and Denly, 2021, 366) to the detriment of other goals.[2] More recently, scholars have attempted

[2] Though, it is worth noting that which type of validity has been over-emphasized is a matter of perspective (and contention). Writing in 2011, McDermott (2011*a*, 27) argued that in political science, concerns with external validity "often border on the mono-maniacal."

to develop the concept of external validity theoretically and to generate methods for probing the concept empirically, examining issues including the design of experiments, nature of the sample, and other factors (Hainmueller, Hall and Snyder Jr, 2015; Bisbee and Larson, 2017; Kertzer, 2022).

In conceptualizing generalizability, we build on Egami and Hartman (2022), who decompose external validity into four components, $X-$, $T-$, $Y-$, and $C-$ validity, referring respectively to populations, treatments, outcomes, and contexts/settings. Our interest is in assessing $C-$ validity: "Do experimental results generalize from one context to another context?" (Egami and Hartman, 2022, 5). We focus, in particular, on cross-country variation, a notable instance of $C-$ validity that is challenging because geographic variation typically involves "a question about covariates that have no variation within an experiment." Put differently: "$C-$validity is the main concern when we ask whether the experimental result is generalizable to a context where no experimental data exist" (Egami and Hartman, 2022, 7).

Broadly, our investigation of generalizability focuses on external validity across countries. We consider a particular finding to be *generalizable* if support for it—in the form of precisely estimated ATEs in the theoretically expected direction—can be found across a variety of contexts that are within the bounds of a theory's scope conditions. If, for example, a theory makes predictions about dynamics within democracies but not within nondemocracies, the scope of that theory might be all democratic countries. Thus we would consider a theory generalizable if we found consistent experimental support for it across an array of democratic countries.

2 Generalizability in IR

Having defined generalizability, we now assess the explanatory goals of typical IR theories. Since the early days of the field, IR theories have been concerned with "analysis of the contemporary multi-state system and. . . the conduct of its *components*" (emphasis added; Wolfers, 1947, 26). A more recent treatment emphasizes that IR theories are intended to explain international politics in "*general* causal terms" (emphasis added; Walt, 2005, 26). IR scholars typically portray their theories as providing general explanations of inter-state relations rather than insights into one specific country or region. This is true whether we focus on "grand" frameworks such as realism or "middle-range" theories explaining the democratic peace, and whether the actors under discussion are states, leaders, voters, or IOs. For example, theories of reputation

(Downs and Jones, 2002; Wolford, 2007) and resolve (Kertzer, 2016) aim to generate predictions about states, leaders, and perceptions on a general level, not restricted to any one particular state, leader, or specific empirical context. If a given theory applies in only one country, it would be considered a theory of that country's foreign policy rather than a "theory of international relations." In sum, IR theories are usually intended to provide broad insights about interstate relations that apply across a wide range of countries.

Given these goals, it is important to ask whether, for any given theory, a sufficient base of evidence from multiple contexts exists to support the theory—ideally, an accumulation of empirical tests from a broad range of countries. To document the extent to which past experimental studies in IR have relied on evidence from samples in particular countries, we conducted a quantitative literature review identifying all IR articles containing experimental studies ($N = 369$) published in the top Political Science journals (APSR, AJPS, JOP) and IR sub-field journals (IO, ISQ, JCR) over the past two decades (2000-2021). Figure 1 plots these studies by country "site" (location). Strikingly, sixty percent of all experiments in our sample focused on U.S. subjects. Moreover, U.S.-based experiments were eight times more common than experiments in Israel, the next most popular site.[3] Evidently, experimental research on the micro-foundations of prominent IR theories relies predominantly on studies of U.S. foreign policy attitudes, behaviors, and perceptions.

As previously demonstrated, concerns about generalizability are relevant to cross-national observational studies (Aronow and Samii, 2016), and more generally, U.S.-centrism is not unique to experimental research (Colgan, 2019*a*). Rather, it seems to be part of a long-running trend in which IR research is dominated by an American approach to studying politics (Hoffmann, 1977; Colgan, 2019*a*; Levin and Trager, 2019; Kristensen, 2015), whether in terms of scholars' countries of residence (Wæver, 1998), how PhD programs train graduate students (Kang and Lin, 2019), patterns of publishing and citation (Kristensen, 2012), or even the content of prominent cross-national datasets (Colgan, 2019*a*). Per Hendrix and Vreede (2019, 311), the U.S. "is not the eight-hundred-pound gorilla in the literature, but the three-hundred-thousand-pound blue whale."

Regardless, one could reasonably worry that relying predominantly on U.S.-based samples to carry out micro-level empirical tests of IR theories does not provide much insight into the generalizability of a theo-

[3] This dovetails with Hendrix and Vreede (2019, 311), who points out that Israel and the U.S. broadly receive attention in the scholarly literature far out of proportion to their population, GDP, etc.

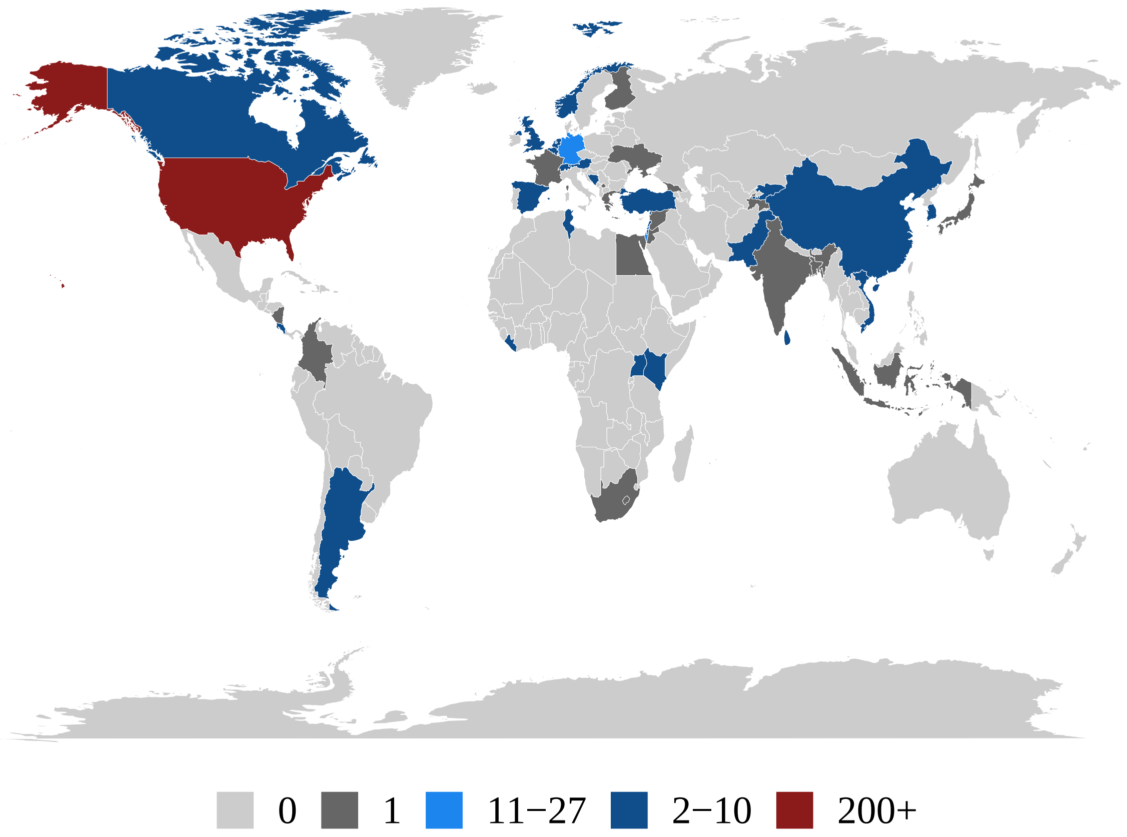

Figure 1: Geographic distribution of IR experiments published in six top journals between the years 2000-2021.

rized empirical relationship. The U.S. is wealthier, has enjoyed longer-standing democratic institutions,[4] is more geographically-protected, more conflict-prone, and more powerful and authoritative than most other countries. To the extent that country-level factors like these could impact the ideologies, perceptions, or judgments of respondents, experimental findings from U.S.-based subjects might shed little light on the preferences of populations in other places. U.S.-based subjects might also differ from residents of other countries at an individual level. For example, Americans tend to be less knowledgeable than their peers in other locations (Levin and Trager, 2019, 352; Dimock and Popkin, 1997), and the U.S. is different demographically even from other large, powerful countries (Brooks et al., 2018). In fact, psychologists have

[4] It is worth noting, though, that the U.S. democracy scores may have been (and may still be) inflated due to American-centric bias in democracy codings (Colgan, 2019*b*, 301; Levitsky and Ziblatt, 2019)

developed a term—Western, educated, industrialized, rich and democratic (WEIRD)—to denote societies like the U.S., arguing that people from such places ”are some of the most psychologically unusual people on Earth” (Henrich, Heine and Norenzayan, 2010*b*, 29), with the U.S. standing out even from its WEIRD peers as “an outlier in an outlier population” (Jones, 2010, 1627; see also Henrich, Heine and Norenzayan, 2010*a*).

On the other hand, concerns about the risk posed by focusing on U.S. samples may be overblown. Coppock, Leeper and Mullinix (2018), for example, replicate 27 (largely non-IR) experiments that had originally been carried out on nationally-representative samples and find strong correspondence between the original results and replications performed using online convenience samples. They use these results to contend that many social science experiments exhibit low “treatment effect heterogeneity”: i.e., for many studies, subgroup treatment effects do not differ that much, which might lead one to expect that effects will not differ much across national contexts, either. In an IR context, there is also suggestive evidence that at least some results are relatively robust to different contexts and samples. For example, Renshon, Yarhi-Milo and Kertzer (2023) find similar effects of “democratic reputations” across six national samples, and Suong, Desposato and Gartzke (2020) find that evidence on the micro-foundations of the democratic peace theory from the U.S. and the U.K. generalize to Brazil. However, without a more systematic harmonized research program assessing the generalizability of prominent IR theories, it is impossible to say whether the U.S. focus of existing IR experiments represents an acceptable base for broader knowledge or an evidentiary crisis.

3 Research Design

3.1 Overview

We build on our conception of generalizability to inform the design of our harmonized multi-site replications. Four key features of our design are worth noting. First, and perhaps most importantly, our study is designed to answer a different set of questions than typical IR replication studies. Previous works have tended to probe a single study’s “external validity” by fielding the same instrument—at times with design variations—at one or more alternative sites to explore whether an effect identified in an initial context replicates in a different population (examples in IR include Tomz and Weeks, 2013; Lupu and Wallace, 2019).

In contrast, we focus on the following two questions: (1) what is the *underlying level* of support for a given IR theory across a broad range of countries? (2) In *how many* of the countries do we find treatment effects in the theoretically expected direction?

The second important feature of our study is our use of "purposive variation" for selecting sites (i.e., countries). This selection strategy is designed to yield variation across sites along theoretically-important moderators (Egami and Hartman, 2022). This approach has the advantage of being both principled and empirically verifiable, while lending itself directly to analytical methods (sign generalization tests and meta-analysis) that enable us to answer the two key questions outlined above.

Third, our design is harmonized, reducing the possibility that idiosyncrasies in timing, logistics, or design variations could render studies incomparable (Slough and Tyson, 2021). We harmonized our study in terms of treatment (identical across countries), outcomes (identical across countries), timing (all experiments implemented at the same time to hold constant external information environment), and samples (single survey aggregator to increase comparability across countries). Fourth and finally, we pre-registered our study to reduce the risk of selective reporting, a general problem that is particularly salient when the goal is to evaluate generalizability.

With our main goal and these design features in mind, we selected studies that test the micro-foundations of general IR theories that should apply beyond the U.S., employ relatively simple designs, were found to produce robust effects in the U.S., and cross substantive boundaries within the field of IR. This approach led us to include experiments on the democratic peace (Tomz and Weeks, 2013), audience costs (Tomz, 2007; Kertzer and Brutger, 2016), international law (Wallace, 2013), and reciprocity in foreign direct investment (Chilton, Milner and Tingley, 2020). More information on the four studies is provided in Appendix B, and details of treatments and outcomes are depicted in Table 1. Below, we describe our method of site selection in more detail and then summarize our analytical strategy and outputs.

3.2 Choosing Country Contexts Based on Purposive Variation

Case selection is rarely discussed explicitly or critically interrogated in experimental research. However, when the goal is to learn about generalizability, the question of site selection takes on added importance (see also Allcott 2015). Below, we detail a purposive country selection process (Egami and Hartman, 2022), which we use to select seven country sites.

Approaches to case selection can generally be characterized as either random or non-random. Random approaches have obvious benefits but provide little leverage here, given that an N of 7 countries does not permit inferences about a broader population of interest (i.e., all countries within the scope of the theory). On the other hand, popular non-random approaches have their own problems. For example, convenience sampling—selecting sites based on ease of access—perpetuates the disadvantages of relying on U.S. samples in the first place: sites that are easiest for scholars to access may resemble the U.S. and differ in systematic ways from less convenient sites. Alternatively, experimentalists often invoke the concept of "least-likely" (or "hard") cases from the the qualitative methods literature. However, the "least-likely" approach is actually meant to shed light on causal effects in the presence of confounding, which is not relevant in randomized experiments.[5]

We opt for a different non-random approach, based on "purposive variation," in which one selects sites in order to ensure variation along theoretically-important moderators. This approach addresses two key issues. First, it provides a framework for investigating and understanding variation in treatment effects across countries that results from moderators that contribute to treatment effect heterogeneity. Moreover, purposive variation also addresses a second issue, namely how to generalize existing evidence to *unobserved* contexts. Even when a study is conducted in a variety of countries, its findings are inherently "local" and require additional assumptions to generalize to unobserved contexts (Egami and Hartman, 2022, 11-12). Using theoretically informed purposive variation allows researchers to more credibly make the "range assumption," which states that the true causal effect lies within the range of purposively varied countries under investigation. Under the range assumption, researchers can use analytical strategies such as sign generalization tests (described below) to extrapolate from the "local" findings to state more general conclusions.

In light of the two reasons above, it's useful useful to choose cases, at least in part, to reflect a diversity of theoretically-relevant moderators. We therefore specified four key theoretical components of our four

[5] In the qualitative methods literature, "least-likely" cases provide a "hard test" for a theory, where finding support for a hypothesis provides particularly strong evidence in favor of the theory. Although experimentalists often result to using the "least likely" language when describing case selection, this approach is unsuitable for selecting sites for experiments as it addresses an issue, confounding, that is already solved through experimental designs. Consider a theory involving an independent variable (Z), a dependent variable (Y), and potential confounding variables (X) related to both Z and Y. In the qualitative literature, a case is "least likely" if the observed level of Z predicts a particular value of Y, but an alternative explanation (X) predicts a different value of Y. If Y takes on the value predicted by Z even though other background variables predict a different outcome, the theory passes a "hard test," increasing confidence in its predictive power. Thus, least-likely designs are meant to shed light on causal effects in the presence of potential confounding. Obviously, this problem is already solved in experiments by randomizing the treatment within each country.

studies: i) Treatment, ii) Mechanism, iii) Outcome, and iv) Moderators (Findley, Kikuta and Denly, 2021).[6] For three out of four studies, our interpretation of existing research revealed theoretically-relevant moderators – strength of democratic norms in the democratic peace experiment; hawkishness in the audience costs experiment; and international legal obligation in the international law experiment. Table 1 summarizes the theoretical components of all four studies and specifies the expected direction of the moderating effect.

	Treatment	**Mechanisms**	**Outcome**	**Moderator**
Democratic peace	Adversary regime type	Conflict perceived as immoral/costly	Support for military attack	Democratic Norms (+)
Audience costs	Leader backs down after initiating a threat	Leader perceived as inconsistent and/or belligerent	Approval of leader	Hawkishness (-)
International law	Information that torture violates international law	Perceived legitimacy of law or expected cost of violation	Support for the use of torture	International legal obligation (+)
FDI Reciprocity	A foreign country's FDI policy	Concern for fairness	Support for FDI policy	NA

Table 1: Theoretical components of our studies. Sign in parentheses indicates the direction of the moderating effect.

After parsing the theories, our country selection process proceeded systematically through five steps, depicted in Figure 2 (detail in Appendix C). First, we *determined the scope conditions* of each theory and excluded countries that fell outside of these conditions. Since two of our selected studies—audience costs and democratic peace—-make predictions unique to voters in democracies, and given that public opinion likely plays a larger role in democracies, we focus on countries above a minimum threshold of democracy (Polity ≥ 6). Second, we sorted all countries that met this scope condition based on *policy importance*, allowing us to focus on powerful countries that are more consequential in world politics. This entailed sorting democracies based on their GDP and prioritizing more powerful countries over less powerful ones, all else equal, but *without sacrificing the key variation along moderators described below*.

Third, we aimed to maximize variation along *unmeasured* moderators like religious and cultural fac-

[6] All four original studies specified clear treatments and outcomes, but the mechanisms and moderators were sometimes not discussed in detail. In such cases, we built on the authors' theoretical framework to lay out the causal mechanisms and moderators. We also contacted all authors in order to verify that our reading of their theory's implications was reasonable. We note that although moderators vary at the national level, they can also be measured at the individual level.

tors by selecting countries from each major region of the world.[7] As a fourth step we verified variation along our *predefined moderators*: military expenditures, years since becoming a democracy and territorial integrity rights.[8] As demonstrated in the bottom-right panel of Figure 2, our selected countries yielded significant variation: there are at least two countries above and two countries below the cross-national mean of each moderating variable. As a final step, we considered *practical constraints* by verifying that our survey provider – Lucid/Cint – operates in the selected countries and is able to match country samples on key demographics of the general population of interest (i.e., gender and age). This step did not constrain our case selection procedure, and Lucid/Cint was able to offer samples from all selected countries – Brazil, Germany, India, Israel, Japan, Nigeria, and the U.S.

3.3 Expectations and Analytical Strategies

Above, we identified two key questions we seek to answer: (1) what is the *underlying level* of support for a given IR theory across a broad range of countries? (2) In *how many* of the countries do we find treatment effects in the theoretically expected direction? To answer these questions, we focus on two key estimations: a meta-analysis and a sign generalization test.

Meta-Analyses

First, to identify the underlying support for a given theory across all countries in our sample, we identify a cross-country meta-analytic effect, representing the average of effects across all countries under investigation (Borenstein et al., 2021). This involves two steps. First, we estimate simple country-specific OLS regressions to identify country-average treatment effects (and their corresponding standard errors) for each experiment. We then aggregate these ATEs using a meta-analytic random-effects model, which essentially provides a weighted average of effects from all countries (Borenstein et al., 2021). Weights in the random effects model are determined by the inverse of the variance of each study's average treatment effect (representing sampling variability), as well as by the variance of effects across studies (representing the

[7] We rely on seven regions utilized by the World Bank – Latin America, North America, South Asia, East Asia, Europe, Sub-Saharan Africa, and the Middle East.

[8] These three moderators correspond—roughly—to the theoretically relevant moderators in Table 1. We use data from the Stockholm International Peace Research Institute (SIPRI) on military expenditure as a proportion of government spending as a proxy for hawkishness. We use the number of years a country has been a democracy and the Physical Integrity Rights Index to indicate the strength of democratic norms. Finally, we use the number of international treaties a country has ratified to proxy for the level of international legal obligation.

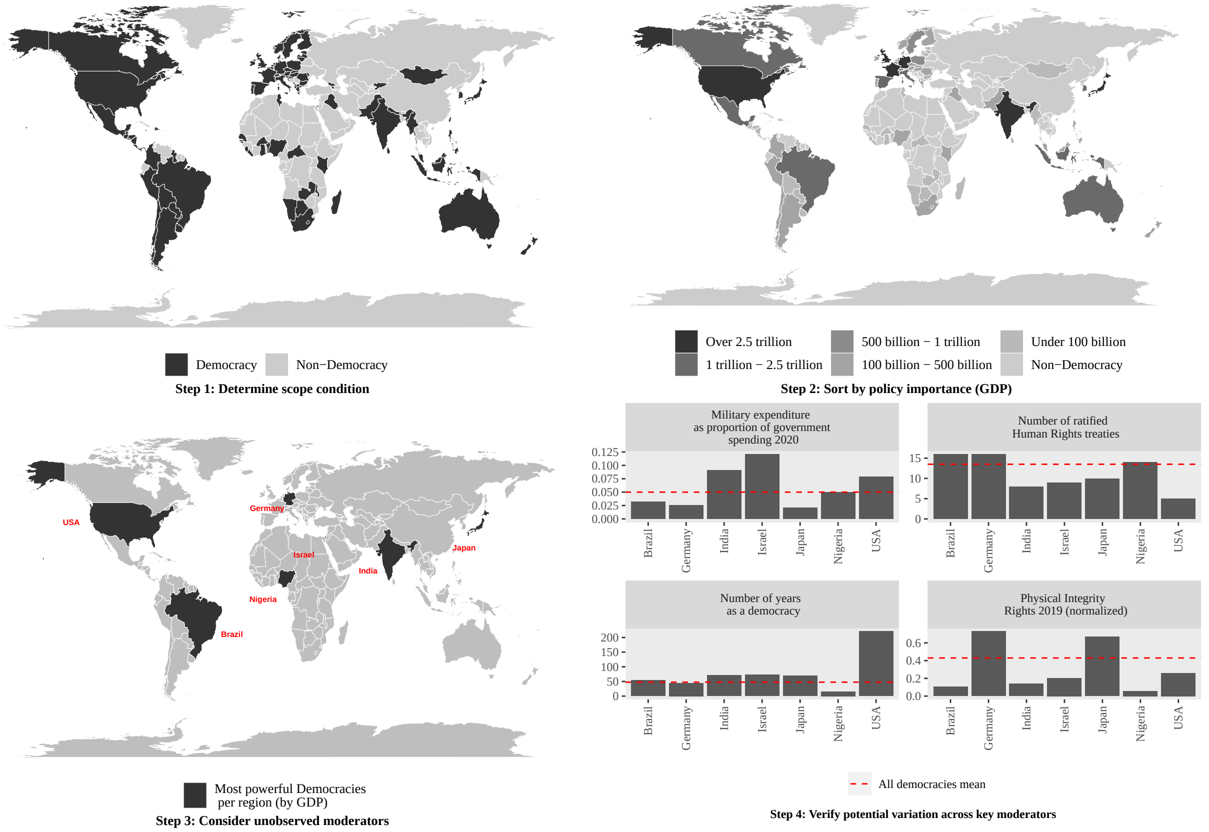

Figure 2: Steps in selecting countries for replication. GDP data is from the World Bank. Military expenditure data is from Stockholm International Peace Research Institute (SIPRI). Physical integrity rights data is from Fariss (2019).

heterogeneity of the true effect across countries).[9]

[9] By employing the meta-analytic random-effects model we assume that for our population of interest—-countries that fall within our theoretical scope conditions—there exists a distribution of effect sizes for a given treatment. Under the assumption that the country-specific ATEs we identify represent a random draw from the broader distribution of ATEs, our random effects model provides the mean and variance of the overall distribution of ATEs (Borenstein et al., 2021). While this is probably an overly strong assumption given the number of countries in our sample—even a random draw of 7 countries out of the overall population would probably not suffice—our approach nevertheless allows us to learn about a general and substantively important quantity of interest—the average of ATEs across countries within our scope conditions, and the variance of this ATE. Alternative approaches, namely fixed effects meta-analyses, assume that there exists one true value of the ATE across all countries (rather than a distribution of ATEs), and that any observed variance in ATEs across countries is due *entirely* to sampling variability. In contrast, our models make the more plausible assumption that variance across country ATEs is due to a combination of sampling variability and true cross-site variation in ATEs (Borenstein et al., 2021). Since we believe that ATEs might subtly vary across countries and are skeptical that there exists a single value for ATEs across all countries, we employ a random effects model.

Sign-Generalization Test

Second, we seek to directly test the extent to which experiments generalize across specific contexts. To do so, we use a sign generalization procedure proposed by Egami and Hartman (2022), which leverages design-based purposive variation (in our case, country of implementation) and employs a partial conjunction test to estimate the share of experiments that yield a precisely estimated effect in the theoretically expected direction.

Using a sign generalization test is subtly different from simply looking at the country ATEs by study and tallying up the number of precisely estimated results. The latter approach would be on shaky ground because each country ATE represents the p-value of a particular test in a particular country (after correcting for the six other tests), and as such should be interpreted on its own. In contrast, the sign-generalization test is more appropriate as it considers all implemented studies and directly tests the *share of studies* for which we can reject a null in favor of a result in the theoretically expected direction. Using this test, we consider a particular experimental finding to be more generalizable the larger the share of countries in which we can reject a null hypothesis when implementing the sign generalization test.

Interpreting individual null results

Of course, any given study-country combination may produce a null or even opposite result for any number of reasons, including random chance. One or two unexpected findings per experiment would not necessarily undermine our conclusion that an experimental finding "generalizes," in the sense of a statistically significant meta-analytic ATE or a sign generalization test showing generalizability in most contexts. However, null results would provide evidence that a finding does not hold in a *particular* context.

The interesting question then becomes, *why* would a study replicate in some country contexts but not others? Within the confines of space and resource constraints, we designed our studies to allow us to probe such a result. In advance of fielding the study, we pre-registered secondary analyses related to questions of attentiveness, respondents having a particular country in mind, the plausibility of the scenarios, and effect heterogeneity along theoretically-relevant moderators.

4 Results

We fielded our harmonized study in all seven countries (Brazil, Germany, India, Israel, Japan, Nigeria, and the U.S.) in late January and early February 2023 using Cint.[10] For each country, we collected data from around 3,000 attentive respondents.[11] We programmed country-specific surveys using Qualtrics and allowed respondents to choose between English and the dominant national language.[12] Each survey started with a consent form, followed by attention checks embedded in a battery of pre-treatment survey questions measuring social and political dispositions. Respondents who correctly answered the attention checks then proceeded to our four experimental studies. To avoid order effects, we randomized the sequencing of experiments. Appendix D describes our survey instruments in detail. In Appendix E, we report descriptive statistics of our cross-national sample. Like a majority of survey experiments in political science, our results are based on convenience samples (Mullinix et al., 2015; Coppock, Leeper and Mullinix, 2018) designed to mirror the local population in terms of gender and age distribution.[13]

4.1 Strong Underlying Support for Generalizability of IR Theories

In Figure 3, we report our main meta-analyses for all four experiments, assessing the underlying support for each theory across the whole sample of countries. Figure 3 is comprised of three panels: the bottom panel shows the point estimate and 95% confidence interval reported in the originally published papers (all implemented amongst samples of U.S. survey respondents). The middle panel of Figure 3 reports our country-specific average treatment effects with 95% confidence intervals and Benjamini-Hochberg adjusted p-values to account for false discovery rates (Benjamini and Hochberg, 1995).[14] Finally, the top panel of

[10] Cint acquired Lucid in 2021, and a range of recent studies reporting cross-national experiments use Lucid see: Frederiksen (2022); Arechar et al. (2022); Bor and Petersen (2022). Our research procedures were reviewed by the relevant Institutional Review Boards (IRBs) and determined to be "exempt." We followed all principles and guidance for human subjects research published by the American Political Science Association. Cint/Lucid compensated respondents for their participation in the study using their usual procedures. Further details about human subjects considerations are available in the appendix. All replication data will be made available on Dataverse upon publication.

[11] We determined our sample size by power analyses ensuring that we are well-powered (>80%) to identify original point estimates within each country ($\alpha = 0.05$).

[12] We made the survey instrument available in the one or two dominant languages in all countries, requiring translation for all countries aside from the United States and Nigeria. A list of minor wording changes we introduce to address cross-site comparability can be found in Appendix D.2.

[13] In Appendix H we follow Devaux and Egami (2022) and examine the sensitivity of our main results to external validity bias, and consider the extent to which reweighing our sample using different covariate profiles would explain away identified treatment effects. Our results from this exercise suggest that in most country-experiment pairs, employing samples with varying covariate profiles would not explain away identified effects.

[14] Benjamini-Hochberg corrections are applied at the experiment level, accounting for seven tests of the same hypothesis.

Figure 3 reports the meta-analytic average treatment effect for each experiment based on our country-specific point estimates and standard errors from the middle panel.

To help calibrate our replication studies, one can compare the ATEs from the original studies (fielded in the United States) with the ATE from our U.S. sample (bottom row of the middle panel). Here, one can see that the ATEs from our U.S. sample converge with the original study ATEs in both statistical significance and direction. This increases confidence in our larger set of studies, suggesting both that our studies (as fielded) were appropriately comparable to the original studies and ruling out any temporal changes that might have affected support for the theory in the U.S. in the interim between the original studies and our replications.

The general pattern of results in Figure 3 is both striking and reassuring: all four point estimates from our meta-analyses are precisely estimated in the same direction as the original point estimates from published experiments implemented in the U.S. There is some variation in the magnitude, as might be expected. For the audience costs and democratic peace studies, the original ATEs were larger than our meta-analytic ATE, while in the international law study the ATEs were similar and in the reciprocity study, our meta-analytic ATE is larger than the originally-estimated ATE.[15]

We interpret the overall pattern of results reported in Figure 3 as suggesting that average treatment effects drawn from experiments implemented in the U.S.—whether as part of our replications or in the original studies—are representative of the underlying level of support for a given theory in our cross-national sample. Indeed, in terms of the direction of effects, the substantive conclusions that one would draw from experimental studies in the U.S. are identical to the conclusions one would draw from multi-site experiments implemented in a diverse set of countries with varying institutional, cultural, and economic characteristics. Notably, the directional congruence between original point estimates and point estimates from our meta-analyses is not an artifact of a small number of countries generating large effects and compensating for null or negative findings in most countries. Indeed, across our twenty-eight country-experiment dyads, there is no instance in which there is support for any effects in the opposite direction, and only two instances where point estimates are not statistically significant. However, in order to get a more systematic view of the

[15] In the audience costs experiment, the magnitude of the original average treatment effect ($\beta = -0.347$) is approximately double the size of the point estimate from our meta-analysis ($\beta = -0.646$). Similarly, the original average treatment effect from the democratic peace experiment ($\beta = -0.282$) is more than double the size of the point estimate in our meta-analysis ($\beta = -0.122$). Turning to the international law experiment, we find that the original point estimate ($\beta = -0.106$) is rather similar to the point estimate from our meta-analysis ($\beta = -0.130$). Finally, in the reciprocity FDI experiment, we find that the original point estimate ($\beta = 0.120$) is almost a third of the magnitude of the point estimate from our meta-analysis ($\beta = 0.332$).

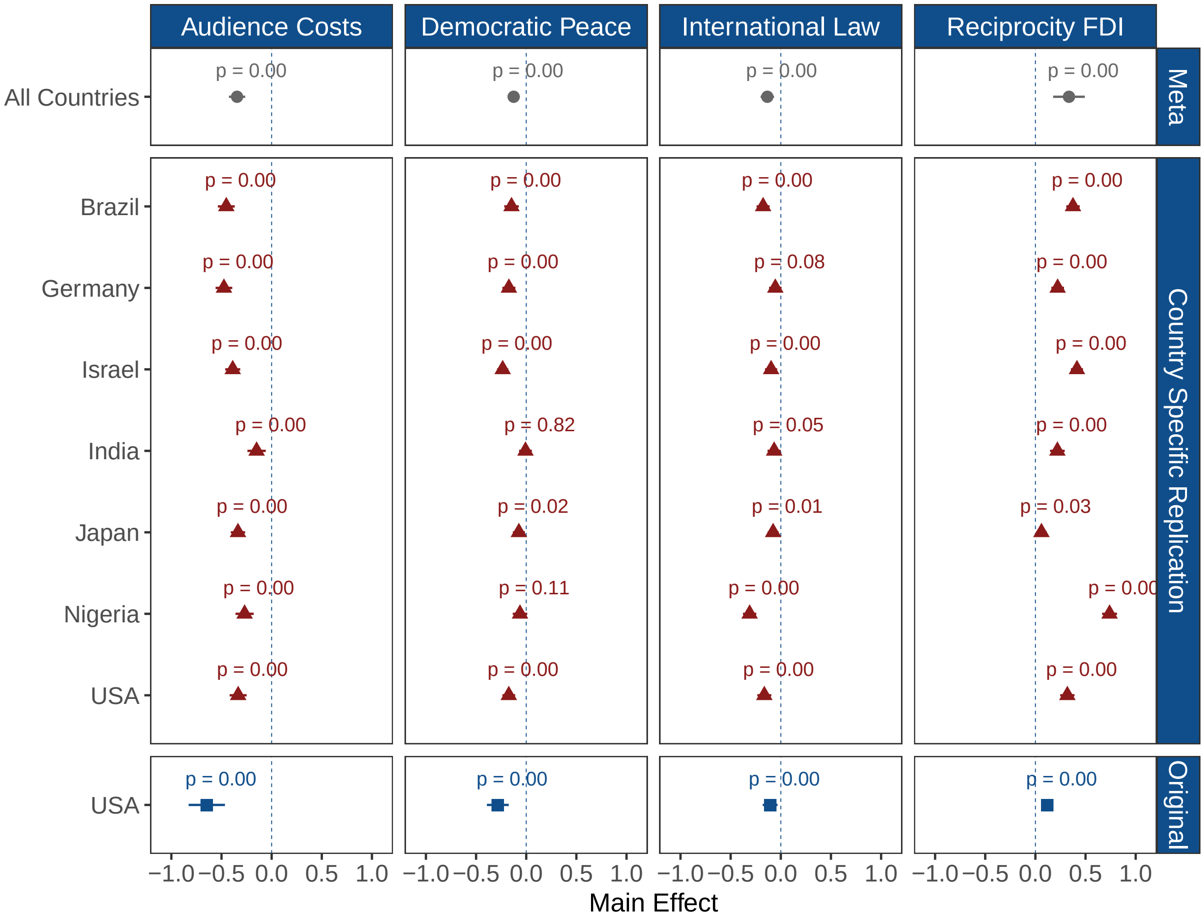

Figure 3: **Meta-Analysis.** For each experiment, we report original point estimates and p-values from published studies, alongside country-specific ATEs (and Benjamini-Hochberg adjusted p-values) from our replications and a meta-analytic average treatment effect based on our harmonized studies. To increase comparability across experiments, all outcomes are standardized ($\mu = 0$, $\sigma^2 = 1$).

generalizability of each experiment across countries, we now turn to report sign generalization tests (Egami and Hartman, 2022).

4.2 Strong Support for Sign-Generalization Among IR Experiments

In Figure 4, we report results from sign generalization tests for our four experiments. Recall that the sign generalization test leverages purposive variation in experiments (in our case, variation in the country of implementation) to systematically quantify the external validity of experiments across different contexts (Egami and Hartman, 2022). We estimate partial conjunction p-values for each experiment and assess the generalizability of a given experiment based on the number of countries in which we obtain a positive result with a corresponding partial conjunction p-value < 0.05.

In the top-left panel of Figure 4, we report the sign generalizability test for our audience costs replication. As indicated by the flags and associated p-values in this panel, the audience costs experiment yields a high level of sign generalizability: we obtain p-values < 0.05 across all seven of our countries. We obtain similar findings for our reciprocity FDI experiment (bottom right panel), in which p-values for all seven countries are again estimated to be < 0.05. The bottom-left panel of Figure 4 presents results for our International Law experiment, where our sign-generalization test yields five p-values < 0.05 suggesting sign generalizability of over 71%. Notably, however, the two other p-values in this test that are > 0.05 are very small ($p = 0.05$). We thus construe the overall pattern of results for the international law experiment to imply relatively high levels of sign generalization across countries.

Turning to the democratic peace experiment in the upper-right panel of Figure 4, we find general support for sign generalization. Indeed, we obtain partial conjunction p-values < 0.05 for five out of seven countries. The countries in which we obtain partial conjunction p-values > 0.05 are Nigeria ($p = 0.09$) and India ($p = 0.41$). We construe the relatively small p-value in Nigeria ($p < 0.1$) as providing suggestive evidence for sign generalization in that context as well. However, our data suggest that findings on the micro-foundations of the democratic peace theory do not generalize to our India sample, a finding which we further interrogate in Section 4.3.

Taken together, the results in Figures 3-4 provide optimistic insights into the generalizability of IR experiments outside the U.S. Specifically, our meta-analytic average treatment effects from four separate experiments across seven diverse geographical contexts yield similar substantive results when compared

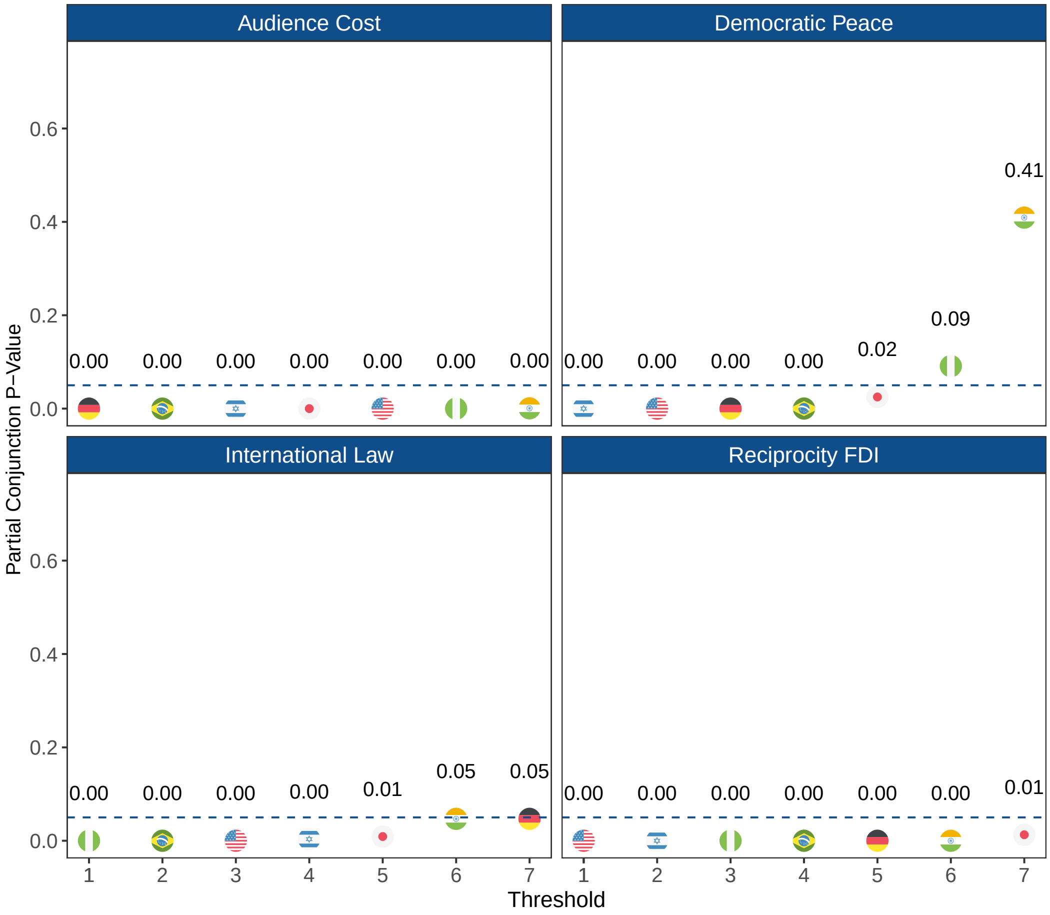

Figure 4: **Sign Generalization Test.** For each experiment, we report the proportion of country replications that generalize in the theoretically expected direction. Country replications are denoted by flags and are generalizable in cases where the partial conjunction p-value is smaller than 0.05. Partial conjunction p-values are denoted above each flag.

to original average treatment effects from experiments implemented in the U.S. Moreover, our sign generalization tests suggest that for all four experiments, original findings from U.S. experiments generalize to a majority of other geographical contexts. Under the range assumption detailed by Egami and Hartman (2022), our sign generalization tests suggest that the experimental findings we replicate have a high degree of sign generalizability. That said, we do find one instance where a study does not replicate in a particular context: the democratic peace experiment implemented in India. We now report a number of pre-registered analyses through which we interrogate this intriguing result.

4.3 Probing the Null: Explaining Democratic Peace in India

In our meta-analyses, we identify wide-ranging micro-level support for the four IR theories we tested. However, there is one exception to this pattern of generalizable results: the democratic peace study in India, where the effect of our democracy treatment on supporting an attack yielded a precise null ($\beta = -0.00$, $p = 0.82$, $CI = [-0.07, 0.05]$). In designing our study, we anticipated potential nulls and pre-registered a series of analyses to rule out some potential explanations—and explore others—for what appears to be a deviation from the overall pattern of findings.

In Appendix K, we provide a detailed overview of our additional pre-registered analyses. These analyses provide strong evidence against scenario implausibility, low attentiveness, ceiling or floor effects, and priming of specific countries (Dafoe, Zhang and Caughey, 2018) as potential explanations for the null effect in the democratic peace experiment in India. We find some evidence that our theoretical moderator (democratic norms) significantly moderates the democracy treatment effect and that India is a context with relatively low support for democratic norms. This pattern suggests that our null result in India is partially a function of low democratic norms dampening average treatment effects. However, as we explain in Appendix K, given the limited magnitude of treatment effect heterogeneity, such moderation is unlikely to fully account for the null effect.

We conjecture that the micro-level mechanisms tested in the original democratic peace experiment (Tomz and Weeks, 2013) do not apply in India because of historical dynamics involving ongoing conflict with neighboring Pakistan. Given that Pakistan is often considered to be a democracy, individuals in India may have concluded that democracies do not adhere to norms of peaceful conflict resolution and pose significant threats, undermining the mechanisms underlying the democratic peace. We see this single result as a call for empirical studies that probe scope conditions—both empirical and theoretical—and a reminder of the importance of empirical research for theory-building. Ultimately, however, it is important to contextualize this null result within the broader pattern of findings, which reveals a high degree of generalizability for our experiments across seven countries.

4.4 Explaining Generalizability: Treatment Effect Homogeneity and IR Experiments

Overall, our experiments provide strong evidence in support of generalizability. In the previous section, we interrogated one exception to these findings—India in the democratic peace experiment—but it is equally important to examine the overall pattern of results and ask what explains the strong degree of correspondence between estimates from the U.S. and other countries. We now evaluate two explanations for these patterns relating to *sample characteristics* and *treatment effect homogeneity*. We conclude that the most plausible explanation for the general pattern of results relates to (limited) treatment effect heterogeneity within our studies.

One explanation for our strong pattern of generalizable results could involve the characteristics of the samples we collected. Since we relied on online convenience samples, one might worry that we selected particularly "WEIRD" subjects who resemble U.S. citizens along key covariates. This could bias our results in favor of identifying effects similar to original studies implemented in the U.S. We conclude, however, that this explanation is unlikely. Our Appendix reports distributions of key covariates and demonstrates a meaningful degree of cross-country variation along hawkishness, international legal obligation, and support for democratic norms (Figure A8).[16] Moreover, in Table A8, we formally test differences between country samples by regressing the moderators as well as a host of theoretically important demographic variables (education, ideology, and age) over country indicators. Since a majority of estimates are significantly different from the U.S., we conclude that our country samples vary along demographic and theoretically relevant covariates, and thus cross-country similarity in samples is unlikely to explain our main pattern of results.

A second explanation for our consistently significant results relates to treatment effect homogeneity. According to this explanation, the experiments we selected produce relatively constant effects across subjects and are thus likely to generalize across different samples from various countries (Coppock, 2019). Indeed,

[16] Variation along some of these moderators strikingly matches our country-level proxies – see, for example, the country distributions of our Hawkishness measure compared to the military expenditure proxy in Figure 2. Other proxies proved less precise, but we nonetheless observe variation across countries, as seen in Table A8.

in line with the substantive interests of IR scholars, we intentionally chose studies testing the observable implications of *general IR theories* that are theorized to hold — i.e., to produce treatment effects in the same direction — across different contexts.[17]

However, some IR theories explicitly hypothesize about *conditional* effects, where the existence and even direction of an effect is expected to depend on particular individual-level, country-level, or situational moderators. While "general" theories produce predictions that are largely expected to hold across different contexts; theories that are more "conditional" predict that average treatment effects differ across groups and contexts, sometimes even differing in sign. One should *expect* experiments that test such theories to exhibit treatment effect heterogeneity, sometimes to the point of producing effects in different directions among different subsamples and across different contexts.

Our full replication of the Audience Cost study by Kertzer and Brutger (2016) allows us to consider an experiment designed to test a conditional theoretical expectation and compare its results to our main pre-registered experiments. The results we presented in our main analyses compared only the "back down" to "stay out" conditions, but as we explain in Appendix J, in our replication, respondents were assigned to three experimental conditions, allowing us to decompose the general audience cost into belligerence and inconsistency costs, the costs that leaders pay for engaging in bellicose behavior and the costs the leaders pay for not following through on their statements, respectively. Kertzer and Brutger (2016) theorize and provide evidence that there is respondent-level variation in who punishes versus rewards belligerent leaders, and country-level and situational factors should also influence reactions to belligerence. Given that these individual-, country-, and situation-level factors vary across our samples, we should find effects in different directions across subgroups and contexts.

In our secondary analyses reported in Appendix J, we replicate the full audience costs design to identify belligerence costs and their sensitivity to respondents' predispositions. To compare our general results with the belligerence cost extension, we implement a treatment effect homogeneity test proposed by Ding, Feller and Miratrix (2016). This test allows us to see how individual-level moderators influence treatment effects within countries, one important form of heterogeneity. Specifically, this empirical exercise leverages a Fisher Randomization Test procedure to test a null hypothesis of homogeneity in average treatment effects.

[17] To the extent that average treatment effects in these experiments vary as a function of moderators (some of which we specify in Table 1), one would expect such moderation to attenuate (or amplify) effects rather than lead to opposite results.

In Table 2, we report results from this homogeneity test. When examining our four main experiments implemented across seven countries, we can reject the null hypothesis of homogeneity in about 28% of country-experiment pairs. In other words, treatment effects appear to be mostly homogeneous in our four main experiments. Moreover, as we show in Appendix G, to the extent that heterogeneity along theoretical moderators does exist, it attenuates effects but does not yield directionally opposite results across subgroups.

Study	**N Comparisons**	**N Significant**	**N Significant (BH Adjustment)**
Audience Costs	7	3	2
Democratic Peace	7	1	0
International Law	7	3	2
Reciprocity FDI	7	4	4
Belligerence Cost (AC Extension)	7	7	7

Table 2: This table reports results for treatment effect heterogeneity tests developed by Ding, Feller and Miratrix (2016).

In contrast, when examining treatment effect homogeneity in the belligerence cost extension, we find strong evidence against homogeneity. Indeed, in the bottom row of Table 2, we reject the null hypothesis of homogeneity across all countries in our sample. Moreover, as we show in Appendix J, the consequences of treatment effect heterogeneity, specifically as it relates to respondents' hawkishness, go beyond effect magnitude and bear consequences for the direction of effects. For example, the belligerence treatment yields positive effects for hawkish respondents and negative effects for dovish respondents, in line with Kertzer and Brutger's expectation that hawks will reward belligerence while doves punish it.

Given the theoretical expectation of conditional effects, the evidence that individual-level hawkishness shapes reactions to belligerence, and the fact that the countries in our sample vary in terms of average hawkishness, one would expect reactions to the belligerence cost treatment to vary across countries. We provide empirical support for this expectation in Appendix J by showing that the belligerence treatment (the effect of engaging versus staying out) yields null effects in two countries, negative effects in two countries, positive effects in three countries, and an overall null meta-analytic average treatment effect. Comparing our general pattern of results discussed in Section 4.1, where treatment effects are largely homogeneous, with results from the belligerence costs extension, where treatment effects are heterogeneous, further suggests that the generalizability of our main findings is driven by treatment effect homogeneity. When treatment effects are heterogeneous, they do not generalize as well across a set of purposively varied countries.

5 Conclusion

This paper was motivated by the concern that the breadth of experimental evidence in international relations does not match the scope of its underlying theories. Although most theories of international relations make predictions that should apply to a wide array of countries, past experimental studies on the micro-foundations of IR theories have overwhelmingly relied on U.S.-based samples. To examine the extent to which prominent experimental findings generalize to a diverse set of countries, we fielded a pre-registered and harmonized multi-site replication study across a set of seven purposively varied countries.

In our main analyses, we found that all four of the experiments we replicated produced consistent results across a wide array of countries. We purposively selected countries to ensure variation in key variables that could moderate the treatment effects we set out to test. Nonetheless, in 26 out of the 28 (93%) country-experiment pairs we examined, we found significant treatment effects in the anticipated direction. Moreover, we found no cases of contradicting results in which treatments had positive effects in some contexts and negative effects in others. In only one situation—the democratic peace experiment in India—did treatments yield a precisely estimated null effect, deviating from the overall pattern of results.

We draw three key conclusions from these findings. First, our findings provide encouraging insights about the generalizability of IR-related experiments. We cannot know without additional replications whether a different set of experiments would have yielded equally consistent results across countries, and indeed, secondary analyses indicate that tests of theories with more "conditional" predictions may not replicate as widely. However, the fact that four experiments testing general IR theories, with varying substantive focuses, replicated consistently across seven diverse countries without producing a single example of contradictory treatment effects suggests grounds for cautious optimism. Of course, future research can and should test whether our optimism is warranted, and we hope future scholarship will examine whether particular features of experiments might generate more versus less consistent results across country contexts.

Second, consistent with other replication studies (Coppock, Leeper and Mullinix, 2018; Coppock, 2019), we find that the most plausible explanation for our general pattern of results relates to limited treatment effect heterogeneity. This has implications for researchers who wish to theorize about the generalizability of their studies. We encourage scholars to interrogate – both theoretically and empirically – the extent to which their treatment effects are homogeneous versus heterogeneous. In practice, this entails theorizing ex-ante about variables that could moderate average treatment effects and incorporating measures of these moderators into the experimental design. Ex-post, researchers should employ treatment effect heterogeneity tests (Ding, Feller and Miratrix, 2016) to estimate whether their treatment effects are homogeneous and use these tests to inform arguments about generalizability. One caveat to this approach is that the analyses of treatment effect heterogeneity are often limited to the set of covariates collected by researchers – which further emphasizes the importance of considering and measuring moderators a priori. If treatment effects appear heterogeneous, scholars should carefully consider the distribution of the effects before making strong claims about generalizability. However, when treatment effects are generally homogeneous – as we find in our study – scholars may wish to make bolder claims regarding generalizability.

Third, our aggregation of results from a range of countries suggests that—however WEIRD Americans may be—the U.S. is not an outlier when it comes to experimental results on the micro-foundations of IR theories. Though American respondents may differ from respondents in other countries in terms of key demographic attributes (Henrich, Heine and Norenzayan, 2010*b*), and though Americans might be unusual in their foreign policy preferences (see Figure A8 in the Appendix), their reactions to informational treatments in four survey experiments testing the micro-foundations of general IR theories were not qualitatively unique. This insight is in line with other research identifying a strong degree of correspondence between different samples in political science experiments (Coppock, Leeper and Mullinix, 2018; Coppock, 2022*b*; Kertzer, 2022). Thus, while it remains true that past experimental work has focused rather narrowly on U.S.-based samples, we find little evidence to suggest that this reliance has led to wildly distorted conclusions about the micro-foundations of prominent theories of international relations.

Crucially, these findings also indicate that scholars testing general IR theories are potentially able to draw broad insights when they field studies in non-U.S. contexts, particularly if they follow our advice about measuring moderators and testing for treatment effect heterogeneity. This conclusion has important implications by improving the accessibility of experimental research for scholars based outside the U.S. as well as for de-centering the U.S. as the "standard" site for experimental research.

Our work also highlights a broader point about the merits of pre-registered and harmonized multi-site replication studies as a rigorous and transparent approach to accumulating knowledge and learning about the generalizability of experimental evidence in the field of international relations and in political science more broadly. In contrast to uncoordinated single-site replications, the coordinated approach we take sidesteps common challenges of design inconsistency that pose analytical hurdles for aggregating findings across contexts. Moreover, a coordinated approach limits the potential for selective reporting and file drawer problems, which ultimately result in publication bias. By allocating significant resources and coordinating multiple simultaneous replication studies across various countries, we can point to how specific findings generalize, pinpoint one local instance of failed replication, and substantiate our interpretation that broader patterns of generalizability are explained by effect homogeneity in IR experiments testing general, rather than conditional, theories. Our empirical exercise emphasizes the importance of pre-registered and harmonized multi-site replication as part of the research cycle in IR, and suggests that scholars should devote more theoretical and empirical attention to levels of effect heterogeneity (or lack thereof).

Importantly, we do not suggest that every experimental test of the micro-foundations of IR theories should attempt to report results from a wide range of countries. Indeed, doing so will often be prohibitively costly, and thus beyond the reach of many scholars. We see much merit in studies that provide rigorous evidence on the micro-foundations of clearly stipulated IR theories based on single-country experiments. Such studies should, however, theorize about potential sources of effect heterogeneity, test for heterogeneity whenever possible, and consider the implications of heterogeneity for the generalizability of their results. Our findings suggest that in many cases, one would be likely to find similar effects across diverse contexts. That said, pre-registered harmonized multi-site replications are an ideal approach to empirically explore the scope conditions of particular mechanisms. Thus researchers can—and should—pool resources to build on exciting single-country experiments and explore their empirical scope across a range of purposively varied countries. In that sense, we agree with previous accounts of generalizability that emphasize the importance of replication and accumulation of knowledge over time (McDermott, 2011*a*; Samii, 2016).

References

Allcott, Hunt. 2015. "Site selection bias in program evaluation." *The Quarterly Journal of Economics* 130(3):1117–1165.

Arechar, Antonio A, Jennifer Allen, Adam J Berinsky, Rocky Cole, Ziv Epstein, Kiran Garimella, Andrew Gully, Jackson G Lu, Robert M Ross, Yunhao Zhang Stagnaro et al. 2022. "Understanding and Combatting COVID-19 Misinformation Across 16 Countries on Six Continents.".

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

Aronow, Peter M, Jonathon Baron and Lauren Pinson. 2019. "A note on dropping experimental subjects who fail a manipulation check." *Political Analysis* 27(4):572–589.

Axelrod, Robert and William D Hamilton. 1981. "The evolution of cooperation." *Science* 211(4489):1390–1396.

Benjamini, Yoav and Yosef Hochberg. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal statistical society: series B (Methodological)* 57(1):289–300.

Bisbee, James and Jennifer M Larson. 2017. "Testing social science network theories with online network data: An evaluation of external validity." *American political science review* 111(3):502–521.

Blair, Graeme, Alexander Coppock and Margaret Moor. 2020. "When to worry about sensitivity bias: A social reference theory and evidence from 30 years of list experiments." *American Political Science Review* 114(4):1297–1315.

Bor, Alexander and Michael Bang Petersen. 2022. "The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis." *American political science review* 116(1):1–18.

Borenstein, Michael, Larry V Hedges, Julian PT Higgins and Hannah R Rothstein. 2021. *Introduction to meta-analysis*. John Wiley & Sons.

Brooks, Deborah Jordan, Stephen G Brooks, Brian D Greenhill and Mark L Haas. 2018. "The demographic transition theory of war: why young societies are conflict prone and old societies are the most peaceful." *International Security* 43(3):53–95.

Brutger, Ryan, Joshua D Kertzer, Jonathan Renshon and Chagai M Weiss. 2022. *Abstraction in experimental Design: testing the tradeoffs*. Cambridge University Press.

Brutger, Ryan, Joshua D Kertzer, Jonathan Renshon, Dustin Tingley and Chagai M Weiss. 2022. "Abstraction and Detail in Experimental Design." *American Journal of Political Science* .

Chilton, Adam S, Helen V Milner and Dustin Tingley. 2020. "Reciprocity and public opposition to foreign direct investment." *British Journal of Political Science* 50(1):129–153.

Clifford, Scott, Geoffrey Sheagley and Spencer Piston. 2021. "Increasing Precision without Altering Treatment Effects: Repeated Measures Designs in Survey Experiments." *American Political Science Review* 115(3):1048–1065.

Clifford, Scott and Jennifer Jerit. 2015. "Do attempts to improve respondent attention increase social desirability bias?" *Public Opinion Quarterly* 79(3):790–802.

Colgan, Jeff D. 2019*a*. "American bias in global security studies data." *Journal of Global Security Studies* 4(3):358–371.

Colgan, Jeff D. 2019*b*. "American perspectives and blind spots on world politics." *Journal of Global Security Studies* 4(3):300–309.

Coppock, Alexander. 2019. "Generalizing from survey experiments conducted on Mechanical Turk: A replication approach." *Political Science Research and Methods* 7(3):613–628.

Coppock, Alexander. 2022*a*. Persuasion in parallel. In *Persuasion in Parallel*. University of Chicago Press.

Coppock, Alexander. 2022*b*. "Persuasion in parallel." *Chicago studies in American politics* .

Coppock, Alexander and Donald P Green. 2015. "Assessing the correspondence between experimental results obtained in the lab and field: A review of recent social science research." *Political Science Research and Methods* 3(1):113–131.

Coppock, Alexander, Seth J Hill and Lynn Vavreck. 2020. "The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments." *Science advances* 6(36):eabc4046.

Coppock, Alexander, Thomas J Leeper and Kevin J Mullinix. 2018. "Generalizability of heterogeneous treatment effect estimates across samples." *Proceedings of the National Academy of Sciences* 115(49):12441–12446.

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

De Mesquita, Bruce Bueno, James D Morrow, Randolph M Siverson and Alastair Smith. 1999. "An institutional explanation of the democratic peace." *American Political Science Review* 93(4):791–807.

Devaux, Martin and Naoki Egami. 2022. "Quantifying Robustness to External Validity Bias." *Available at SSRN 4213753* .

Dimock, Michael and Samuel L Popkin. 1997. "Political knowledge in comparative perspective." *Do the media govern* pp. 217–24.

Ding, Peng, Avi Feller and Luke Miratrix. 2016. "Randomization inference for treatment effect variation." *Journal of the Royal Statistical Society: Series B: Statistical Methodology* pp. 655–671.

Downs, George W and Michael A Jones. 2002. "Reputation, compliance, and international law." *The Journal of Legal Studies* 31(S1):S95–S114.

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D Hyde, Craig McIntosh and Gareth Nellis. 2019. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*. Cambridge University Press.

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D Hyde, Craig McIntosh, Gareth Nellis, Claire L Adida, Eric Arias, Clara Bicalho, Taylor C Boas et al. 2019. "Voter information campaigns and political accountability: Cumulative findings from a preregistered meta-analysis of coordinated trials." *Science advances* 5(7):eaaw2612.

Egami, Naoki and Erin Hartman. 2022. "Elements of external validity: Framework, design, and analysis." *American Political Science Review* .

Fariss, Christopher J. 2019. "Yes, human rights practices are improving over time." *American Political Science Review* 113(3):868–881.

Fearon, James D. 1994. "Domestic political audiences and the escalation of international disputes." *American Political Science Review* 88(3):577–592.

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

Frederiksen, Kristan Vrede Skaaning. 2022. "Does Competence Make Citizens Tolerate Undemocratic Behavior?" *American Political Science Review* 116(3):1147–1153.

Hainmueller, Jens, Andrew B Hall and James M Snyder Jr. 2015. "Assessing the external validity of election RD estimates: An investigation of the incumbency advantage." *The Journal of Politics* 77(3):707–720.

Hainmueller, Jens and Michael J Hiscox. 2010. "Attitudes toward highly skilled and low-skilled immigration: Evidence from a survey experiment." *American Political Science Review* 104(1):61–84.

Hendrix, Cullen S and Jon Vreede. 2019. "US Dominance in international relations and security scholarship in leading journals." *Journal of Global Security Studies* 4(3):310–320.

Henrich, Joseph, Steven J Heine and Ara Norenzayan. 2010*a*. "Beyond WEIRD: Towards a broad-based behavioral science." *Behavioral and brain sciences* 33(2-3):111.

Henrich, Joseph, Steven J Heine and Ara Norenzayan. 2010*b*. "Most people are not WEIRD." *Nature* 466(7302):29–29.

Hoffmann, Stanley. 1977. An American Social Science: International Relations. In *International Relations — Still An American Social Science? Toward Diversity in International Thought*, ed. Robert M. A. Crawford and Daryl S.L. Jarvis. State University of New York Press pp. 212–241.

Humphreys, Macartan and Alexandra Scacco. 2020. "The aggregation challenge." *World Development* 127:104806.

Hyde, Susan D. 2015. "Experiments in international relations: Lab, survey, and field." *Annual Review of Political Science* 18:403–424.

Jones, Dan. 2010. "A WEIRD view of human nature skews psychologists' studies.".

Kang, David C and Alex Yu-Ting Lin. 2019. "US bias in the study of Asian security: Using Europe to study Asia." *Journal of Global Security Studies* 4(3):393–401.

Keele, Luke, Corrine McConnaughy and Ismail White. 2012. "Strengthening the experimenter's toolbox: Statistical estimation of internal validity." *American Journal of Political Science* 56(2):484–499.

Keohane, Robert O. 1984. *After hegemony*. Princeton University Pressniversity press.

Kertzer, Joshua D. 2016. *Resolve in international politics*. Princeton University Press.

Kertzer, Joshua D. 2017. "Microfoundations in international relations." *Conflict Management and Peace Science* 34(1):81–97.

Kertzer, Joshua D. 2022. "Re-assessing elite-public gaps in political behavior." *American Journal of Political Science* 66(3):539–553.

Kertzer, Joshua D and Ryan Brutger. 2016. "Decomposing audience costs: Bringing the audience back into audience cost theory." *American Journal of Political Science* 60(1):234–249.

Kristensen, Peter M. 2012. "Dividing discipline: Structures of communication in international relations." *International Studies Review* 14(1):32–50.

Kristensen, Peter Marcus. 2015. "Revisiting the "American social science"—Mapping the geography of international relations." *International Studies Perspectives* 16(3):246–269.

Levin, Dov H and Robert F Trager. 2019. "Things you can see from there you can't see from here: blind spots in the American perspective in IR and their effects." *Journal of global security studies* 4(3):345–357.

Levitsky, Steven and Daniel Ziblatt. 2019. *How democracies die*. Crown.

Lupu, Yonatan and Geoffrey PR Wallace. 2019. "Violence, nonviolence, and the effects of international human rights law." *American Journal of Political Science* 63(2):411–426.

McDermott, Rose. 2011*a*. "Internal and external validity." *Cambridge handbook of experimental political science* pp. 27–40.

McDermott, Rose. 2011*b*. "New directions for experimental work in international relations." *International Studies Quarterly* 55(2):503–520.

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

Mutz, Diana C. 2021. "Improving experimental treatments in political science." *Advances in Experimental Political Science* 219.

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

Offer-Westort, Molly, Alexander Coppock and Donald P Green. 2021. "Adaptive experimental design: Prospects and applications in political science." *American Journal of Political Science* 65(4):826–844.

Powers, Kathleen E. 2022. *Nationalisms in international politics*. Vol. 30 Princeton University Press.

Renshon, Jonathan. 2015. "Losing face and sinking costs: Experimental evidence on the judgment of political and military leaders." *International Organization* 69(3):659–695.

Renshon, Jonathan, Keren Yarhi-Milo and Joshua D Kertzer. 2023. "Democratic reputations in crises and war." *The Journal of Politics* 85(1):000–000.

Rosato, Sebastian. 2005. "Explaining the democratic peace." *American Political Science Review* 99(3):467–472.

Samii, Cyrus. 2016. "Causal empiricism in quantitative research." *The Journal of Politics* 78(3):941–955.

Schultz, Kenneth A. 2001. "Looking for audience costs." *Journal of Conflict Resolution* 45(1):32–60.

Schwarz, Susanne and Alexander Coppock. 2022. "What Have We Learned about Gender from Candidate Choice Experiments? A Meta-Analysis of Sixty-Seven Factorial Survey Experiments." *The Journal of Politics* 84(2):655–668.

Shadish, William R, Thomas D Cook and Donald T Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference.* Houghton, Mifflin and Company.

Simmons, Beth. 2010. "Treaty compliance and violation." *Annual Review of Political Science* 13:273–296.

Slough, Tara and Scott A Tyson. 2021. "External Validity and Meta-analysis." *American Journal of Political Science* .

Suong, Clara H, Scott Desposato and Erik Gartzke. 2020. "How'Democratic'is the Democratic Peace? A Survey Experiment of Foreign Policy Preferences in Brazil and China." *Brazilian Political Science Review* 14.

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

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

Wæver, Ole. 1998. "The sociology of a not so international discipline: American and European developments in international relations." *International Organization* 52(4):687–727.

Wallace, Geoffrey PR. 2013. "International law and public attitudes toward torture: An experimental study." *International Organization* 67(1):105–140.

Walt, Stephen M. 2005. "The relationship between theory and policy in international relations." *Annual Review of Political Science* 8:23–48.

Wolfers, Arnold. 1947. "International Relations as a Field of Study." *Columbia Journal of International Affairs* 1(1):24–26.

Wolford, Scott. 2007. "The turnover trap: New leaders, reputation, and international conflict." *American Journal of Political Science* 51(4):772–788.

The Generalizability of IR Experiments

Supplementary Information

A Main Figures in Table Form

In Tables A1 and A2 we report the findings from our main Figures reported in the text.

B Selecting Studies to Replicate

We identified studies that test the micro-foundations of general IR theories, employing relatively simple designs, producing robust effects, and making general theoretical claims that should apply beyond the U.S. We further chose experiments that cross substantive boundaries and research programs: theories of international security and war, international law and human rights, and international political economy. Below, we briefly describe each experiment.

Study I: Democratic Peace Experiment. Democratic Peace theory is a broad theoretical framework predicting that democracies are less likely to engage in conflict with other democracies (De Mesquita et al., 1999; Rosato, 2005). One version of this argument, tested experimentally by Tomz and Weeks (2013), is that an adversary's regime type (i.e., democracy or non-democracy) affects democratic citizens' support for conflict by shaping beliefs about threat and the normative and material costs of conflict. We test whether citizens are less likely to support initiating conflict in a hypothetical vignette when the country is described as a democracy rather than a non-democracy.

Study II: Domestic Audience Costs Experiment. This prominent theoretical framework argues that democratic leaders pay an electoral cost – a domestic audience cost – for backing down from public statements (Fearon, 1994), lending credibility to democracies' threats (Schultz, 2001). In an experimental test of the theory's micro-foundations, Kertzer and Brutger (2016) demonstrate that failing to follow through on a threat reduces public support for leaders, because the public could punish leaders either for revealing their belligerence or for inconsistency between their statements and behaviors. In our primary analyses, we test whether respondents' approval of a leader's performance in a hypothetical scenario declines when the leader issues a threat on which they do not follow through, as opposed to not issuing a threat in the first place. In secondary analyses reported in Appendix J, we decompose the different elements of audience costs.

Study III: International Law and Torture Experiment. Scholars often argue that international laws and treaties influence state policies by shaping popular reactions (Simmons, 2010). Wallace (2013) used a survey experiment to identify the effects of information regarding international law on support for torture. The study provided respondents with a vignette describing torture as a method for obtaining information from captured combatants, randomized whether respondents were informed that torture violates principles of international law to which the U.S. is committed through multiple treaties, and then measured support for using torture. Receiving information about the illegality of torture reduced support for this policy option. We replicated a slightly simplified version of Wallace's original instrument.

Study IV: FDI Reciprocity Experiment. Foundational research in international relations theorizes that reciprocity induces cooperative behavior (Axelrod and Hamilton, 1981; Keohane, 1984). Chilton, Milner and Tingley (2020) fielded several survey experiments in the U.S. and China to test whether reciprocity shapes public opinion on the regulation of foreign investments. In one experiment, Chilton, Milner and Tingley (2020) tell subjects that a foreign country has either made it harder or easier for external companies to acquire local companies and then measure whether respondents think their own country should make foreign acquisition of local companies harder or easier. Chilton, Milner and Tingley (2020) find that respondents' policy preferences follow a reciprocity rationale, rewarding foreign countries who reduce barriers to trade. We replicate a simplified version of the vignette presented in Chilton, Milner and Tingley (2020).

C Selecting Experimental Sites (Countries)

To select our cases, we followed the following steps:

1. Determining Scope Conditions. After parsing the theories, we identified scope conditions, the full set of cases to which a theory is claimed to be applicable (Findley, Kikuta and Denly, 2021). Given our goals, we focused on countries explicitly *within* the stated scope of a given theory, based on the authors' own claims about where a hypothesis should apply. For example, the democratic peace and audience costs studies hypothesize that voters in democracies should behave in specific ways. They limit the scope of their theoretical prediction to democracies, but do not place any further limits on scope, such as specifying that the prediction should apply only to democracies with certain other qualities. While the international law study does not explicitly limit the theoretical scope to democracies, it justifies its focus on public opinion by highlighting the importance of domestic constituents in democratic countries, so it seems most appropriate to test that finding in democracies, as well. The authors of the FDI reciprocity experiment, meanwhile, specified that the theory is applicable regardless of regime type. However, given that public opinion may play a larger role in democracies, and in light of our plan to replicate multiple experiments within each site, we opted to focus on countries that satisfy the scope of all experiments—i.e. democracies—and excluded countries that score below the minimum threshold democracy score (Polity score of ≥ 6).

2. Sorting by Policy Importance. Another potential criterion is policy relevance. To the extent that the goal of IR theory is to explain how global politics work, it may be more useful to verify that IR theories can explain domestic preferences within powerful countries that are more likely to shape global dynamics rather than preferences in isolated and weak nations. This is because global powers tend to shape patterns of security and economic relations to a greater extent than less powerful, smaller countries. For this reason, we sorted all countries meeting our initial scope condition (i.e., democracies) based on GDP, and prioritized more powerful countries over less powerful ones, all else equal (without sacrificing variation on key moderators, which we address in the next step).[18]

3. Maximize Variation along Unobserved Factors by Selecting Countries from each Major Region around the World. After sorting countries by GDP, we select the most powerful country from different regions around the world. Doing so ensures that we maximize variability and heterogeneity along unmeasured factors such as culture and religion.

4. Verifying Variation Across Theoretically Important Moderators. For three of the four studies, our interpretation of existing papers revealed theoretically-relevant moderators. For example, “strength of democratic norms” is a potential moderator in the democratic peace experiment. Similarly, hawkishness is a key moderator in the audience costs experiment. Obligation to international law is a potential moderator in the international law experiment. Our theoretical analysis of the FDI reciprocity study, meanwhile, did not suggest any key moderators. By selecting cases that display variation in potential moderators, we render the range assumption more plausible, and we can increase our knowledge about the generalizability of theories. Moreover, we can carry out exploratory tests of moderation effects at the individual level. This can help place existing evidence in perspective, informing our interpretation of any cross-site variation in average treatment effects. Thus, we verify that our selected countries vary across the moderating variables we specified above, with at least two countries below and two countries above the cross-national mean for each moderating variable.

4. Considering Practical Constraints. Finally, we checked that our case selection yields a consistent approach to data collection across sites. In order to maximize comparability across countries, we worked with one commonly-used platform — Lucid/Cint. We thus verified that Lucid/Cint operates in the countries we selected and would be able to match the sample on key demographics (i.e., gender and age) of the general population in each country of interest. This step did not constrain our case selection procedure as Lucid/Cint was able to offer samples for all countries on our final list, depicted in Figure 2 of the main text: Brazil, Germany, India, Israel, Japan, and Nigeria alongside the U.S.

D Survey Instrument

To implement our study, We programmed our survey on Qualtrics. In all countries except for Nigeria and the U.S., we provided respondents with the option of responding to our survey in either English or the most common native language. Translations were implemented by native translators and evaluated by academics with relevant language proficiency. As a final step, we reverse-translated all instruments via google translate to ensure that no technical errors exist.

Our survey included three components. First, respondents were provided with an informed consent form. Second, after agreeing to participate in our study, respondents answered questions relating demographic variables and moderators. Importantly, in this section, we embedded six attention checks. Failing one or more attention checks resulted in survey termination. Finally, after reporting demographics and moderators, respondents participated in our four experiments, as well as a final auxiliary study implemented by one of the coauthors. Below, we provide an English-language overview of our survey:

[18] Of course, power itself is a potential moderator, though its predicted effect is not clear for the studies we replicate. Our approach nonetheless provides variation with respect to military expenditure, as shown in Figure 2 of the main text.

- **Informed consent**
- **Pre-Treatment Demographics and Moderators:**
  - Below, you will see a series of statements. Please tell us whether you agree or disagree with each statement.
    - 2+2=7
    - Please click the “neither agree nor disagree” response
    - The year 1910 came before the year 1920
    - I would rather be a citizen of my country, than of any other country in the world
    - The use of military force only makes problems worse
  - Below, you will see a series of statements. Please tell us whether you agree or disagree with each statement.
    - My country needs to play an active role in solving conflicts around the world
    - The best way to ensure peace is through military strength
    - Going to war is unfortunate, but sometimes the only solution
    - Please select “somewhat agree”
  - For each of the items below, please choose the response that is closest to your view.
    - I do not mind a politician’s methods if he or she manages to get the right things done.
    - When the country is in great danger, it is often necessary for political leaders to act boldly, even if this means overstepping the usual processes of government decision-making.
    - People should be allowed to vote even if they are badly misinformed on basic facts about politics.
    - People who hate my way of life should still have a chance to talk in a public forum.
    - 2+2=4
  - How strongly do you agree or disagree with the following statements:
    - Complying with international law is an important value.
    - Complying with international law is important, even if it contradicts the national interest.
    - If my country defies international laws and norms, criticism from other countries is justified.
    - 4+3=8
  - In politics people often talk of “left” and “right”. On a scale of 1 (left) and 10 (right), where would you classify your own political views?
  - Are you eligible to vote in the United States? (Y/N)
  - What is the highest level of education you have attained? (Some high school/Some high school/High school graduate/Some college or associate degree/Bachelor’s or equivalent/Master’s or equivalent/Doctoral or equivalent)
  - Now we have a few questions about international organizations. Many people don’t know the answers to these questions. We want to see how much information about international organizations gets out to the public from television, newspapers, and the like. It is important to us that you do NOT use outside sources like the Internet to search for the correct answer. Will you answer the following questions without help from outside sources? (Y/N)
  - Five countries have permanent seats on the Security Council of the United Nations. Which one of the following is not a permanent member? (China/France/India)
  - To what degree are you worried about the following situations?
    - A war involving my country
    - A terrorist attack

- **Experiments:**
  - Now we are going to ask your opinion about some situations that the United States could face in the future. Please read the text carefully and give us your opinions.
    - Randomize order of 4 main experiments (Democratic Peace, Audience Costs, International Law, FDI Reciprocity), followed by the Bassan-Nygate study. Vignette text is provided in Section D.1

D.1 Experimental Vignette

D.1.1 Democratic Peace

Thank you for your response. We will now ask you about a new hypothetical situation.

There is much concern these days about the spread of nuclear weapons. We are going to describe a hypothetical situation that many countries around the world have faced in the past and the United States could face in the future.

For scientific validity, the hypothetical situation is general, and is not about a specific country in the news today. Some parts of the description may strike you as important; other parts may seem unimportant. After describing the situation, we will ask your opinion about a policy option.

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

- The country has not signed a military alliance with the U.S.
- The country does not have high levels of trade with the U.S.
- The country's nonnuclear military forces are half as strong as the U.S. nonnuclear forces.
- The country is a democracy/not a democracy and shows every sign that it will remain a democracy/no sign of becoming a democracy.
- The country's motives remain unclear, but if it builds nuclear weapons, it will have the power to blackmail or destroy other countries.
- The country has refused all requests to stop its nuclear weapons program.

Outcomes and Diagnostics

- By attacking the country's nuclear development sites now, the U.S. could prevent the country from producing any nuclear weapons. Would you favor or oppose using the U.S. military to attack the country's nuclear development sites? (5-point scale)
- By joining a joint international military mission now, the U.S. could prevent the country from producing any nuclear weapons. Would you favor or oppose the U.S. joining a joint international military mission? (5-point scale)
- Did you think of a specific country when you read about the "other country" in the passage? If so, please specify: (Y, specify/N)
- We just asked you to read a scenario in which the U.S. considered preventing another country from producing nuclear weapons. How believable is this situation for the U.S.? (Very believable/Somewhat believable/Not very believable/Not at all believable)
- In the passage you just read, the country developing nuclear weapons was: (A democracy/Not a democracy)

D.1.2 Audience Costs

Thank you for your response. We will now ask you about a new hypothetical situation.

We will now present to you a hypothetical scenario regarding the United States' relations with other countries around the world. You will read about a situation that many countries have faced in the past and will probably face again in the future. Different leaders have handled such situations in different ways. We will describe one approach that leaders have taken, and ask whether you approve or disapprove of the approach.

Imagine that a country sent its military to take over some of its neighbor's territory.

The U.S. President announced that the U.S. would stay out of the conflict. The attacking country continued to invade and the conflict ended with the attacking country taking control of 20% of the contested territory. / The U.S. President announced that if the attacking country continued to invade, the U.S. military would immediately engage and attempt to push out the attacking country. The President sent troops to the region. The attacking country continued to invade and the President ordered the U.S. military to engage. The U.S. did not lose any troops in the conflict and the conflict ended with the attacking country taking control of 20% of the contested territory. / The U.S. President announced that if the attacking country continued to invade, the U.S. military would immediately engage and attempt to push out the attacking country. The President sent troops to the region. The attacking country continued to invade. The President ordered the U.S. military not to engage. The attacking country continued to invade and the conflict ended with the attacking country taking control of 20% of the contested territory.

Outcomes and Diagnostics

- How much would you approve or disapprove of the way the U.S. President handled the situation? (7-point scale)
- Did you think of a specific country when you read about the country in the passage that attacked its neighbor? If so, please specify: (Y, specify/N)
- We just asked you to read a scenario in which the U.S. considered intervening in a foreign dispute. How believable is this situation for the U.S.? (Very believable/Somewhat believable/Not very believable/Not at all believable)
- In the passage you just read, the U.S. President: (Announced the United States [or other country] would stay out of the conflict and stayed out. / Announced that the United States [or other country] military would engage in the conflict, but ultimately did not engage. / Announced that the United States [or other country] military would engage in the conflict, and eventually engaged.)

D.1.3 International Law

Thank you for your response. We will now ask you about a new hypothetical situation.

We will now present you with a hypothetical scenario about how countries treat their enemies' combatants. In military conflicts ranging from World War I to the present, it is common for countries to capture combatants from the opposing side. These combatants may have information of interest for the conflict, such as the opposing side's troop movements or plans for future attacks. Some officials believe interrogating captured combatants through a variety of methods is a useful way to obtain this information.

Imagine that the U.S. is involved in a military conflict with another country. The other country is not a democracy. We will describe some interrogation methods that the U.S. might use. Please consider each of the following points carefully and then tell us what you think.

The interrogation methods would be used on captured combatants that are not part of regular armed forces. The combatants are often organized in small units, do not use standard uniforms or carry their weapons openly, and conduct operations geared toward subversion. The interrogation methods would involve torture, meaning they would cause severe pain or suffering to the captured combatants. The information gained from the interrogation may, or may not, be accurate or relevant to the conflict. The interrogation methods would violate international law. The U.S. has signed international treaties that do not allow the use of these methods under any circumstances against any captured combatants. / The interrogation methods would be used on captured combatants that are not part of regular armed forces. The combatants are often organized in small units, do not use standard uniforms or carry their weapons openly, and conduct operations geared toward subversion. The interrogation methods would involve torture, meaning they would cause severe pain or suffering to the captured combatants. The information gained from the interrogation may, or may not, be accurate or relevant to the conflict.

Outcomes and Diagnostics

- If things happened just as we described, to what extent would you agree or disagree with the following statement: The U.S. should use interrogation methods involving torture on captured combatants. (7-point scale)
- Did you think of a specific country when you read about the country that was in a dispute with the U.S.? If so, please specify: (Y, specify/N)
- We just asked you to read a scenario in which the U.S. was considering using interrogation methods. How believable is this scenario for the U.S.? (Very believable/Somewhat believable/Not very believable/Not at all believable)
- Did the passage you read mention that the U.S. signed international treaties that do not allow the use of torture in interrogation methods? (Y/N)

D.1.4 Reciprocity and FDI

Thank you for your response. We will now ask you about a new hypothetical situation.

We will now present a hypothetical scenario about U.S. policies involving the ability of foreign companies to purchase U.S. companies. The situation is hypothetical but may reflect something that has happened in the past or could happen in the future.

Imagine that the U.S. is considering changing its policies on the ability of foreign companies to purchase U.S. companies. Another country, a democracy, has recently made it much easier for U.S. companies to buy companies in their own country. / Imagine that the U.S. is considering changing its policies on the ability of foreign companies to purchase U.S. companies. Another country, a democracy, has recently made it much harder for U.S. companies to buy companies in their own country.

Outcomes and Diagnostics

- Should the U.S. make it easier or harder for companies from this country to buy U.S. companies? (The U.S. should make it much easier / The U.S. should make it somewhat easier / The U.S. should make no change / The U.S. should make it somewhat harder / The U.S. should make it much harder)
- Did you think of a specific country when you read about the “other country” in the passage? If so, please specify: (Y, specify / N)
- We just asked you to read a scenario in which the U.S. was considering changing its policies on the ability of foreign companies to purchase U.S. companies. How believable is this situation for the U.S.? (Very believable/Somewhat believable/Not very believable/Not at all believable)
- In the passage you read, the other country has made it: (Much easier for U.S. companies to buy companies in their country. / Much harder for U.S. companies to buy companies in their country.)

D.2 Deviations from Original Surveys

In Table A3, we report several minor differences between our instrument and the original studies we replicate. We committed several deviations to ensure that experiments are presented in a simplified manner, maximizing power and consistency across studies. Importantly, despite these deviations, our findings are consistent with the original studies.

E Descriptive Statistics

In this section we report aggregate descriptive statistics of our cross-national sample, as well as country-specific descriptive statistics tables.

F Diagnostics

In this Section, we report several diagnostic tests relating to our four experiments. First, in Table A5 we report manipulation checks to assess treatment take-up. To do so, we regress responses to our factual manipulation checks over respondents' treatment status. Doing so, we demonstrate that in the democratic peace experiment, respondents assigned to the democracy condition are more likely to report that the country described in their vignette was a democracy. In the audience cost experiment, respondents assigned to the back-down condition are more likely to report that the leader described in the vignette backed down on a threat. In the international law experiment, respondents receiving information about a treaty signed by their country were more likely to report that their country signed a treaty banning the use of torture. Finally, in the reciprocity treatment, respondents receiving information that a country made it harder for companies from their country to buy companies in another country were more likely to report that said country had increased barriers for companies from their country to buy companies in the other country.

In Figure A1, we report data on respondents' evaluations of scenario plausibility by country and experiment. Respondents have mostly found scenarios plausible with the exception of the international law experiment, where we speculate that respondents interpreted the question as being about the plausibility that their country will use torture, rather than consider using it. In Figure A2 we report the response time to our full survey by country. In Table A6, we examine whether our treatments in all four experiments increased the probability of reporting specific countries when asked whether the respondents thought of a specific country while reading the vignette. In the democratic peace experiment, we find some evidence that respondents in the non-Democracy condition were more likely to report that they thought of a specific country. We examine reported countries by country, experiment, and treatment condition in Figures A3-A6. As further discussed, in Appendix K we find no evidence for clear differences in countries between treatment and control conditions at the country-experiment level. Finally, in Figure A7 we report treatment and control means (and 95% confidence level) for each experiment-country combination as well as for our full sample and the original studies we replicate. This figure, as well as the other diagnostics detailed above, are discussed when interrogating the India democratic peace null result in Appendix K.

G Heterogeneous Treatment Effects

In Figure A8, we report the distribution of individual-level moderators across countries. As expected, we uncover significant variation along key theoretical dimensions. We thus explore treatment effect heterogeneity along these dimensions, in our full sample, in Figure A9-A11, as well as in Table A7.

As expected, we find that support for democratic norms moderates the effects of democracy on support for conflict. Democracy has a larger negative effect on support for war among people with higher levels of support for democratic norms. We do not find much evidence that hawkishness moderates the main treatment in the audience cost experiment. However, we show meaningful and consequential treatment effect heterogeneity when we decompose the treatment into belligerence and inconsistency costs in Appendix J. Finally, we find evidence in support of moderation when focusing on the legal obligation index. In other words, respondents with high levels of legal obligations are more opposed to the use of torture when assigned to the information treatment regarding government commitment to a treaty banning torture.

H Sensitivity to External Validity Bias

In line with an overwhelming majority of survey experiments in political science, we employ a range of convenience samples across countries. Previous investigations suggest that doing so, does not have substantial consequences for the main inferences we draw (Coppock, Leeper and Mullinix, 2018). However, in this section, we implement a general tests to consider sensitivity to external validity bias. Specifically, we follow Devaux and Egami (2022) and examine the sensitivity of our main results to external validity bias, and consider the extent to which reweighing our sample using different covariate profiles would explain away identified treatment effects. These analyses lend insight to the following question: how different would a nationally representative sample need to be in order to change our substantive findings? Our results from this exercise suggest that in most country-experiment pairs, employing samples with varying covariate profiles would not explain away identified effects. In figure A12, we plot the estimated external robustness and the distribution of estimated CATEs for each study. We mark in red any cases in which the estimated external robustness is below the proposed upper-bound benchmark by Devaux and Egami (2022) (0.57).

I Democratic Peace Extension

In this section, we report an extension to our original democratic peace experiment. Specifically, we use an alternative outcome measuring respondents' support for their country joining a joint international military mission that would prevent the country from producing any nuclear weapons. We introduced this secondary outcome due to a concern regarding floor effects, by which respondents from weaker countries may be hesitant to support unilateral foreign intervention but might consider a multilateral one. The results in Table A9 using this alternative outcome measure are largely consistent with the results presented in the main text.

J Audience Costs Extension

In this Section, we report a series of pre-registered secondary analyses in which we decompose the general audience cost treatment into two components: belligerence costs (i.e., the costs or rewards citizens impose on leaders for issuing threats rather than remaining aloof) and inconsistency costs (i.e., the cost citizens impose on leaders for not following through on threats). Notably, as theorized and demonstrated by Kertzer and Brutger (2016), such costs may vary as a function of individual and situational factors. For example, they find that doves punish belligerence while hawks reward it. Other individual factors could include risk aversion or other dispositional variables that could shape respondents' views on using force in a particular situation. Situational factors would include variables, including those that vary across either vignettes, countries, or time, that influence how respondents perceive the costs and benefits of intervening versus staying out in particular situation.

In Figure A13, we report our main estimates for these additional analyses. We find broad support for inconsistency costs – point estimates from all countries, as well as our meta-analytic ATE, are directionally similar to the original point estimates from Kertzer and Brutger (2016). However, when estimating belligerence costs, we find substantial variation across countries in ATEs, which yield a null meta-analytic ATE.

As we argue in Section 4.4 of the main text, treatment effect heterogeneity likely explains why the belligerence treatment yields diverging effects across countries. Indeed, in their theory, Kertzer and Brutger (2016) argue that the ATE of belligerence — support for using force versus support for remaining out of the conflict altogether — should vary across subjects depending on their level of hawkishness. More hawkish subjects should be more likely to reward leaders who use force, while more dovish subjects should be more likely to punish belligerent leaders. We confirm this prediction in Figure A14. The belligerence treatment is the only treatment in our study for which a given theoretically motivated individual-level moderator (i.e., hawkishness) shapes not only the magnitude but also the direction of ATEs. As shown in the left-hand side of Figure A14, belligerence reduces leader support among respondents' reporting low levels of hawkishness and increases leader support among respondents reporting high levels of hawkishness. Moreover, as discussed in Section 4.4 of the main text, homogeneity tests proposed by Ding, Feller and Miratrix (2016) produce strong evidence of heterogeneity in all countries with regard to the belligerence treatment.

Given this evidence, we conclude that much of the cross-country variation in reactions to belligerence reported in Figure A14 is due to individual-level treatment effect heterogeneity originally theorized and empirically demonstrated by Kertzer and Brutger (2016). Since individual attributes both moderate responses to treatment and vary substantially across countries, the effect of belligerence varies across countries. That said, while hawkishness appears to contribute to treatment effect heterogeneity, other unmeasured individual-level moderators may also play a role, as could situational factors such as current events that potentially influenced interpretations of the vignette.[19]

[19] We suspect that at least two results from Figure A13 cannot be explained by hawkishness-induced heterogeneity alone. For example, we observe rewards for belligerence in the U.S. replication (in contrast to a negative effect in the original U.S. study), and the Israeli sample tends to punish belligerence even though it is relatively hawkish. Though we cannot provide conclusive evidence either way, one possibility is that these patterns are due to current events and country-level variables shaping respondents' views about the utility of using force versus staying out in the hypothetical vignette, which describes a situation in which a country invades a neighbor. In the U.S. sample, it is possible that ongoing U.S. engagement in the Russia-Ukraine war made the "engage" option more popular, and the "stay out" option less popular, compared to the original U.S. study. In the Israeli context, we suspect that other unmeasured factors (e.g., Israelis seeing little national interest in intervening in far-off disputes, given their country's own security challenges) may explain why Israelis punish belligerent leaders despite being relatively hawkish. We emphasize that these interpretations are only suggestive and encourage researchers to build on our findings and the insights of Kertzer and Brutger (2016) to further examine the conditions under which belligerence provokes punishments versus rewards.

K Probing the Null: Explaining the Absence of Democratic Peace in India

Our findings suggest that the micro-foundations of the democratic peace theory did not generalize to our India sample. As we note in section 4 of our manuscript, the effect of our democracy treatment on supporting an attack amongst our India sample was null ($p = 0.82$). However, we designed our study in a way that would allow us to probe such results, and in this section we review and evaluate several potential explanations:

1. *Implausible scenario:* One explanation for a null result may be that respondents in India found the democratic peace scenario implausible. That is, the idea that India would face a situation in which it considered attacking another country for pursuing nuclear weapons is not realistic – either when compared to other countries, or in comparison to other studies fielded in India. We conclude that this explanation is improbable since, as reported in Figure A1, over 85% of respondents in India said the scenario is plausible. This score is high both in absolute terms, and in comparison to other countries, and is consistent with other studies fielded in India.

2. *Information leakage:* Respondents in India may have had a particular country in mind while reading the vignette – a version of confounding (Dafoe, Zhang and Caughey, 2018) – either across experimental conditions or differentially. First, we do not find evidence for differential beliefs about the country in the scenario. In Figure A3 we demonstrate that respondents in India thought of similar countries across both conditions, with most respondents thinking of Pakistan and China. However, it is possible that if respondents in India always thought of an adversary like Pakistan, then perhaps they were prone to strike in both experimental conditions, muting the treatment effect. We note that in other countries – Israel and Japan – the proportion of respondents who name the same country (Iran and North Korea, respectively) was much higher in comparison to India, making them more obvious candidates for muted effects due to confounding. Nonetheless, it is possible that the 'true effect' of democracy in Israel and Japan is much larger than in India, allowing us to identify the effect regardless of information leakage. We are thus unable to fully rule out information leakage as a potential explanation.

3. *Floor or ceiling effects:* We examine whether our sample in India is prone to floor or ceiling effects due to particularly high or low levels on our outcome of interest – support for attacking the other country's nuclear facilities. We determine that this is an improbable explanation for two reasons. First, while the mean of the India sample on our main outcome in the democratic peace experiment was relatively high (3.75 on a scale of 1 to 5) it is not as high as the mean in the Israel sample (3.99) or as low as the mean in the Japan sample (2.33) which would be more obvious candidates for ceiling and floor effects, respectively (see Figure A7). Second, we also report a null effect in India on an alternative outcome, asking respondents whether they supported joining a joint international mission (see Table A9).

4. *Inattentive sample:* Another explanation for our null result in India may be that respondents in India were much less attentive when compared to samples in other countries and have thus failed to take-up the treatment, biasing effects towards zero. There is some evidence to suggest that our sample in India was less attentive than samples in other countries. First, a larger proportion of subjects in India failed our pretreatment screeners. This suggests that the broader pool of subjects in India from which our sample was drawn was less attentive, and if we assume that our pretreatment screeners were imperfect then it is likely that the subjects who managed to pass our screeners were also less attentive. Second, as is evident from Table A5, subjects from India passed our manipulation checks at substantially lower rates than subjects from other countries. While subjects in India passed manipulation checks at lower rates across all four studies, it is possible that the 'true effect' in the democratic peace experiment in India was particularly low in comparison to the other studies. Since it is not advisable to drop experimental subjects who fail manipulation check (Aronow, Baron and Pinson, 2019) we screen out respondents who have failed manipulation checks in the *other* studies, using them as a proxy (albeit imperfect) for attentiveness. While this slightly increases our estimate (to -0.03) and reduces our p value ($p = 0.69$), we still report null effects (Table A11). Hence, while we cannot rule it out completely, we conclude that attentiveness cannot serve as the sole explanation for the null effect in India.

5. *Ineffective mechanisms:* Finally, it is possible that the mechanisms outlined in the original democratic peace experiment by (Tomz and Weeks, 2013) do not generalize to India. Perhaps due to the ongoing conflict with

Pakistan, a country which is occasionally labeled as a democracy, subjects in India have learned that democracies are not less threatening or costlier to attack, and that it is not normatively 'wrong' to attack a democracy. Our current design does not allow us to evaluate this explanation, but future research may wish to survey respondents in India about their beliefs about democracies with respect to threats, morality or cost of war.

L Ethics Statement

This study conformed to principles for human subjects research published by the American Political Science Association. We did not collect any identifying information, and subjects remained completely anonymous to us. The survey procedures employed in this study were reviewed by the relevant Institutional Review Board (IRB) and determined to be exempt under category CFR 46.101(b)(2).

We informed subjects that they were taking part in a research study, that their participation was voluntary, and that they could exit the survey at any time. To ensure subjects were able to give informed consent (and understood all aspects of the survey), all survey materials were translated into the primary languages (Brazilian Portuguese, German, Hindi, Hebrew and Japanese) in respondents' countries by native translators, and were further evaluated by academics with relevant language proficiency. We provided the informed consent form in respondents' native language at the beginning of the survey to ensure each respondent understood what they were agreeing to and their rights regarding the storage and use of their data. We also confirmed that each respondent was above the age of 18 before continuing with the survey. After reading the consent information, subjects decided whether to proceed with the survey. Given that the research was exempt with minimal risk of harm, we were not required to obtain signed consent from individuals who opted to take the survey.

Our research procedures did not involve deception. We informed subjects that the situations we posed were hypothetical. To reinforce this idea, we measured our dependent variables using hypothetical language.

This study did not intervene in political processes as described in Principle 10 of the APSA Principles and Guidance for Human Subjects Research.

The survey was administered via Qualtrics, with subjects recruited by Cint. Cint (often known as Lucid in the US) is a professional survey firm that recruits respondents on the Internet for surveys about politics, public affairs, products, brands, and other topics of general interest. Cint compensated subjects according to their prorprietary system. Cint contracts with suppliers who handle incentives to participants directly. Researchers pay Cint a cost per completed interview (CPI) and Cint pays suppliers who then provide a portion of those earnings to participants in the form of cash, gift cards, or loyalty reward points.

Our participant pool was diverse: Cint recruited a diverse sample of adults in each country that was constructed to resemble the local adult population with respect to gender and age. Our research did not intentionally target vulnerable or marginalized groups; any inclusion of such individuals was incidental. Our research procedures did not differentially benefit or harm particular groups.

	Democratic Peace				Audience Costs				International Law				Reciprocity (FDI)			
Samples	Estimate (DP)	SE (DP)	P value (DP)	N (DP)	Estimate (AC)	SE (AC)	P value (AC)	N (AC)	Estimate (IL)	SE (IL)	P value (IL)	N (IL)	Estimate (FDI)	SE (FDI)	P value (FDI)	N (FDI)
Brazil	-0.15	0.04	0.00	3060	-0.45	0.04	0	2004	-0.18	0.03	0.00	3053	0.37	0.03	0.00	3059
Germany	-0.17	0.03	0.00	3000	-0.48	0.04	0	1951	-0.05	0.03	0.08	3005	0.22	0.03	0.00	3013
India	-0.01	0.03	0.82	3075	-0.15	0.05	0	2019	-0.07	0.03	0.05	3070	0.22	0.04	0.00	3072
Israel	-0.23	0.03	0.00	3072	-0.39	0.04	0	2089	-0.10	0.03	0.00	3080	0.41	0.03	0.00	3068
Japan	-0.07	0.03	0.02	3056	-0.34	0.04	0	2029	-0.08	0.03	0.01	3063	0.06	0.03	0.03	3064
Nigeria	-0.06	0.04	0.11	3130	-0.27	0.05	0	2079	-0.31	0.03	0.00	3137	0.74	0.04	0.00	3137
USA	-0.17	0.03	0.00	3019	-0.33	0.04	0	2012	-0.16	0.04	0.00	3023	0.32	0.04	0.00	3019
All Countries	-0.13	0.03	0.00	21412	-0.34	0.04	0	14183	-0.13	0.03	0.00	21431	0.34	0.08	0.00	21432
Original (USA)	-0.28	0.06	0.00	1271	-0.65	0.09	0	451	-0.11	0.04	0.00	2792	0.12	0.01	0.00	2763

Table A1: Meta analysis (Figure 3) in table form.

	Democratic Peace		Audience Costs		International Law		Reciprocity (FDI)	
Samples	Threshold (DP)	P Value (DP)	Threshold (AC)	P Value (AC)	Threshold (IL)	P Value (IL)	Threshold (FDI)	P Value (FDI)
Brazil	4	0.00	2	0	2	0.00	4	0.00
Germany	3	0.00	1	0	7	0.05	5	0.00
India	7	0.41	7	0	6	0.05	6	0.00
Israel	1	0.00	3	0	4	0.00	2	0.00
Japan	5	0.02	4	0	5	0.01	7	0.01
Nigeria	6	0.09	6	0	1	0.00	3	0.00
USA	2	0.00	5	0	3	0.00	1	0.00

Table A2: Sign Generalization (Figure 4) in table form.

Study	Deviation	Reasoning
Democratic Peace	Holding constant additional features of the vignette: In the original study, Tomz and Weeks (2013) randomized additional features of the vignette such as whether the country developing nuclear weapons is an ally of the U.S. We held these additional features constant, where the other country was described as a non-ally of the respondents' country (did not sign a military alliance and does not have high levels of trade with the country).	We kept these features constant to increase statistical power and simplify the experiment
	Additional outcome: we replicate the main outcome analyzed by Tomz and Weeks (2013), measuring support for attacking the country's nuclear sites. We include an additional outcome asking respondents whether they support joining a joint international mission.	We added the additional outcome to examine whether there are floor effects in the original DV, since one concern is that respondents from countries with a weak military will always oppose attacking the facilities
Audience Costs	Title of leader: In the original study by Kertzer and Brutger (2016), the title of the leader is 'President', we changed this word to the title of the leader in each country (e.g. 'Prime Minister' in Israel and 'Chancellor' in Germany)	Ensure compatibility across countries.
	Unspecified leader's party: In the original study, Kertzer and Brutger randomized the party of the President (Republican/Democrat). We did not specify what party the leader is from.	We did not specify the leader's party to simplify the vignette and ensure compatibility.
International Law	Holding constant the nature of the conflict: In the original study, Wallace (2013) varied the nature of the conflict, randomizing information on whether combatants against which torture is used are/are not from regular armed forces. We fix this information at non-regular forces	We fix the nature of the conflict to increase statistical power.
	Removed additional information on reciprocity: In the original study, Wallace further randomized information on whether the opposing side uses torture on the U.S. We removed this information from the vignette.	We remove information on reciprocity to simplify the vignette.
Reciprocity FDI	Minimizing treatment categories: In the original study, Chilton, Milner and Tingley (2020) employ multiple treatment conditions, varying both past (low, medium, high) and present (low, medium, high) score for the ability of U.S. companies to buy companies in the other country. We simplified this into two categories, where the other country either made it easier or harder for companies from the respondents' country to buy companies.	Simplify the scenario and increase power by removing additional treatment conditions.

Table A3: Deviations from Original studies

Statistic	N	Mean	St. Dev.	Min	Max
DP outcome (rescaled)	21,281	0.001	1.000	−1.571	1.276
DP outcome 2 (rescaled)	21,275	0.001	1.000	−2.239	1.017
AC outcome (rescaled)	21,303	0.001	1.001	−1.783	1.340
IL outcome (rescaled)	21,293	−0.0004	1.000	−1.200	1.604
FDI outcome (rescaled)	21,433	0.0001	1.000	−1.507	1.510
Manipulation DP	21,266	0.456	0.498	0	1
Manipulation AC	21,290	0.290	0.454	0	1
Manipulation IL	21,282	0.551	0.497	0	1
Manipulation FDI	21,415	0.465	0.499	0	1
Democratic norms	21,433	3.180	0.630	1.000	5.000
Hawkishness	21,433	2.943	0.988	1.000	5.000
Intl legal obligation	21,433	3.948	0.782	1.000	5.000
Gender	21,433	0.501	0.500	0	1
Education	21,433	4.640	1.469	1	11
Eligable to vote	21,433	0.983	0.131	0	1
Age	21,433	41.151	15.160	18	74

Table A4: Descriptive Statistics - All Countries

	Country is Democracy							Leader back down							Torture Violates Law							Investment Made Harder						
	USA	BRZ	GRM	IND	ISL	JPN	NGR	USA	BRZ	GRM	IND	ISL	JPN	NGR	USA	BRZ	GRM	IND	ISL	JPN	NGR	USA	BRZ	GRM	IND	ISL	JPN	NGR
Democracy	0.49^*	0.45^*	0.54^*	0.29^*	0.53^*	0.42^*	0.58^*																					
	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)																					
Engage								0.48^*	0.55^*	0.53^*	0.26^*	0.57^*	0.40^*	0.56^*														
								(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)														
IL Law															0.43^*	0.41^*	0.43^*	0.21^*	0.50^*	0.41^*	0.54^*							
															(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)							
Harder Invest																						0.49^*	0.43^*	0.49^*	0.26^*	0.54^*	0.39^*	0.56^*
																						(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)
Adj. R^2	0.24	0.21	0.30	0.09	0.30	0.20	0.34	0.25	0.33	0.29	0.09	0.34	0.17	0.35	0.19	0.17	0.19	0.05	0.25	0.17	0.29	0.24	0.18	0.24	0.08	0.30	0.15	0.31
Num. obs.	3014	3052	2997	3065	3071	3056	3126	2011	2000	1950	2015	2089	2030	2073	3016	3045	3003	3070	3077	3060	3132	3015	3055	3011	3071	3068	3062	3133

$^*p < 0.05$

Table A5: Manipulation Test: Treatment Effects on Correct Recall

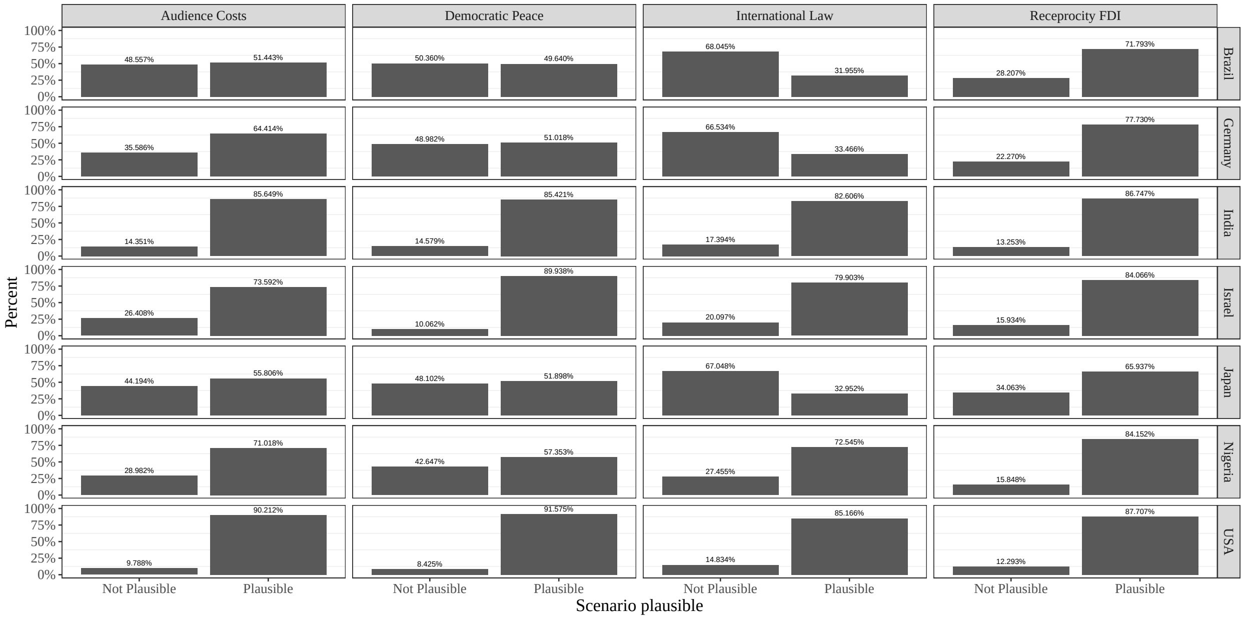

Figure A1: **Plausibility of Experimental Scenarios Across Countries.** This plot reports respondents' evaluation of how plausible an experimental vignette is, by experiment, per country.

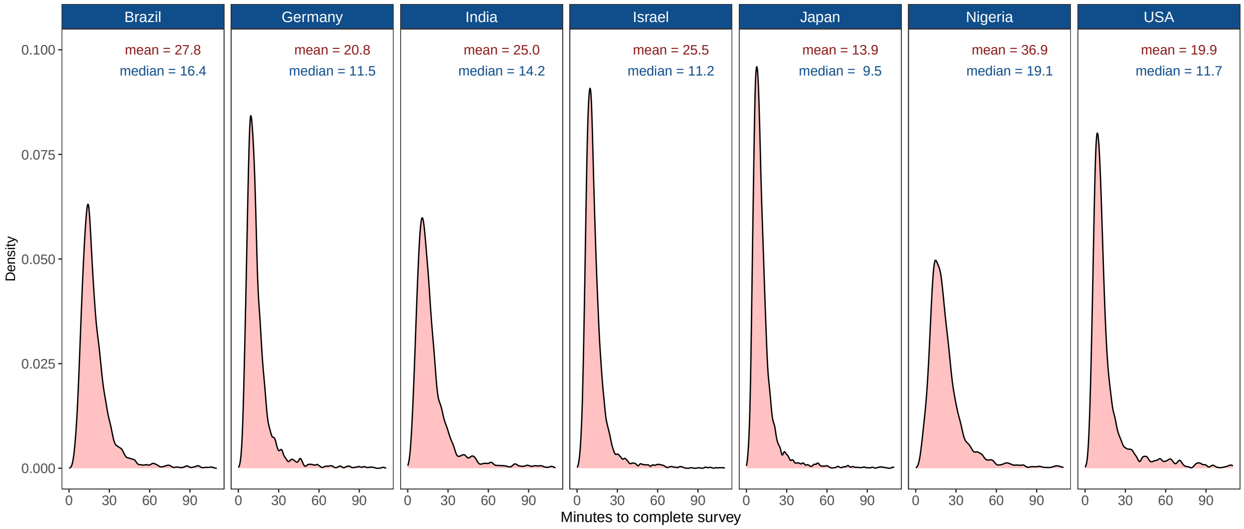

Figure A2: **Response Latency.** Figure reports the density of duration (in minutes) it took respondents to complete the survey. Averages and medians (in minutes) for each country are reported in red and blue, respectively, at the top of each figure.

	Info leak DP							Info leak AC							Info leak IL							Info leak FDI						
	USA	BRZ	GRM	IND	ISL	JPN	NGR	USA	BRZ	GRM	IND	ISL	JPN	NGR	USA	BRZ	GRM	IND	ISL	JPN	NGR	USA	BRZ	GRM	IND	ISL	JPN	NGR
Democracy	−0.11*	−0.03	−0.14*	−0.04*	−0.18*	−0.17*	−0.04*																					
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)																					
Engage								−0.03	−0.04*	−0.09*	0.04	−0.12*	−0.08*	−0.01														
								(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)														
IL Law															0.01	−0.01	0.00	0.02	−0.02	0.03*	0.02							
															(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)							
Harder barrier																						0.04*	−0.02	0.01	−0.01	−0.04*	−0.01	−0.02
																						(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)
Adj. R^2	0.01	0.00	0.02	0.00	0.03	0.03	0.00	0.00	0.00	0.01	0.00	0.01	0.01	−0.00	−0.00	−0.00	−0.00	0.00	0.00	0.00	−0.00	0.00	0.00	−0.00	−0.00	0.00	−0.00	−0.00
Num. obs.	3019	3055	2998	3068	3072	3058	3130	2014	2004	1953	2018	2090	2031	2077	3023	3051	3007	3071	3081	3063	3139	3018	3058	3014	3073	3070	3063	3137

$^* p < 0.05$

Table A6: Information Leakage: ATE on thinking of a specific country

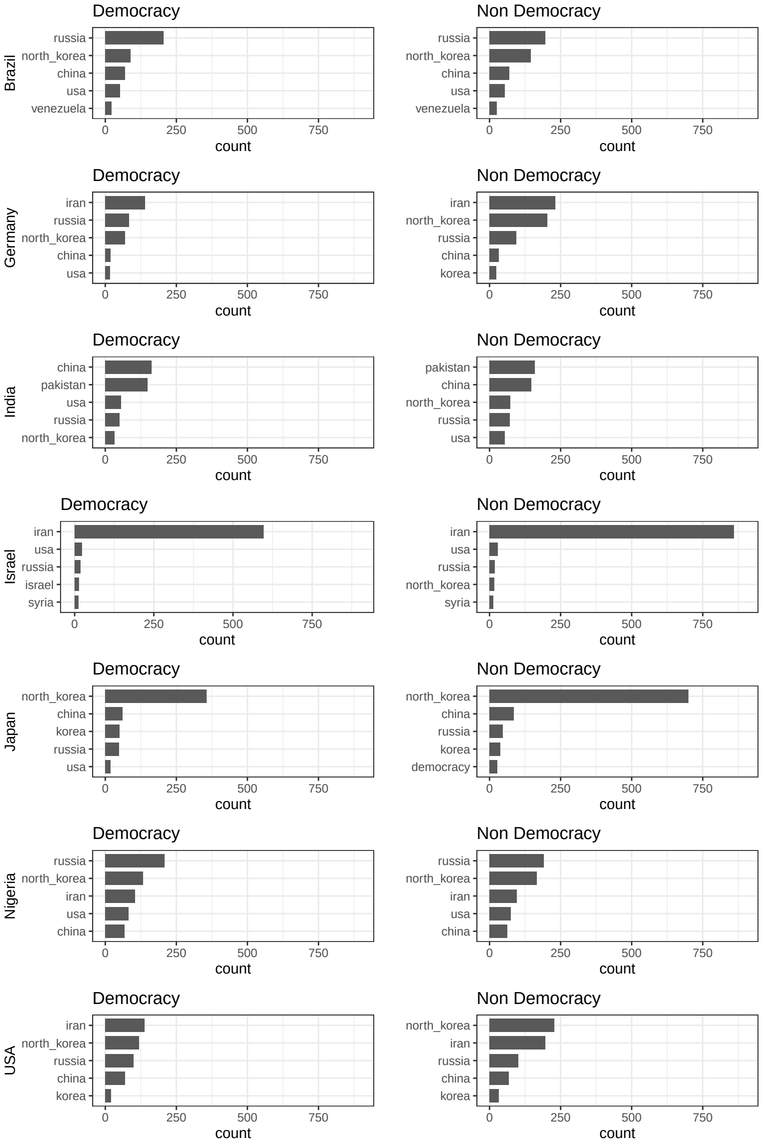

Figure A3: **Top 5 countries respondents thought of in the Democratic Peace study.** We report the top 5 most used words in an open-ended question asking respondents whether they thought of a specific country when reading the democratic peace scenario. X-axis orders the words from most to least mentioned. Plot is faceted by country and by condition.

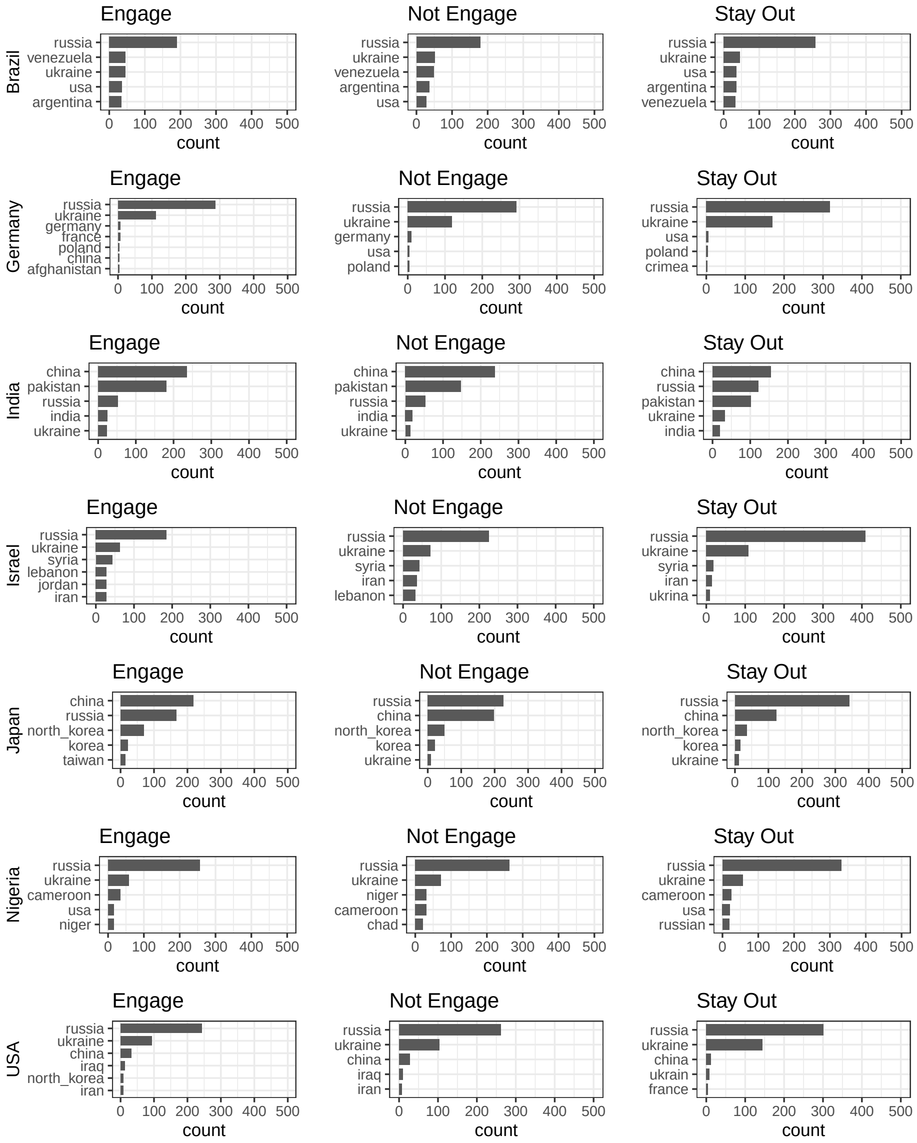

Figure A4: **Top 5 countries respondents thought of in the Audience Costs study.** We report the top 5 most used words in an open-ended question asking respondents whether they thought of a specific country when reading the audience costs scenario. X-axis orders the words from most to least mentioned. Plot is faceted by country and by condition.

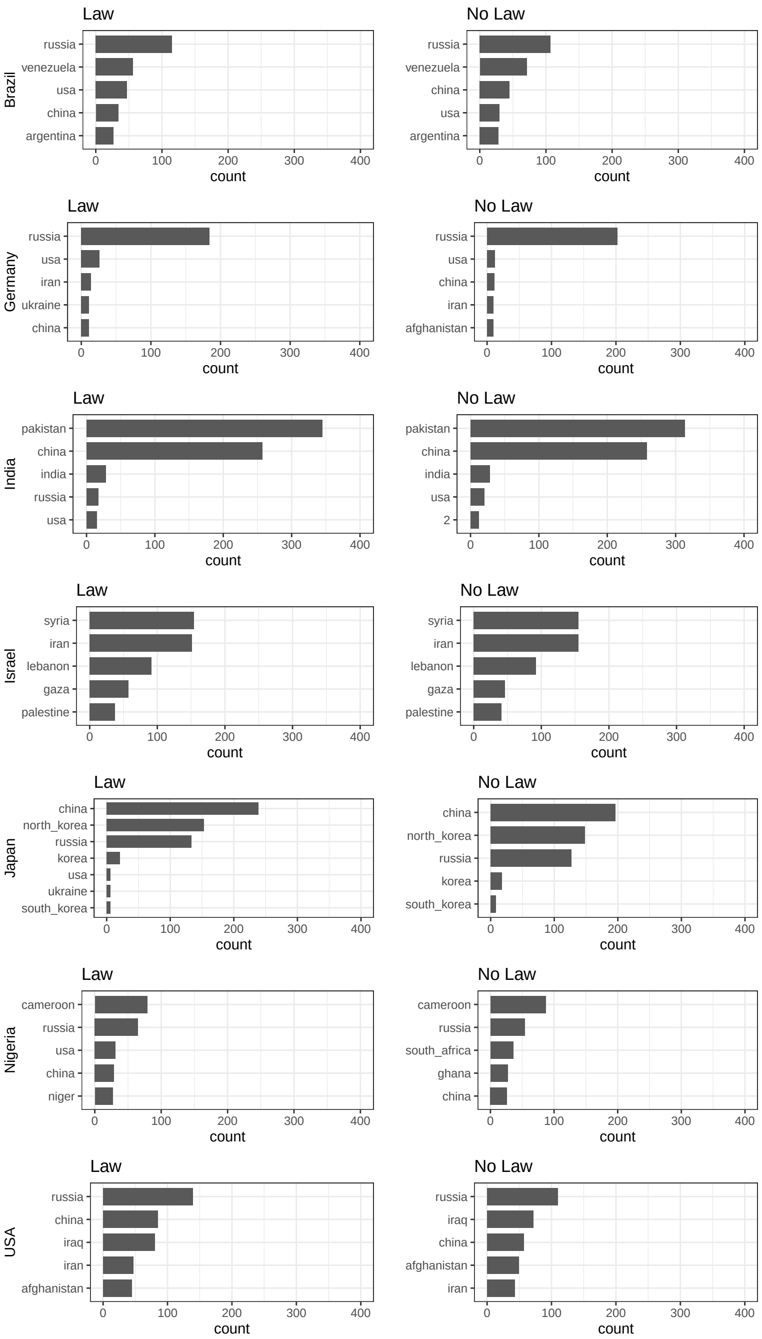

Figure A5: **Top 5 countries respondents thought of in the International Law study.** We report the top 5 most used words in an open-ended question asking respondents whether they thought of a specific country when reading the international law scenario. X-axis orders the words from most to least mentioned. Plot is faceted by country and by condition.

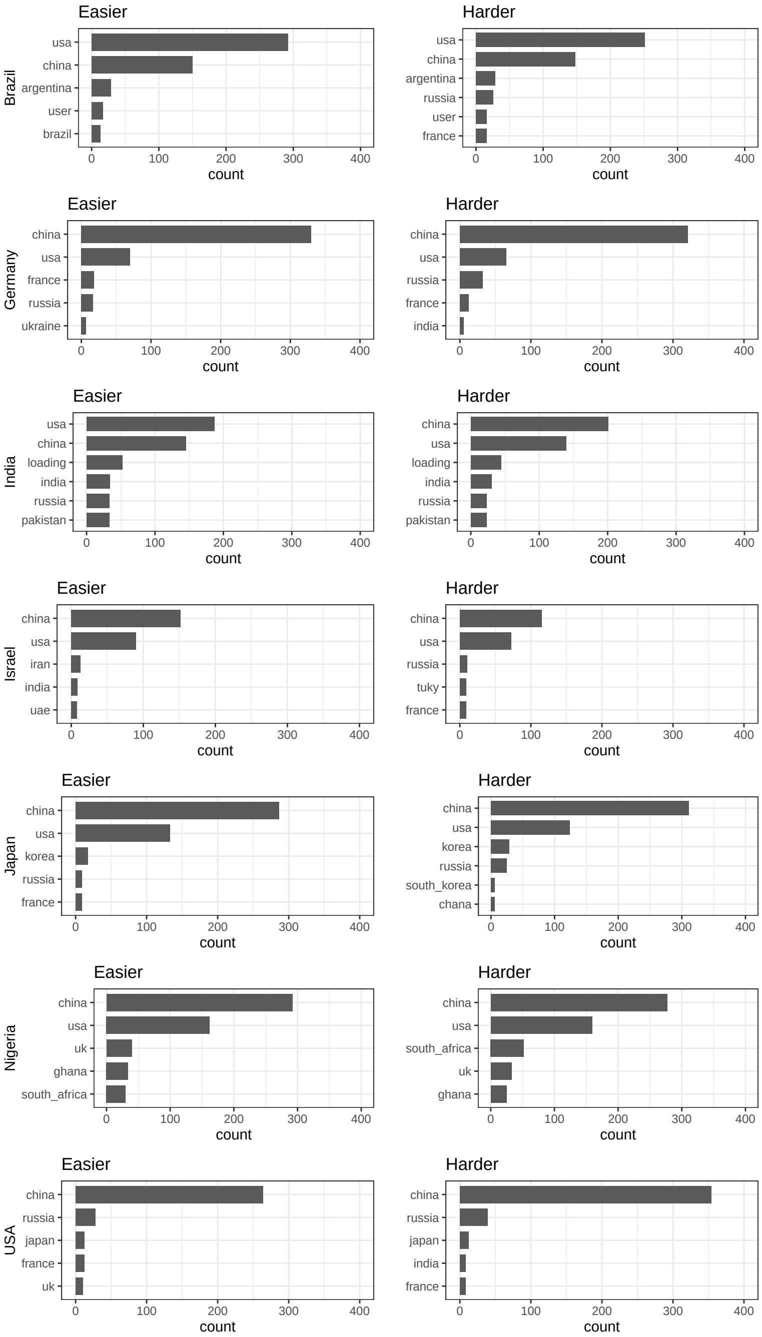

Figure A6: **Top 5 countries respondents thought of in the Reciprocity FDI study.** We report the top 5 most used words in an open ended question asking respondents whether they thought of a specific country when reading the reciprocity FDI scenario. X-axis orders the words from most- to least- mentioned. Plot is faceted by country and by condition.

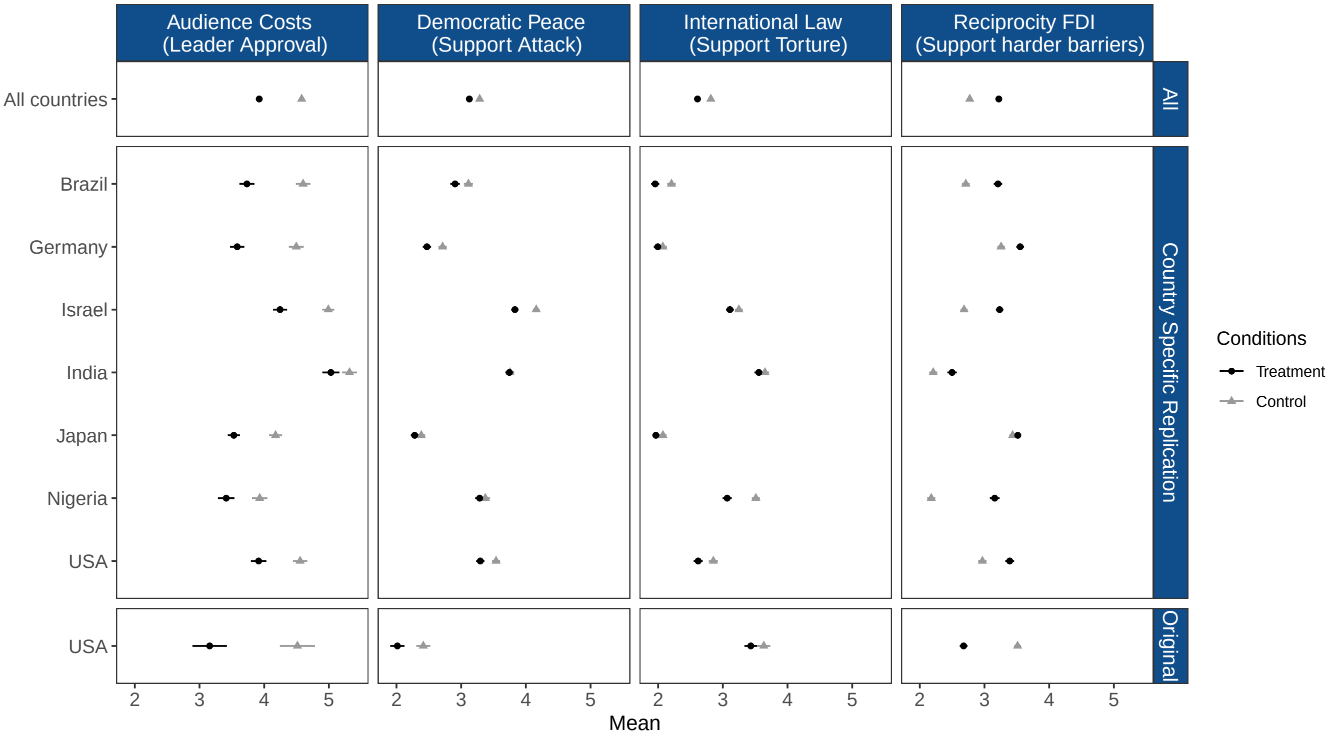

Figure A7: **Means by condition.** For each experiment, we report original means and confidence intervals of main treatment and control conditions from published studies, alongside means and confidence intervals from our country-specific replication and from the full multi-national sample (”Meta”). To ease interpretation, we use the original outcomes employed in the studies. The Democratic Peace, International Law, and Reciprocity FDI experiments employed outcomes ranging from 1-5, whereas the Audience Costs experiment employed an outcome ranging from 1-7.

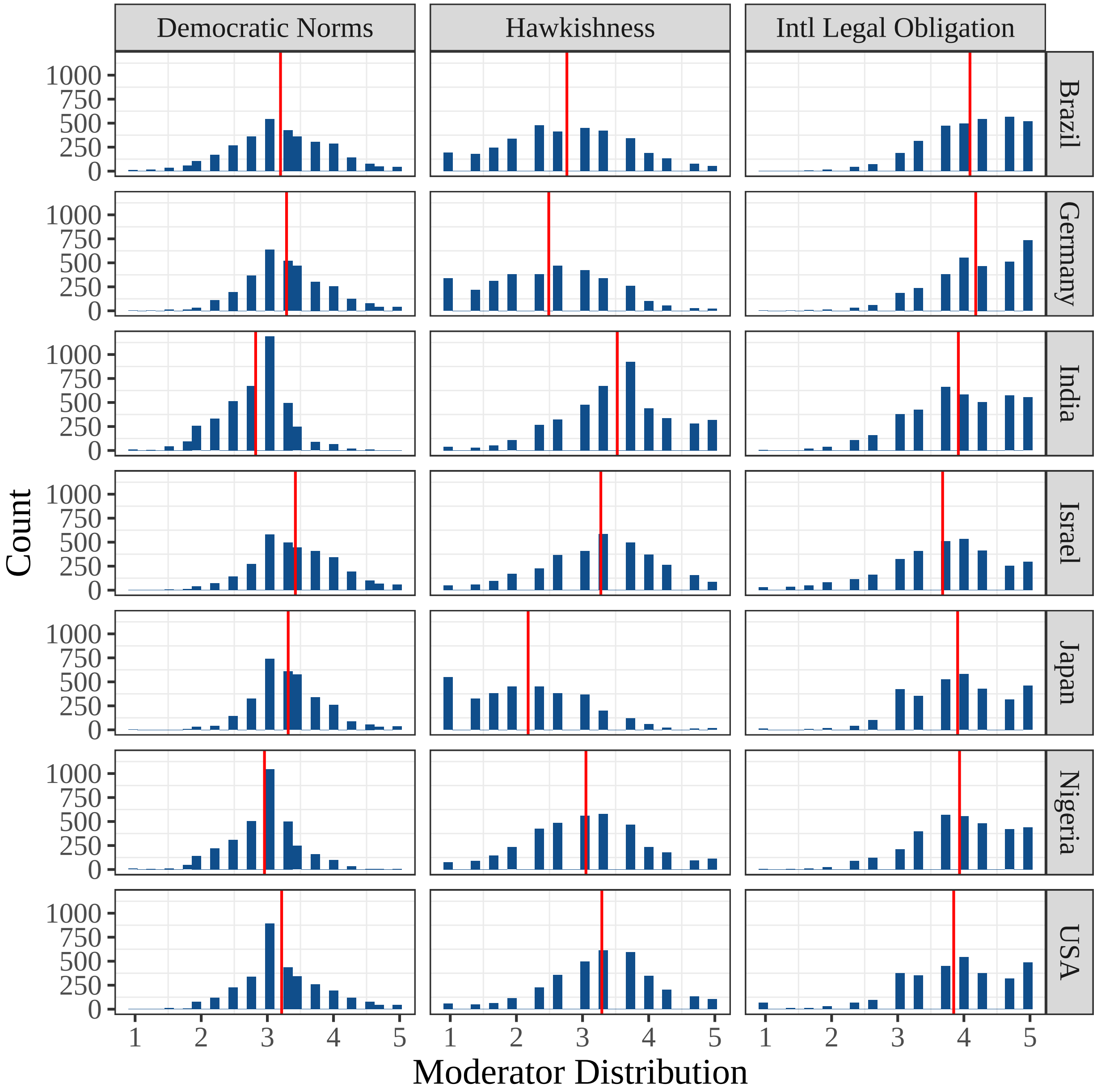

Figure A8: **Distribution of Moderators Across Countries.**

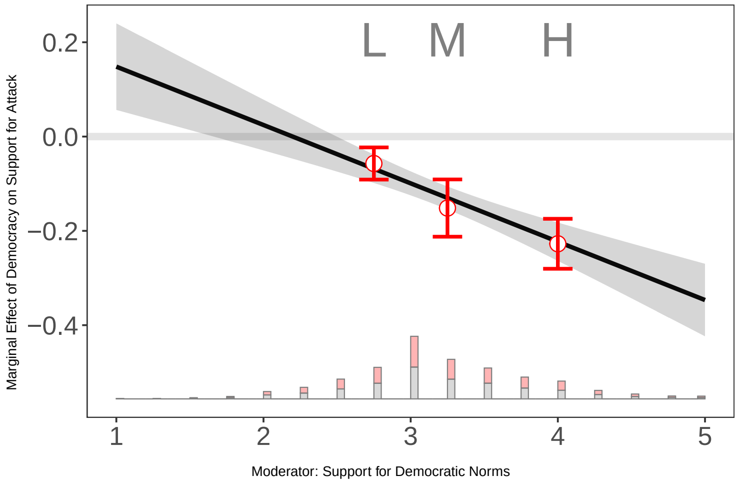

Figure A9: **Moderating Effect of Support for Democratic Norms Index in the Democratic Peace Experiment.** This figure demonstrates the negative moderation of support for democratic norms on the democracy treatment effects. That is, the effect of describing a country as a democracy reduces support for attacking the said country, and the effects are smaller (larger) for respondents with low (high) levels of support for democracy.

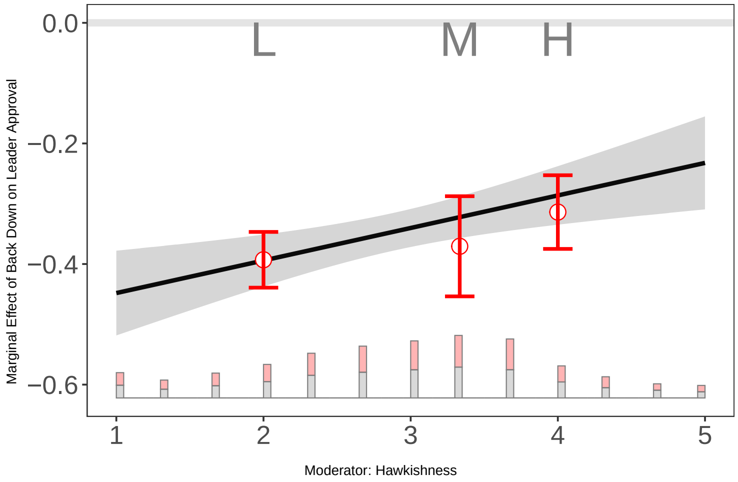

Figure A10: **Moderating Effect of Hawkishness Index in the Audience Costs Experiment.** This figure demonstrates there is no strong evidence for a moderation of hawkishness on the audience costs experiment.

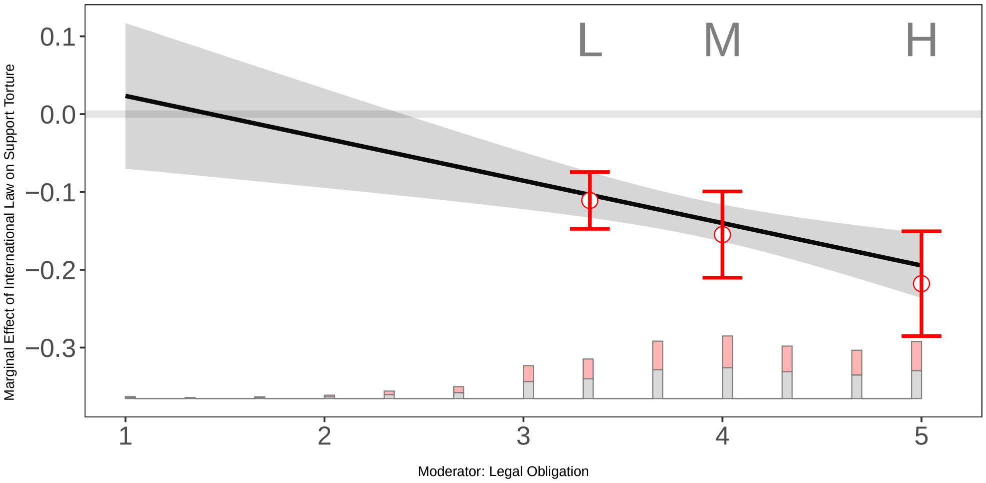

Figure A11: **Moderating Effect of International Legal Obligation Index in the International Law Experiment.** This figure demonstrates the negative moderation of legal obligation on the international law treatment effects. That is, mentioning that the respondent's country signed international law treaties prohibiting the use of torture reduces support for the use of torture, and the effects are smaller (larger) for respondents with low (high) levels of international legal obligation.

	Democratic Peace	Audience Costs	International Law
	Model 1	Model 2	Model 3
Democracy	-0.121^{*}		
	(0.012)		
Dem Norms	-0.142^{*}		
	(0.016)		
Dem*Norms	-0.095^{*}		
	(0.022)		
Back Down		-0.343^{*}	
		(0.016)	
Hawkish		0.041^{*}	
		(0.014)	
BD*Hawk		0.029	
		(0.020)	
Intl Law			-0.137^{*}
			(0.012)
Legal Obligation			-0.026^{*}
			(0.012)
IL*Oblig			-0.053^{*}
			(0.017)
Adj. R^2	0.181	0.098	0.221
Num. obs.	21426	14197	21445

$^{*}p < 0.05$. Regressions interact treatment with covariates (gender, age, education, voting status, country).

Table A7: Moderation Tests

	Hawkishness	Legal Oblig	Demo Norms	Age	Ideology	University Educated
	(1)	(2)	(3)	(4)	(5)	(6)
Brazil	-0.528^{***}	0.246^{***}	-0.017	-5.743^{***}	-0.629^{***}	0.097^{***}
	(0.021)	(0.019)	(0.015)	(0.308)	(0.062)	(0.012)
Germany	-0.803^{***}	0.333^{***}	0.074^{***}	0.976^{***}	0.193^{***}	-0.123^{***}
	(0.021)	(0.019)	(0.015)	(0.317)	(0.062)	(0.012)
India	0.234^{***}	0.071^{***}	-0.393^{***}	-9.053^{***}	0.077	0.439^{***}
	(0.020)	(0.018)	(0.014)	(0.281)	(0.062)	(0.012)
Israel	-0.015	-0.167^{***}	0.208^{***}	-3.918^{***}	0.384^{***}	0.076^{***}
	(0.021)	(0.019)	(0.015)	(0.311)	(0.062)	(0.012)
Japan	-1.114^{***}	0.060^{***}	0.099^{***}	2.263^{***}	0.507^{***}	0.108^{***}
	(0.021)	(0.019)	(0.015)	(0.304)	(0.062)	(0.012)
Nigeria	-0.240^{***}	0.088^{***}	-0.259^{***}	-12.007^{***}	0.262^{***}	0.306^{***}
	(0.021)	(0.019)	(0.015)	(0.297)	(0.062)	(0.012)
N	24,781	23,442	23,581	33,428	22,097	22,082

Notes:

Table A8: Estimating Differences Between Country Samples. Each model regresses relevant outcomes over country indicators compared to the US (which serves as a reference category).

	Joint International Mission						
	BRZ	GRM	IND	ISL	JPN	NGR	USA
Democracy	-0.15^{*}	-0.27^{*}	-0.06	-0.27^{*}	-0.11^{*}	-0.00	-0.20^{*}
	(0.05)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Adj. R^2	0.00	0.01	0.00	0.02	0.00	-0.00	0.01
Num. obs.	3057	3001	3070	3072	3058	3131	3020

$^{*}p < 0.05$

Table A9: Alternative Outcome (DP): Treatment Effect on Joining Intl Mission

	Support attack
	Model 1
Democracy	-0.012
	(0.032)
Dem Norms	-0.012
	(0.049)
Democracy*Norms	-0.124
	(0.066)
Adj. R^2	0.017
Num. obs.	3077

$^{*}p < 0.05$. Regressions interact treatment with covariates (gender, age, education, voting status).

Table A10: Moderation Tests (India DP)

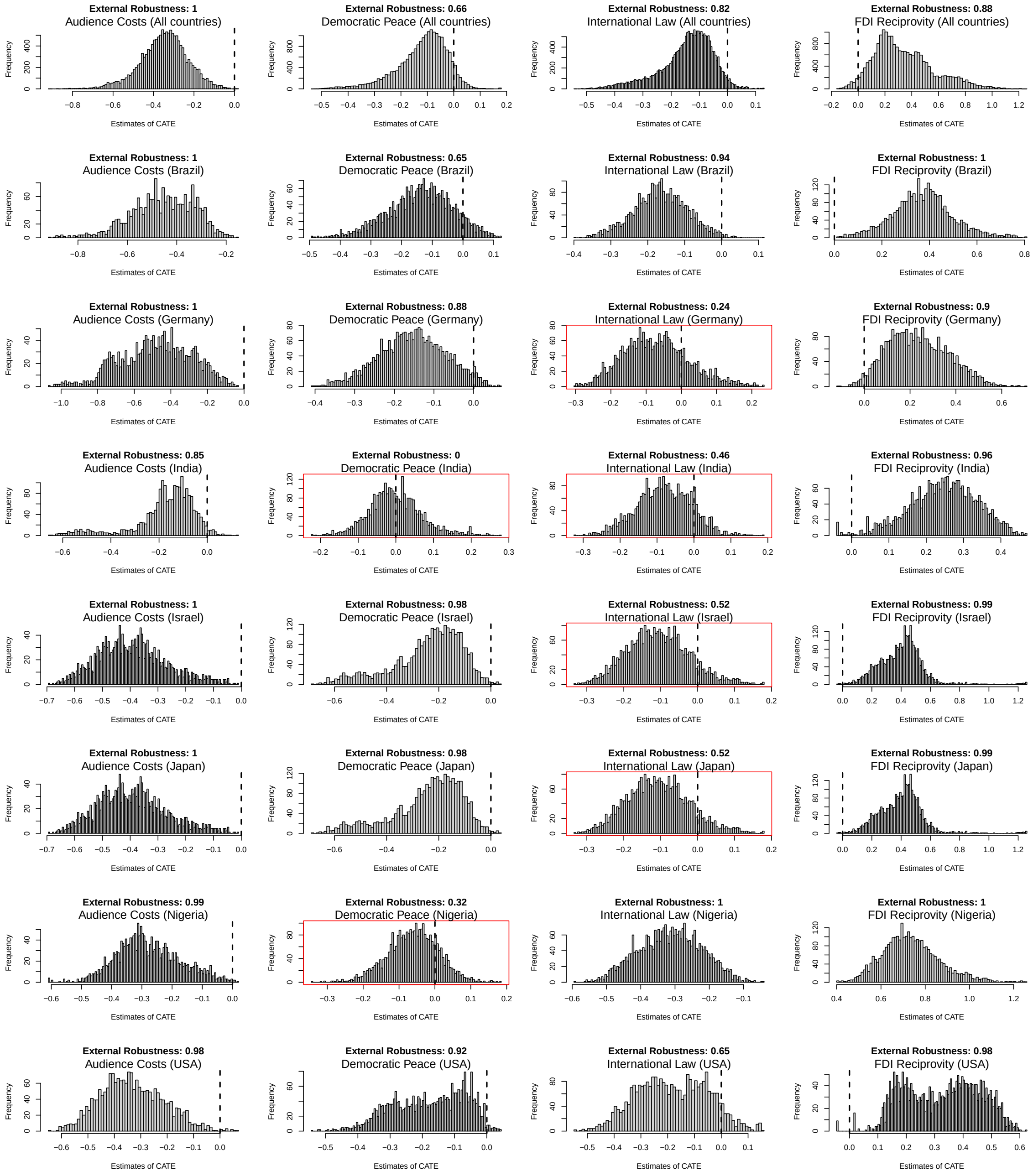

Figure A12: **External Validity Bias Test.**

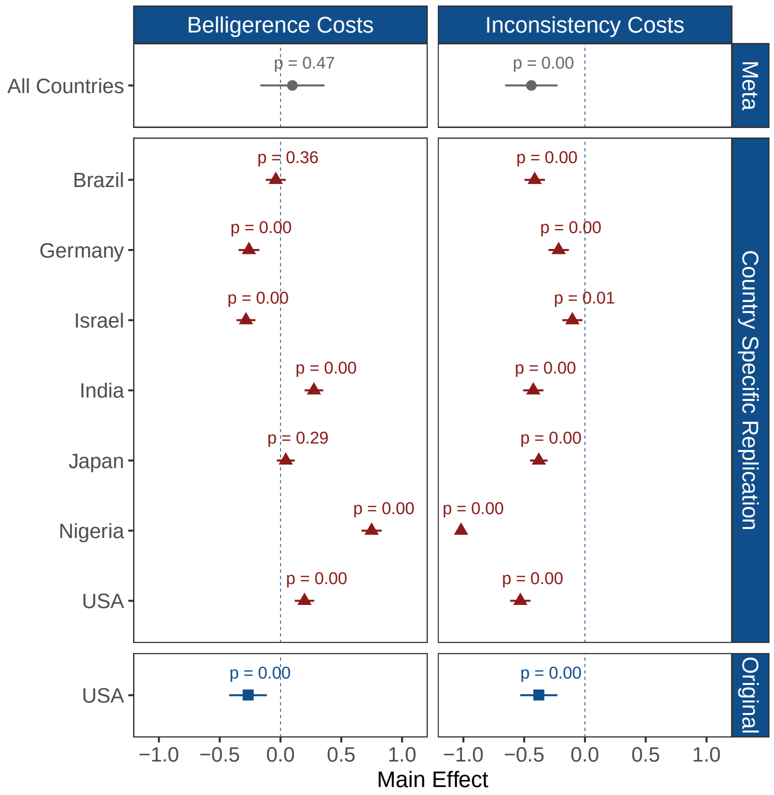

Figure A13: **Audience Costs extension.** We report the original estimates and p-values of the belligerence and inconsistency costs, as calculated in the original study. We further report the country-specific ATEs (and BH-adjusted p-values) from our replications, and a meta-analytic average treatment effect based on our harmonized studies.

	Support attack
	Model 1
Democracy	−0.026
	(0.066)
Adj. R^2	−0.001
Num. obs.	824

${}^{*}p < 0.05$.

Table A11: Screening out failed manipulation from other studies (India DP)

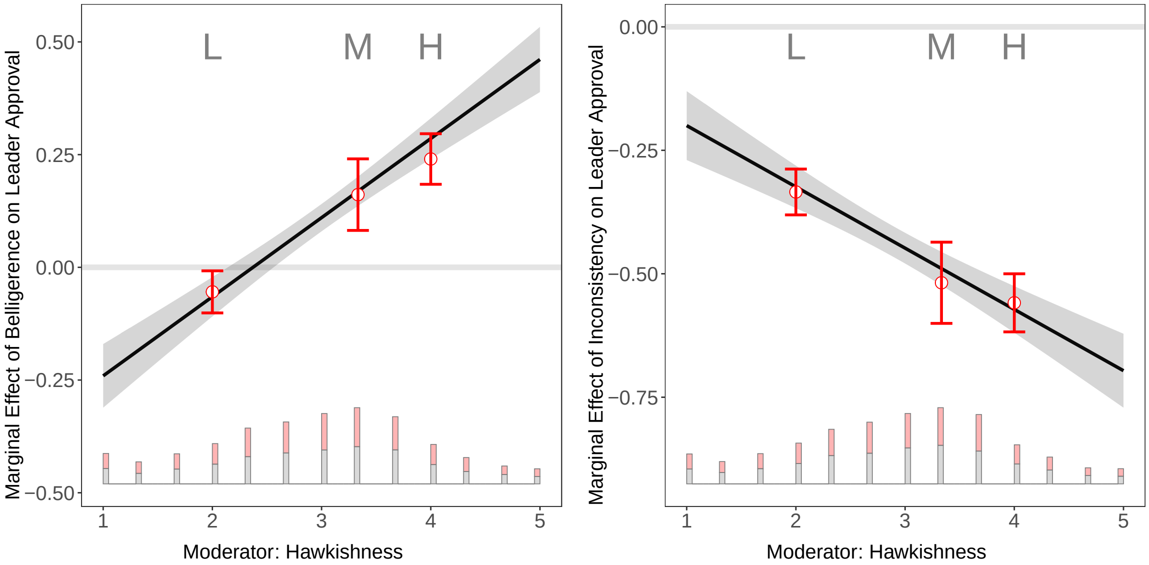

Figure A14: **Moderating Effect of Hawkishness Index in the extension of the Audience Costs Experiment.** This figure demonstrates the negative moderation of hawkishness on the inconsistency treatment effects, and the positive moderation of hawkishness on belligerence treatment effect. These results are consistent with findings from the original study.

CONFIDENTIAL - FOR PEER-REVIEW ONLY

Generalizability of IR Experiments (#119402)

Created: 01/20/2023 10:27 AM (PT)

1) Have any data been collected for this study already?
It's complicated. We have already collected some data but explain in Question 8 why readers may consider this a valid pre-registration nevertheless.

2) What's the main question being asked or hypothesis being tested in this study?
We replicate four prominent International Relations vignette experiments in seven countries: USA, Germany, Brazil, Japan, Nigeria, India, Israel. The four experiments test the following hypotheses:

Democratic Peace: respondents are less likely to support attacking another country if that country is described as a democracy, compared to a condition in which the country is described as an autocracy (Tomz and Weeks 2013).

Audience Costs: respondents will evaluate a leader less favorably if said leader does not follow through on their threat towards an aggressor, compared to a condition in which the leader stays out of conflict in the first place (Tomz 2007; Brutger and Kertzer 2016).

International Law: respondents are less likely to support the use of torture when informed that using torture violates international treaties signed by their country, compared to a condition in which international treaties are not mentioned (Wallace 2013).

Reciprocity: respondents are more likely to support increasing barriers to foreign investment on another country if said country increased barriers to investment, compared to a condition in which a country lowered barriers (Chilton et al 2020).

3) Describe the key dependent variable(s) specifying how they will be measured.
Each of our four experiments has its own dependent variables drawn from the original study except where mentioned:

Democratic Peace: support for attacking other country (approval, scaled from 1-5); secondary outcome (not from original study): support for joining a mission attacking other country (approval, scaled from 1-5)

Audience Costs: leader approval (approval, scaled from 1-7)

International Law: support for employing torture (scaled from 1-5)

Reciprocity: support for reducing/increasing investment barriers on other country (scaled from 1-5)

4) How many and which conditions will participants be assigned to?
Each respondent completes all four studies, but we randomize the order of the studies. Within each experiment, respondents are assigned to the following conditions (drawn from original studies):

Democratic Peace: country is described as either: a) democracy b) non-democracy.

Audience Costs: leader is described as either: a) staying out of the dispute, b) engaging in dispute but not following through on threat, c) engaging in dispute and following through on threat. Only conditions (a) and (b) are used for main analysis (see Section 2 above), consistent with Tomz 2007.

International Law: either: a) torture is described as a violation of international law, b) international law is not mentioned.

Reciprocity: other country is described as making it either: a) easier b) harder for the respondent's country to purchase a company in the other country.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
Our main questions will be examined in 3 (related) parts. First, for each experiment, a "country-specific" ATE will be calculated (for each country-outcome combination) using OLS regressions (with robust standard errors) where each study's outcome is regressed over the study's main randomized treatment contrast. We report adjusted p-values using the Benjamini-Hochberg correction accounting for seven tests (1 test for each country) of each hypothesis. We reject the null hypothesis for a given test if the adjusted p-value < 0.05.

Second, for each experiment, those country-specific ATEs will be aggregated into a "meta-analytic" ATE using a meta-analytic random effects model (Borenstein et al. 2021), implemented using the "rma" command in the "metafor" package in R. We report unadjusted p-values for the meta-analyses. Third, and to complement our analysis of meta-analytic average treatment effects, we will employ a "sign-generalization" test designed by Egami & Hartman (2022).

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
Our study has four pre-treatment attention checks. Subjects who fail any one of four pre-treatment attention checks will not be allowed to continue in the survey and thus be excluded from the analysis.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
Based on a power analysis using effect sizes and outcome SDs from original studies, we aim to collect 3,000 complete, attentive subjects per country, resulting in a sample size of 21,000 subjects across 7 countries: USA, Japan, India, Nigeria, Israel, Brazil, and Germany. In case of excess respondents, we will use all data delivered by the survey company.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
Information on question 1: Pilot data (N=416) were collected in Nigeria in August 2022 but will not be used in our main analyses.

We plan to implement several additional sets of analyses, outlined below.

1) Diagnostics, including:
a) Evaluate treatment take-up: For each experiment within each country, we regress a response to factual manipulation check over treatment condition.
b) Evaluate vignette plausibility by probing variation in plausibility (by study) across countries. This will be accomplished by plotting distribution of post-treatment questions asking about plausibility of scenario for each experiment in each country.
b) Evaluate whether respondents have in mind a particular country for each scenario. This will be accomplished by plotting distribution of answers to the question "did you have a specific country in mind while reading this vignette?" for each country and each experiment by treatment condition.

2) Heterogeneous Treatment Effects: For each experiment, we consider a key moderator by focusing on interacting our treatment with a moderator as well as with pre-treatment controls (gender, age, education, voting eligibility, country) in our pooled sample, as follows:
a) Democratic Peace: respondents' support for democratic norms (based on Kingzette 2021).
b) Audience Costs: respondents' hawkishness (based on Brutger and Kertzer 2016).
c) International Law: respondents' legal obligation (based on Bayram 2017).
3) External Validity Bias: we evaluate issues related to demographics and external validity in using a procedure proposed by Egami and Devaux 2022 for estimating external validity bias for each experiment by country. We implement the procedure proposed by Egami and Devaux 2022 for all experiments across all countries. For each experiment in each country, this approach employs all pre-treatment covariates to estimate heterogeneity in average treatment effects (using a generalized random forest approach), and report an external validity score (between 0-1) depending on the amount of reweighting necessary to explain away the average treatment effect.

4) Audience Cost Extension: In our secondary analysis we follow Brutger and Kertzer 2016 and decompose audience cost into a "belligerence" cost and an "inconsistency cost." We plan to plot the decomposed audience cost average treatment effects across countries, using Benjamini-Hochberg adjusted p-values to account for the 14 tests (2 outcomes across 7 countries).