The Generalizability of IR Experiments Beyond the U.S.

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

March 14, 2024

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 especially important to know whether particular findings hold across countries. To examine the generalizability of IR experimental findings (in terms of direction and significance) of effects 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 limited treatment effect heterogeneity. Our study reveals that findings from the U.S. are similar to findings from a wide range of democracies, offering important theoretical and empirical implications to inform the design and interpretation of future experimental research in IR.

*Corresponding Author. Postdoctoral Research Associate, Niehaus Center for Globalization and Governance, Princeton University ✉: lb3167@princeton.edu, 🌐: lotembassanygate.com

†Professor of Political Science, Department of Political Science, UW-Madison, ✉: renshon@wisc.edu, 🌐: jonathanrenshon.com

‡Professor of Political Science, Department of Political Science, UW-Madison, ✉: jweeks@wisc.edu, 🌐: jessicalpweeks.com

§Postdoctoral fellow, Conflict and Polarization Lab, Stanford University ✉: cmweiss@stanford.edu, 🌐: chagaiweiss.com

In recent years, scholars of international relations (IR) have 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 (Chaudoin, 2014; Mutz and Kim, 2017), nationalism (Powers, 2022), and immigration (Hainmueller and Hiscox, 2010), among others. Over time, a cottage industry has emerged to further improve the internal validity of experimental research, shoring up one of the method's key strengths.[1]

A new wave of political science research 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). 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, lags behind these important endeavors. To the extent that scholars have examined the "generalizability" of IR experiments, they have tended to evaluate findings from a single study in one or several additional contexts (Tomz and Weeks, 2013; Renshon, Yarhi-Milo and Kertzer, 2023; Suong, Desposato and Gartzke, 2020), often introducing design changes across countries and providing limited motivation for case selection. Existing multi-site experiments in IR are thus often unable to evaluate the extent to which findings generalize to other countries.

Here, we define generalizability as denoting whether existing evidence—in this case, from a series of

[1] E.g. 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); Chaudoin, Gaines and Livny (2021); Brutger, Kertzer, Renshon, Tingley and Weiss (2022); Brutger, Kertzer, Renshon and Weiss (2022).

prominent IR papers—"apply to other sets of individuals, to other types of interventions, and in other contexts" (Blair and McClendon, 2021, 411). More specifically, we are interested in "*C*-validity" (Egami and Hartman, 2022, 7), which focuses on extensions of findings to contexts ('*C*') in which theories have not yet been tested. We examine the extent to which replications of IR experiments in new contexts produce statistically significant effects in the same direction as original results, sidestepping the issue of the magnitude of effects, which is typically more relevant for theories about particular policy interventions (Egami and Hartman, 2022, 1080-86).

Assessing the generalizability of IR findings is crucial for remedying the mismatch between the scope of IR theories and the breadth of their underlying evidence. Although the broad predictions of "theories of international relations" make it particularly important to evaluate their explanatory power across different country contexts, the vast majority of existing experimental evidence stems 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*). It is thus difficult to judge whether IR theories are truly *international*, or merely explain the foreign policy preferences of Americans. Moreover, assessing experimental results in other countries reveals 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 designed to sidestep challenges such as publication bias (i.e., selective reporting of positive results) and study comparability (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 democracies (the U.S., India, Germany, Japan, Brazil, Nigeria, and Israel), which we selected using a strategy of "purposive variation" (Egami and Hartman, 2022). Our empirical tests address two key questions about generalizability: (1) in how many (and which) countries is the sign of the result consistent with theoretical predictions? (*sign-generalizability*) and (2) is there support for a given theory in the pooled population of respondents across all our countries? (*meta-analysis*). Our sign-generalizability test also allows us to make (speculative) inferences about the direction of effects in countries we did not study,

subject to plausiblity of additional assumptions.

Our study makes three central contributions. First, our results suggest the 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. We provide rigorous evidence bolstering the generalizability of prominent experiments testing the micro-foundations of important IR theories. 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 important and well-known experiments from different substantive domains and our empirical focus on a set of purposively varied countries suggests reasons for optimism in considering the external validity of IR experiments more generally.

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. Americans are indeed different from other populations in many ways, but our main results suggest that such differences may not dramatically shape experimental finding across countries for common IR theories. More broadly, and in line with recent studies in American politics (Coppock, 2022), the theories we tested appear to exhibit low treatment effect heterogeneity (Coppock, 2019): samples with considerable heterogeneity along a number of measured covariates respond similarly to treatment, and we estimate low amounts of *unmeasured* heterogeneity as well.

Third, our study has important implications for future experimental research in both IR and other subfields. On the one hand, our findings suggest that researchers can learn much from single-country studies, whether in the U.S. or elsewhere. This conclusion has important practical and normative implications by reducing barriers to entry for non-U.S. based scholars and for correcting the impression that the U.S. should be the "normal" site for experimental research. However, our findings emphasize the importance of theorizing *ex ante* about variables that could moderate treatment effects and incorporating measures of these moderators at the design stage. Doing so can reveal whether treatment effects are heterogeneous *within* their sample. If treatment effects are homogeneous, scholars can increase their confidence in cross-country generalizability. If treatment effects are heterogeneous, particularly in ways that change the sign (rather than the magnitude) of a hypothesized effect, scholars should consider how samples in other contexts might differ and be cautious about making more general claims.

At the same time, our study demonstrates the value of harmonized multi-site replication studies in circumstances where such efforts are possible. In those cases, others may rely on our approach, which brings together an innovative suite of tools for choosing sites, analyzing experimental data—based on Egami & Hartman's (2022) framework of "purposive variation" and sign-generalization—and designing research around theoretically relevant moderators and investigating possible null results. And while our empirical exercise provides reassuring insights regarding the generalizability in IR experimental research, 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). A given design may or may not be internally valid, a quality specific to a particular study (McDermott, 2011*a*, 28; 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 specific to individual experiments (Renshon, 2015, 667). Rather, insights about external validity emerge as replications "across time and populations" reveal the extent to which conclusions generalize (McDermott, 2011*a*, 28). Although critics have lamented that political science has "fallen down an internal validity rabbit hole" at the expense of other goals, (Findley, Kikuta and Denly, 2021, 366)[2] scholars have begun to develop the concept of external validity theoretically and generated 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).

We follow Renshon (2015), Blair and McClendon (2021, 411) and others by defining external validity, i.e. generalizability, as whether existing findings "apply to other sets of individuals, to other types of inter-

[2]Though, see McDermott (2011*a*, 27).

ventions, and in other contexts." More specifically, 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. We aim to assess $C-$ validity: "Do experimental results generalize from one context to another context?" (Egami and Hartman, 2022, 5). 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). We focus specifically on cross-country variation, which is challenging because geographic variation may involve covariates that do not vary within individual studies.[3]

We consider a particular finding *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 within the bounds of a theory's scope conditions. Our focus on direction and significance (rather than magnitude of effect) is motivated by Egami and Hartman (2022, 1086), who recommend generalizability tests of direction/sign for "synthesizing scientific knowledge," while reserving tests that implicate magnitude for evaluating direct policy implications. Further justification comes from the nature of the theories we test, which do not (implicitly or explicitly) feature predictions about effect sizes. The *scope* of the theory matters by helping to bound our empirical tests: 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 the theory generalizable if we found consistent experimental support for it across an array of democratic countries.

2 Generalizability in IR

IR theories are usually intended to provide broad insights about interstate relations across a wide range of countries. Since the field's inception, IR theories have been concerned with "analysis of the contemporary multi-state system and…the conduct of its *components*" (emphasis added; Wolfers, 1947, 26). Likewise, contemporary IR theories seek 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 both for "grand" frameworks

[3]Note however that although moderators vary at the national level, they can often be measured, as we do, at the individual level.

such as realism and "middle-range" theories such as the democratic peace, and whether the theorized actors are states, leaders, voters, or IOs. For example, theories of reputation (Downs and Jones, 2002; Wolford, 2007) and resolve (Kertzer, 2016) make predictions about states, leaders, and perceptions on a general level, not restricted to any one particular state, leader, or specific empirical context. If a theory applied in only one country, it would be considered a theory of that country's foreign policy rather than a "theory of IR."

Given these goals, it is important to assess whether a given theory is validated by a sufficient base of evidence from multiple contexts—ideally, an accumulation of empirical tests from a broad range of countries. Scholars have observed a long-running trend in which IR research has focused on the U.S. (Hoffmann, 1977; Colgan, 2019*a*; Levin and Trager, 2019; Kristensen, 2015), not only in terms of evidence but also 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."[4]

To assess whether the micro-level experimental literature is similarly U.S.-centric, 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 provides a heat map of these studies by country "site" (location). Strikingly, nearly sixty percent of all the experiments utilized U.S. subjects. Moreover, the U.S. was eight times more popular than the next most common site, Israel.[5] Evidently, experimental research on the micro-foundations of prominent IR theories relies predominantly on studies of U.S. foreign policy attitudes, behaviors, and perceptions.

[4]More broadly, observational work is not free from concerns over generalizability (Aronow and Samii, 2016)

[5]This echoes Hendrix and Vreede (2019, 311), who point out that Israel and the U.S. receive scholarly attention far out of proportion to their population, GDP, etc.

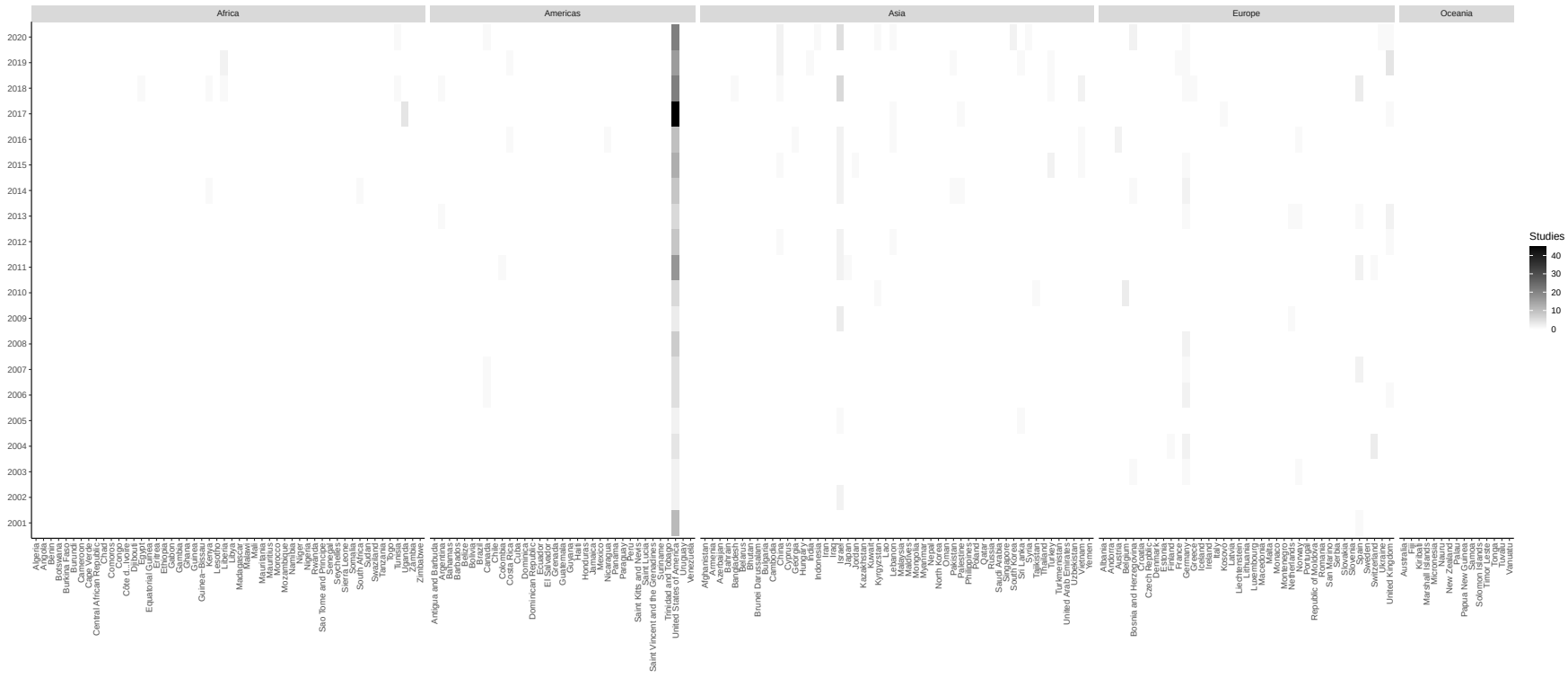

Figure 1: Heatmap of IR experiments published in six top journals between 2001-2020 by country. As is demonstrated by the single vertical line signifying U.S.-based studies, IR survey experiments are predominantly conducted in the U.S.

One could reasonably worry that carrying out micro-level empirical tests of IR theories nearly exclusively on U.S.-based samples provides little insight into broader empirical relationships. The U.S. is wealthier, has longer-standing democratic institutions,[6] is more geographically-protected, more conflict-prone, and more powerful and authoritative than most other countries. To the extent that such country-level factors affect residents' ideologies, perceptions, or judgments, experimental findings from U.S.-based subjects might shed little light on causal theories that apply to populations in other places. U.S.-based subjects might also be unusual at the individual level. For example, Americans tend to be less knowledgeable than peers in other locations (Levin and Trager, 2019, 352; Dimock and Popkin, 1997), and the U.S. stands out demographically even from other large, powerful countries (Brooks et al., 2018). In fact, psychologists have developed a term—Western, educated, industrialized, rich and democratic (WEIRD)—to denote people from societies like the U.S., characterizing them as "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 might be overblown. Coppock, Leeper and Mullinix (2018), for example, use online convenience samples to 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 the convenience-sample replications. They interpret these results as suggesting that many social science experiments exhibit low "treatment effect heterogeneity": i.e., for many studies, subgroup treatment effects do not differ that much. This might suggest that effects would not differ much across national contexts, either. In IR, at least some results have been found to be 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; 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; and Tomz, Weeks and Bansak (2023) find that formal military alliances have robust causal effects across 13 NATO countries. However, without systematic harmonized research 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 empirical crisis.

[6]Though, U.S. democracy scores may be biased (inflated) (Colgan, 2019*b*, 301; Levitsky and Ziblatt, 2019)

3 Research Design

3.1 Overview

The conception of generalizability developed above informs the design of our harmonized multi-site replications. We note four key features of our design. First, and perhaps most importantly, our study is designed to answer different questions than typical IR replication studies. Previous IR 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 (e.g. Tomz and Weeks, 2013; Lupu and Wallace, 2019). In contrast, we focus on the following two questions, each linked to an appropriate statistical test and research design:

1. In how many (and which) countries do we find statistically significant results in the theoretically expected direction? (Sign-generalization test)

2. Is there support—in the form of statistically significant results in the theoretically-expected direction—for a given theory in the pooled population of respondents across all our countries? (Meta-analysis)

A second important feature is our use of "purposive variation" for selecting country sites. This approach is designed to yield variation across sites along theoretically-important moderators (Egami and Hartman, 2022). It 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. It also allows us to make inferences about countries outside of the sample, subject to certain assumptions we discuss below.

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 sought harmony in terms of treatment (identical across countries), outcomes (identical across countries), timing (all experiments implemented simultaneously 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, which is particularly salient when evaluating generalizability.

Given our goals and these design features, 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 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, much less interrogated critically, in experimental research. However, when the goal is to learn about generalizability, site selection takes on added importance (Allcott 2015). Below, we detail the purposive country selection process (Egami and Hartman, 2022) 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 would provide little leverage here as an N of 7 countries does not permit strong inferences about a broader population of interest (i.e., all countries within the scope of the theory). On the other hand, non-random approaches have their own limitations. For example, convenience sampling—selecting sites based on ease of access—perpetuates the disadvantages of relying on U.S. samples: sites that are easiest for scholars to access may resemble the U.S. and differ systematically from less convenient sites. Alternatively, experimentalists might consider invoking the concept of "least-likely" (or "hard") cases from the qualitative methods literature. However, the "least-likely" approach is mainly designed to shed light on causal effects in the presence of confounding, which is not relevant in randomized experiments.[7]

We opt for a different non-random approach, using "purposive variation" to select sites that ensure vari-

[7] In qualitative methods, "least-likely" cases provide "hard tests" in that finding support for a hypothesis provides particularly strong evidence in favor of the relevant theory. For example, consider a theory involving an independent variable (X), a dependent variable (Y), and potential confounding variables (Z) related to both X and Y. In qualitative methods, a case is "least likely" if the observed level of X predicts a particular value of Y, but an *alternative* explanation (X) predicts a different value of Y (e.g., Gerring and Cojocaru, 2016). If Y takes on the value predicted by X even though other background variables predict a different outcome, the theory passes a "hard test," increasing confidence in its predictive power. Put differently, least-likely designs are meant to shed light on causal effects given potential confounding. However, confounding is already addressed in experiments by randomizing the treatment within each country: there are, by design, no uncontrolled variables that would predict a value of Y other than that predicted by the value of the independent variable. A better analogy from the qualitative literature is the concept of "causal heterogeneity," where, even if there is no confounding, the theory might predict the independent variable to have one effect in one context, and a different effect in another context (Seawright, 2016).

ation along theoretically-important moderators. This approach addresses two key issues. First, it provides a framework for investigating heterogeneity in treatment effects across countries due to *observed* moderators. Second, it addresses how to generalize existing evidence to *unobserved* contexts. Even when a study is conducted in multiple countries, its findings are inherently "local" and require additional assumptions to generalize elsewhere (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 sites 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 more general conclusions.

Given our interest in investigating variation in treatment effects *among* our selected countries and making inferences about countries *outside* of our data, it was critical to choose cases with sufficient variation in theoretically-relevant moderators. We specified four key theoretical components of each study (Findley, Kikuta and Denly, 2021): i) Treatment; ii) Mechanism; iii) Outcome; and iv) Moderators.

For three out of four studies, we identified 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.[8] Table 1 summarizes the theoretical components of all four studies and specifies the expected direction of the moderating effect.

After parsing the theories, our country selection proceeded systematically through through the process depicted in Figure 2 (detail in Appendix C). First, we *determined the scope conditions* of each theory and excluded countries outside those 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 meeting this scope condition by *policy importance*, prioritizing more powerful countries that are more consequential in world politics. This entailed sorting democracies based on their GDP and ranking more powerful countries over less powerful ones, all else equal, though without sacrificing key variation along moderators as described below.

Third, we aimed to maximize variation along traditional demographic factors (both measurable and

[8]When mechanisms and moderators were not discussed in detail, we built on the authors' theoretical framework to identify them. We contacted all authors to confirm our interpretations.

	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.

latent) by selecting one country from each major region of the world.[9] Fourth, we verified variation along our *predefined moderators*: military expenditures (to proxy for hawkishness), years since becoming a democracy (to proxy for democratic norms) and number of ratified human rights treaties (to proxy for international legal obligations). As demonstrated in the bottom-right panel of Figure 2, our selected countries yielded substantial variation, with at least two countries above and two below the cross-national mean of each moderating variable. Finally, we verified that Lucid/Cint operated in the selected countries and was able to match country samples on key demographics of the general population of interest (i.e., gender and age). Luckily, this step did not constrain case selection—and is thus not depicted in Figure 2—-as Lucid/Cint was able to offer samples from all selected countries (Brazil, Germany, India, Israel, Japan, Nigeria, and the U.S.).

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

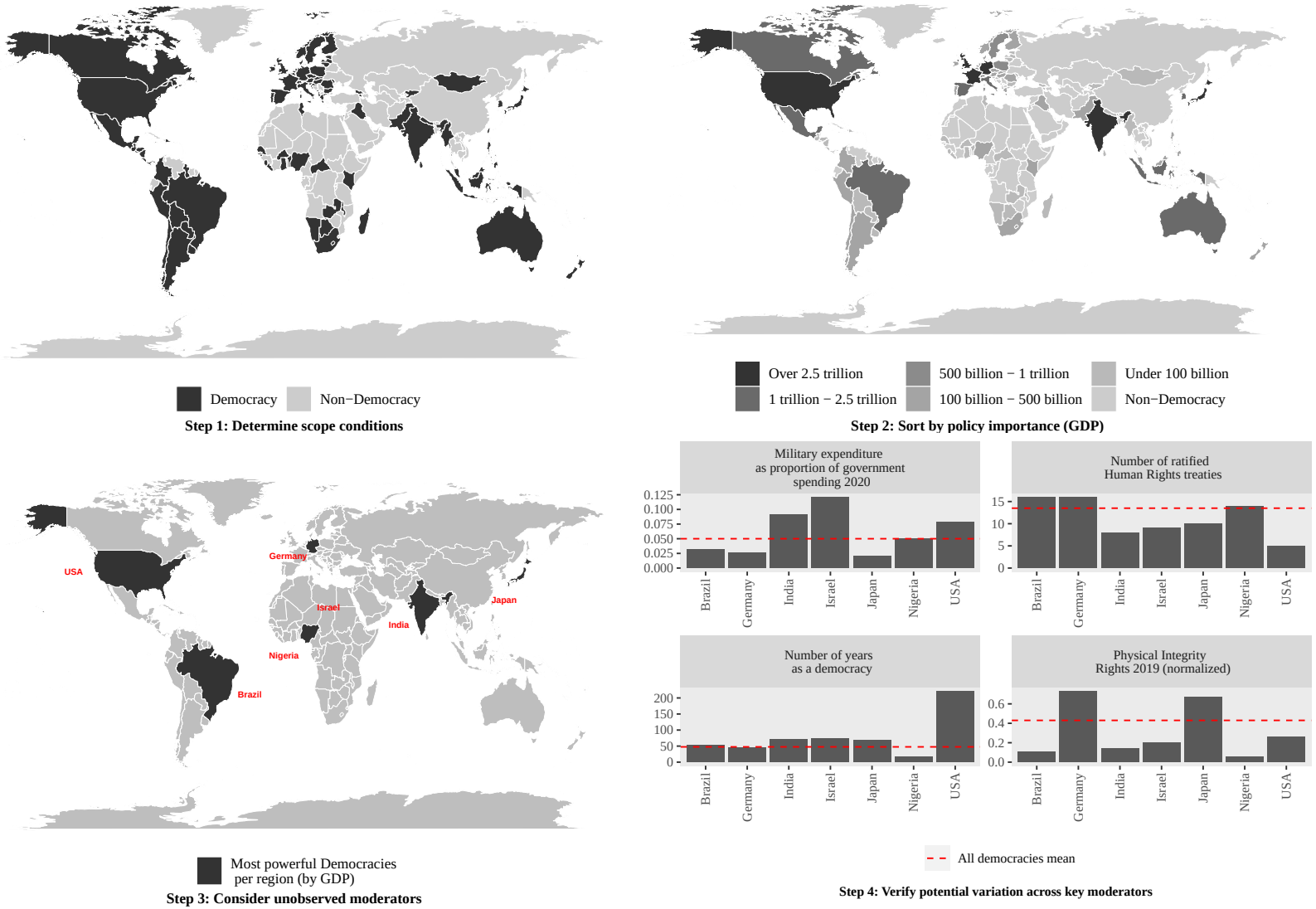

Figure 2: Steps in selecting countries for replication. GDP data is from the World Bank. Military expenditure data is from Stockholm International Peace Research Institute (SIPRI). Physical integrity rights data is from Fariss (2019).

3.3 Expectations and Analytical Strategies

Above, we identified two key questions about generalizability: (1) In *how many* (and which) of the countries do we find treatment effects in the theoretically expected direction? (2) Is there support for a given theory in the pooled population of respondents from all seven countries? To answer these questions, we report two key estimations—a sign generalization test and a meta-analysis—in both cases focusing on direction rather than magnitude of effects.

Sign-Generalization Test

First, to test the extent to which the direction of causal effects is generalizable, we use the sign generalization procedure proposed by Egami and Hartman (2022). This approach leverages design-based purposive variation (in our case, across countries) and employs a partial conjunction test to estimate the share of experiments yielding a precisely estimated effect in the theoretically expected direction. We consider a particular finding generalizable to the extent that support for it—in the form of precisely estimated ATEs in the theoretically expected direction—can be found across a variety of contexts within the bounds of a theory's scope conditions. The more various the contexts in which those results are found, the more generalizable a result would be.

The sign-generalization test has two key advantages. First, it allows us to directly answer our question of interest while properly accounting for multiple comparisons.[10] The intuition is that (for each study) we compute one-sided p-values separately for each country, sort them in order of size ($p_{(1)} \leq p_{(2)} \ldots \leq p_{(k)}$) and implement a partial conjunction test (for which no further adjustment for multiple comparisons is necessary).[11] The output is a percentage estimating the number (and identity) of countries in which a given treatment has a significant effect in the same direction.

The test's second advantage is its ability to generalize outside of our sample of countries. As Egami and Hartman (2022, 1081) explain, concerns about external validity are fundamentally about variation that is *not observed*. Even in a study such as ours with harmonized experiments across seven countries, we would like to know the extent to which our results generalize outside of our sample(s) to the broader population. Sign

[10] This is subtly different from tallying up the number of significant ATE's by study, which 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 thus should be interpreted on its own.

[11] For details, see Egami and Hartman (2022, 1082).

generalization lets us justify these inferences outside of our sample to the extent that the "range assumption" holds: the target population ATE (unobserved) is within the range of causal effects identified in our purposively selected countries. Because we selected countries to generate variation along key moderators, the range assumption is plausible (though inherently unprovable).

Meta-Analysis

Second, to identify the underlying support for a given theory across our (pooled) respondents, we use a meta-analytic research design, the generally-recognized gold standard for "combining data from multiple experiments..." (Blair and McClendon, 2021, 412)—in order to "obtain a more precise estimate of the ATE in the population" (Gerber and Green, 2012, 362). In contrast to many meta-analyses, which are "post-study" designs in which data from existing research is combined, we create our data through fielding a set of coordinated, simultaneous experiments (Blair and McClendon, 2021, 414).

The output is 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 bivariate (outcome $\sim$ treatment) 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 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 heterogeneity of the true effect across countries).[12]

[12]The meta-analytic random-effects model assumes that for our population of interest—-countries within our theoretical scope conditions—there exists a distribution of effect sizes for a given treatment. Under the assumption that our country-specific ATEs 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 likely not suffice—our approach 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 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 random effects 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).

Power and Interpreting Individual Null results

We determined our sample size by power analyses ensuring that we are well-powered (>80%) to identify original point estimates for each "input" into the metanalysis (i.e., within each country ($\alpha = 0.05$; see Figure A20). Because this is a more demanding standard, our estimates ensured that our other empirical test (sign-generalization) was extremely well powered (>99%, see Figure A21). While our visualizations below can sometimes draw attention towards differences in magnitudes of effects across various country or study combinations, we are not powered to detect those differences.

Of course, any given study-country combination may produce a null or even opposite result for various reasons, including random chance. Our conception of generalizability is not binary, so the existence of null results for a given experiment would not automatically yield the conclusion that a "study does not generalize." However, null results would provide evidence that a finding does not hold in a *particular* context and the more null findings that accumulate, the more circumspect our conclusions about generalizability would be.

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 probe such results. We pre-registered secondary analyses related to attentiveness, respondents having a particular country in mind, the plausibility of the scenarios, and effect heterogeneity along theoretically-relevant moderators.

4 Generalizability of IR Experiments: Results

We fielded our harmonized study in all seven countries in late January and early February 2023 using Cint.[13] For each country, we collected data from around 3,000 attentive respondents recruited to mirror the local population in terms of gender and age distribution. We allowed respondents to choose between English and the dominant national language.[14] Each survey started with a consent form, followed by attention checks embedded in a battery of pre-treatment measures of social and political dispositions. Attentive respondents

[13]Cint acquired Lucid in 2021. Our procedures were reviewed by the relevant Institutional Review Boards (IRBs) and designated "exempt." We followed all guidance for human subjects research published by the American Political Science Association. Cint/Lucid compensated respondents using their usual procedures. The appendix provides further details about human subjects considerations. All replication data available on Dataverse upon publication.

[14]We made the survey instrument available in the one or two dominant languages in all countries, requiring translation for all countries aside from the U.S. and Nigeria. The appendix shows that most subjects took the survey in their home-country language. A list of minor wording changes to address cross-site comparability is in Appendix D.2.

proceeded to our four experimental studies, whose order was randomized . Appendix D details our survey instruments. Appendix E reports descriptive statistics of each sample.

4.1 Strong Support for Sign-Generalization Among IR Experiments

Figure 3 reports results from sign generalization tests for each of the four experiments to evaluate in how many (and which) countries the sign of the result is consistent with theoretical expectations. As indicated by the flags and associated p-values, the audience costs experiment (top-left panel) yields a high level of sign generalizability with p-values < 0.05 across all seven countries. We obtain similar findings for the reciprocity FDI experiment (bottom right panel); p-values for all seven countries are again estimated to be < 0.05. The bottom-left panel of Figure 3 shows that for international law, the sign-generalization test yields five p-values < 0.05, suggesting sign generalizability of over 71%. Notably, however, the two remaining p-values > 0.05 are right around $p = 0.05$. 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 3, we find general support for sign generalization, with partial conjunction p-values < 0.05 for five out of seven countries. The countries with p-values > 0.05 are Nigeria ($p = 0.09$) and India ($p = 0.41$). We interpret the relatively small p-value in Nigeria ($p < 0.1$) as providing only suggestive evidence for sign generalization in that context. However, our data suggest that findings on the micro-foundations of the democratic peace theory do not generalize to our India sample, a finding we further interrogate in Section 4.3.

Overall, our sign generalization tests suggest that the experimental findings we replicate have a high degree of generalizability within our selected countries. The relatively high levels of generalizability found in our studies also engender some confidence that these findings would generalize outside of our sample to countries where the range assumption (detailed by Egami and Hartman 2022 and in section 3.3) is plausible.

4.2 Strong Underlying Support for Generalizability Using Meta-Analysis

Figure 4 reports the meta-analyses for all four experiments, assessing the underlying support for each theory across the pooled sample of countries. The top panel reports the meta-analytic average treatment effect for each experiment. These are based on the country-specific average treatment effects, shown in the middle

Figure 3: **Sign Generalization Test.** For each experiment, we report the proportion (r out of k) of country replications that generalize in the theoretically expected direction. Countries are denoted by flags and partial conjunction p-values are denoted above each flag.

panel along with 95% confidence intervals.[15] The bottom panel shows the point estimate and 95% confidence interval from the original studies—all using U.S. survey respondents—for reference.

To calibrate our replications, one can compare the direction and precision of the ATEs from the original studies (fielded in the U.S.) with the ATE from our U.S. sample (bottom row of middle panel). The ATEs from our U.S. sample converge with the original study ATEs in both statistical significance and direction. This suggests both that our studies (as fielded) were appropriately comparable to the original studies and helps rule out any temporal changes that might have occurred (and affected respondents' reactions) in the interim between the original studies and ours.

The general pattern of results in Figure 4 is both striking and reassuring: all four meta-analytic point estimates are precisely estimated in the same direction as those from published U.S.-based experiments. We interpret the overall pattern in Figure 4 as suggesting that average treatment effects 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 one would draw from studies in the U.S. are identical to those 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 those 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 of support for an effect in the opposite direction, and only three where point estimates are not statistically significant.

Although we did not preregister predictions about effect magnitudes, readers might be interested in what our results suggest on that dimension. For the audience costs and democratic peace studies, our meta-analytic ATEs appear slightly smaller than those in the original studies, while in the international law study the ATEs were similar, and in the reciprocity study, our meta-analytic ATE appears larger than the originally-estimated ATE. Future research might further investigate these potential patterns.

Together, the results in Figures 3-4 suggest optimism regarding the generalizability of IR experiments

[15] Our supplementary analyses further adjusted p-values for false discovery rates (Benjamini and Hochberg, 1995), applying Benjamini-Hochberg corrections at the experiment level accounting for seven tests of the same hypothesis, and do not change the interpretations of our findings.

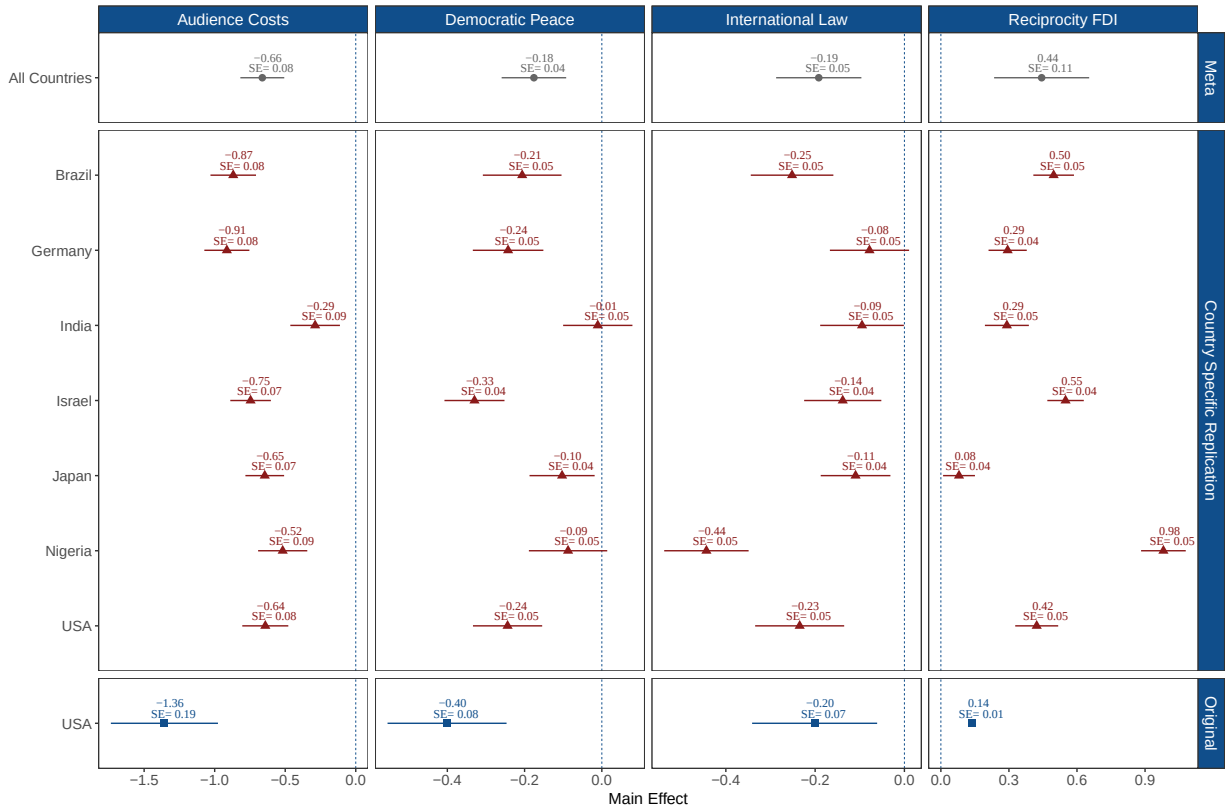

Figure 4: **Meta-Analysis.** For each experiment, we report original point estimates and standard errors from published studies, alongside country-specific ATEs and standard errors from our replications and a meta-analysis, based on our harmonized studies.

outside the U.S., using two different approaches. We now turn to pre-registered analyses designed to interrogate the one instance in which a study clearly did not replicate.

4.3 Probing the Null: Explaining Democratic Peace in India

The one clear exception to the pattern of generalizable results we found was the democratic peace study in India, where the effect of democracy on support for an attack yielded a precise null ($\beta = -0.01$, $p = 0.818$, $CI = [-0.1, 0.08]$). Fortunately, we anticipated potential null results and pre-registered analyses designed to shed light on such situations.

Appendix K provides a detailed overview of these analyses, which provide strong evidence against scenario implausibility, low attentiveness, ceiling or floor effects, or priming of specific countries (Dafoe, Zhang and Caughey, 2018) as explanations for the null democracy effect in India. Among our pre-specified moderators, we do find some evidence that low democratic norms in India might help explain at least part of the null effect, given that (1) support for democratic norms significantly attenuates the effect of the democracy treatment across our full sample of all countries (see Figure A9 and Figure A8, respectively); (2) our India sample exhibits the lowest support for democratic norms amongst our country samples ($\mu = 2.82$ in India compared to $\mu = 3.23$ for all other countries); and (3) we find suggestive evidence—in light of our limited power to detect within-country moderation effects—that norms do moderate treatment effects within India (see Table A15 and A12; the interaction between norms and the democracy treatment in India is estimated at $\beta = -0.14$, $p = 0.058$).

We speculate that the rest of the answer may involve historical dynamics surrounding ongoing conflict with neighboring Pakistan. Given that Pakistan has been considered a democracy for significant parts of its history (Marshall and Gurr, 2020), Indian respondents may have come to believe that democracies do not adhere to norms of peaceful conflict resolution and pose significant threats, undermining key mechanisms of the democratic peace (Tomz and Weeks, 2013). This result highlights the usefulness of empirical studies that probe scope conditions—both empirical and theoretical—and 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 four experiments across seven countries.

4.4 Explaining Generalizability: Limited Treatment Effect Heterogeneity

What explains the strong degree of correspondence between estimates from the U.S. and other countries? Below, we describe exploratory analyses designed to adjudicate between two possibilities: *sample characteristics* (i.e., low variation in the composition of the samples across countries) and *treatment effect heterogeneity*. High levels of treatment effect heterogeneity would describe a situation in which different kinds of people respond to treatment in very different ways. Overall, exploratory analyses suggest high variation in the composition of our samples across countries and limited evidence for treatment effect heterogeneity. Although we cannot definitively provide evidence for the obverse (treatment effect homogeneity), we conclude it is the most plausible explanation for the correspondence we observe. We provide evidence for limited heterogeneity by exploiting variation in *measured* moderators, by estimating *unmeasured* moderators using a test proposed by Ding, Feller and Miratrix (2016), and by contrasting results for our main studies with an extension of our audience cost study that was designed to have high levels treatment effect heterogeneity.

Sample Characteristics: One possible explanation for our consistent results might involve characteristics of the samples we collected. For example, perhaps our online convenience samples inadvertently selected for subjects who are particularly "WEIRD" or resemble U.S.-based respondents along other dimensions. Put simply, perhaps the treatment effects are similar because the people in the studies are similar. However, we find little support for this explanation. Figure A8 in 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.[16] Moreover, in Table A18, we formally test differences between country samples by regressing the moderators as well as a host of demographic variables (education, ideology, and age) over country indicators. If country samples vary along covariates (in comparison to the reference category: the U.S.) then inadvertent cross-country similarity in samples is unlikely to explain our main pattern of results. Since 34/36 of these estimates are significantly different from the U.S., we conclude that our country samples do indeed vary along both demographic and theoretically relevant covariates that we measured, and that cross-country similarity in samples is thus unlikely to explain our main pattern of results.

[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 A18.

Limited Treatment Effect Heterogeneity: A second possible explanation for the consistent pattern of results involves low treatment effect heterogeneity. If treatment effects are homogeneous, then differences between samples (such as the variation established above) do not matter for generalizing: "what is learned from any subgroup can be generalized to any other population of interest" because different people are reacting similarly to treatment (Coppock, 2019, 615). Indeed, 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.

Our first step in investigating this possibility was to evaluate how our results vary across individuals as a function of the theoretically-based moderators we measured (democratic norms for the democratic peace experiment; hawkishness in the audience costs experiment; and perceptions of international legal obligations in the international law experiment). Table A14 in the Appendix displays our results when pooling *across* country samples, demonstrating that (1) there are statistically significant moderating effects in our democratic peace and international law experiments but that (2) even in those cases the moderators never change the direction of the ATE, instead merely attenuating or amplifying the the treatment effect (Figures A9-A11). An alternate way to investigate treatment effect heterogeneity is to consider variation *within* country samples. Doing so, we again fail to find strong evidence of moderation along measured covariates: e.g., in the international law experiment, the moderator "perceptions of international legal obligation" has a significant attenuating effect in only one of seven countries and never reverses the sign of the effect (see Figure A12 and in Tables A15-A17). Overall, the results suggest that the moderation effects are substantively somewhat small and, if anything, shape the magnitude but not the direction of our ATEs.[17] One reason that these analyses are only suggestive, however, is that we may have limited power to examine variation within countries.

While the analyses above consider heterogeneity resulting from measured covariates, another avenue for investigation involves estimating the effect of *unmeasured* variation. To do so, we implement a procedure proposed by Ding, Feller and Miratrix (2016), which tests the null hypothesis that the treatment effect is constant across all units, allowing us to estimate the presence of significant unexplained variation within each country-study pair. Formally, the test leverages a Fisher Randomization Test (requiring minimal assump-

[17] Buttressing our interpretation that causal conclusions would stay the same even in populations very different from our samples are the results from an analysis of "external validity bias" contained in Appendix H.

tions) to test a null hypothesis of homogeneity in average treatment effects. In Table 2, we report results from this test, correcting for multiple comparisons as suggested by Coppock (2019). The table reports the number of models in which we can reject the null of constant treatment effects across units; higher numbers for a given experiment (row) suggest more unmeasured heterogeneity in treatment effects. Out of 28 country-study pairs in the main pre-registered analyses (above the horizontal line), only 8 show evidence of effect heterogeneity. Thus, in the majority of units (20/28 country-study pairs, or 71%), we cannot reject the null of homogeneity. Of course, there is no obvious bar for what constitutes "high" or "low" heterogeneity in this type of analysis, rendering a definitive interpretation difficult. Additionally, we may not be powered to detect small moderation effects across all country-study pairs.[18] Overall, however, these exploratory analyses do suggest that we cannot rule out treatment effect homogeneity as a plausible explanation for our strong pattern of generalizability.

As a final way to approach this issue, we turn to an analysis of an extension to our Audience Cost experiment. Recall that the pre-registered Audience Cost study was chosen in part because its effects were predicted to be relatively unconditional, and we found evidence to support this above: even extreme values of the moderator of "hawkishness" did not flip the sign of the ATE. However, we also fielded an extension to the main study based on the work of Kertzer and Brutger (2016), which differs from the main pre-registered study in that the extension decomposes audience costs 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; put differently, high levels of treatment effect heterogeneity.

By comparing the main audience cost study to the Kertzer and Brutger (2016) version, we can thus compare studies predicted to have varying levels of treatment effect heterogeneity. The results in our main analyses compared the "back down" to "stay out" conditions, but in the extension (described in Appendix J), respondents were assigned to three experimental conditions, allowing us to decompose the general audience cost into belligerence and inconsistency costs. Appendix J shows that, consistent with our expectation of differences across groups and contexts, the belligerence treatment (the effect of "engaging" versus "staying

[18]Nevertheless, our ad-hoc simulation study presented in Figure A13 shows that given our large sample size, we should be well powered to detect treatment effect heterogeneity on the scale of 0.15 SD.

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. The appendix (see Figure A10) further shows that hawkishness not only moderates belligerence costs in the audience costs extension, but that the sign on the treatment effect actually *flips* at high versus low levels of hawkishness, in line with Kertzer and Brutger's expectation that hawks will reward belligerence while doves punish it (see also Figure A17). Furthermore, the treatment effect homogeneity test proposed by Ding, Feller and Miratrix (2016) shows that there is treatment effect heterogeneity on unmeasured moderators in 100% of country-pairs (see bottom row of Table 2).[19] In sum, comparing our general pattern of results discussed in Section 4.2, 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 may be driven by treatment effect homogeneity.

Study	**N Comparisons**	**N Significant**	**N Significant (BH Adjustment)**
Audience Costs	7	3	2
Democratic Peace	7	3	0
International Law	7	3	2
Reciprocity FDI	7	5	4
Belligerence Cost (AC Extension)	7	7	7

Table 2: This table reports results for tests of unmeasured treatment effect heterogeneity developed by Ding, Feller and Miratrix (2016). The total number **N Comparisons** are the number of countries per study, while the next two columns denote the number of countries (per study) in which we can reject the null of homogeneous treatment effects, both raw and (in gray) after adjusting for multiple comparisons. The top four rows denote our main studies, while the last row refers to our Audience Cost extension discussed on Page 24.

[19] See also section G for comparison of I^2 statistics, testing for heterogeneity between country-samples.

5 Conclusion

This paper was motivated by concerns that the breadth of experimental evidence in international relations does not match the scope of its underlying theories. Although most IR theories make predictions intended to apply to a wide array of countries, past experimental studies on the micro-foundations of such 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 of four prominent IR studies across a set of seven democracies purposively chosen to ensure variation in key variables that could moderate the treatment effects we set out to test.

We found that all four experiments produced consistent results—in direction and significance—across a wide array of democracies. Our sign-generalizability analysis revealed that our replications exhibited consistent levels of generalizability—5 out 7 countries for the Democratic Peace and International Law experiments and 7 out of 7 countries for the Audience Cost and Reciprocity experiments. Our meta-analysis revealed significant meta-ATEs in the predicted direction for all four studies, and in no individual country did we find an effect in the "wrong" direction. In only one situation—the democratic peace experiment in India—did treatments yield a clear null effect, deviating from the overall pattern of results. Of course, 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, our replication of 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.

Consistent with other replication studies (Coppock, Leeper and Mullinix, 2018; Coppock, 2019), we found that the most plausible explanation for our general pattern of results relates to limited treatment effect heterogeneity. However WEIRD Americans may be, the U.S. does not appear to be an outlier when it comes to experimental results on the micro-foundations of IR theories. American respondents appear to differ from respondents in other countries in terms of key demographic attributes (Henrich, Heine and Norenzayan, 2010*b*), and may have atypical foreign policy preferences (see Figure A8 and Table A18 in the Appendix), but their reactions in our experiments were similar to those of subjects in other countries. This insight parallels other research documenting a strong degree of correspondence between different samples in

political science experiments (Coppock, Leeper and Mullinix, 2018; Kertzer, 2022). Thus, while it remains true that past experimental work has focused heavily on U.S.-based samples, we find little evidence that this reliance has led to wildly distorted conclusions about the micro-foundations of prominent theories of international relations.

These findings have striking implications for future research in both international relations and political science more broadly. On the one hand, our findings underscore the value of pre-registered and harmonized multi-site replication studies in the potentially limited instances in which scholars have resources to field such studies or are able to pool resources and coordinate their approaches. In contrast to uncoordinated single-site replications, coordinated approaches sidestep common challenges of design inconsistency that pose analytical hurdles for aggregating findings across contexts. Moreover, the transparency of such approaches limit 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 were able to learn how specific findings generalize, pinpoint one local instance of failed replication, and substantiate our interpretation that broader patterns of generalizability are explained by low effect heterogeneity in IR experiments testing general, rather than conditional, theories. Similar studies, when feasible, are a useful part of the research cycle in IR in which knowledge accumulates over time (McDermott, 2011*a*; Samii, 2016).

However, our findings also highlight the perhaps surprising potential value of single-country studies for testing the microfoundations of general IR theories, whether such studies are fielded in the U.S. or in other countries. In almost all of the 28 study-site combinations we examined, we found that the substantive conclusions one *would* have drawn from any one particular site would have been the same had one happened to choose a different site. For scholars with easy and/or inexpensive access to U.S.-based samples, our findings thus provide some reassurance that much can be learned from U.S.-based studies. At the same time, our findings should hearten scholars based outside the U.S., or who have convenient or inexpensive access to non-U.S.-based samples for other reasons, as their findings may have greater generalizability than previously believed. Our findings thus have the potential to improve access to experimental research for both U.S. and non-U.S. based scholars and to de-center the U.S. as the "standard" site for experimental research.

Our approach also offers guidance for how to place claims about generalizability on firmer theoretical

and empirical footing through deliberate choices at the design stage. Whenever possible, scholars should theoretically and empirically interrogate the extent to which their treatment effects are homogeneous versus heterogeneous. Ideally, 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 test for both observed and unobserved (Ding, Feller and Miratrix, 2016) heterogeneity and use these tests to inform arguments about generalizability. If treatment effects appear markedly heterogeneous, scholars should be cautious about making strong claims about generalizability. However, when treatment effects show relatively low heterogeneity – as we find in our study – bolder claims may be warranted.

References

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *The Quarterly Journal of Economics* 130(3):1117–1165.

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.

Blair, Graeme and Gwyneth McClendon. 2021. "Conducting experiments in multiple contexts." *Advances in experimental political science* pp. 411–428.

Borenstein, Michael, Larry V Hedges, Julian PT Higgins and Hannah R Rothstein. 2021. *Introduction to Meta-Analysis*. New Jersey: 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*. New York: 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* .

Chaudoin, Stephen. 2014. "Promises or Policies? An Experimental Analysis of International Agreements and Audience Reactions." *International Organization* 68(1):235–256.

Chaudoin, Stephen, Brian J Gaines and Avital Livny. 2021. "Survey design, Order Effects, and Causal Mediation Analysis." *The Journal of Politics* 83(4):1851–1856.

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. Persuasion in Parallel. In *Persuasion in Parallel*. Chicago: University of Chicago Press.

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. In *Do the Media Govern*, ed. Shanto Iyenger and Richard Reevs. London: Sage London 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*. New York: 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.

Gerber, Alan S and Donald P Green. 2012. "Field experiments: Design, analysis, and interpretation." *New York: W.W. Norton* .

Gerring, John and Lee Cojocaru. 2016. "Selecting cases for intensive analysis: a diversity of goals and methods." *Sociological Methods & Research* 45(3):392–423.

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. Albany: 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." *American Association for the Advancement of Science* .

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: Princeton University Pressniversity press.

Kertzer, Joshua D. 2016. *Resolve in International Politics*. Princeton: 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*. New York: 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.

Marshall, Monty G and Ted Robert Gurr. 2020. "Polity5: Political regime characteristics and transitions, 1800-2018." *Center for Systemic Peace* 2.

McDermott, Rose. 2011*a*. Internal and External Validity. In *Cambridge handbook of experimental political science*, ed. James N. Druckman, Donald P. Greene, James H. Kuklinski and Arthur Lupia. New York: Cambridge University Press pp. 27–40.

McDermott, Rose. 2011*b*. "New Directions for Experimental Work in International Relations." *International Studies Quarterly* 55(2):503–520.

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*. Princeton: 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 dDemocratic 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.

Seawright, Jason. 2016. *Multi-method social science: Combining qualitative and quantitative tools*. Cambridge University Press.

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

Smetana, Michal. 2024. "Microfoundations of domestic audience costs in nondemocratic regimes: Experimental evidence from Putin's Russia." *Journal of Peace Research* p. 00223433231220252.

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, Jessica LP Weeks and Kirk Bansak. 2023. "How membership in the North Atlantic Treaty Organization transforms public support for war." *PNAS nexus* 2(7):pgad206.

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

	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	-0.21	0.05	0.00	3060	-0.87	0.08	0	2004	-0.25	0.05	0.00	3053	0.50	0.05	0.00	3059
Germany	-0.24	0.05	0.00	3000	-0.91	0.08	0	1951	-0.08	0.05	0.08	3005	0.29	0.04	0.00	3013
India	-0.01	0.05	0.82	3075	-0.29	0.09	0	2019	-0.09	0.05	0.05	3070	0.29	0.05	0.00	3072
Israel	-0.33	0.04	0.00	3072	-0.75	0.07	0	2089	-0.14	0.04	0.00	3080	0.55	0.04	0.00	3068
Japan	-0.10	0.04	0.02	3056	-0.65	0.07	0	2029	-0.11	0.04	0.01	3063	0.08	0.04	0.03	3064
Nigeria	-0.09	0.05	0.11	3130	-0.52	0.09	0	2079	-0.44	0.05	0.00	3137	0.98	0.05	0.00	3137
USA	-0.24	0.05	0.00	3019	-0.64	0.08	0	2012	-0.23	0.05	0.00	3023	0.42	0.05	0.00	3019
All Countries	-0.18	0.04	0.00	21412	-0.66	0.08	0	14183	-0.19	0.05	0.00	21431	0.44	0.11	0.00	21432
Original (USA)	-0.40	0.08	0.00	1271	-1.36	0.19	0	451	-0.20	0.07	0.00	2792	0.14	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.

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.[20] 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).[21]

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 outside our selected countries. Moreover, we can carry out exploratory tests of moderation effects at the individual level. This can help place existing evidence in perspective,

[20] That is not to say, of course, that theories could generalize outside the theorized scope (Smetana, 2024), but it is not the purpose of our study to answer this question.

[21] 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.

informing our interpretation of any cross-site variation in average treatment effects. Thus, we use country-level measures to 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. We use data from the Stockholm International Peace Research Institute (SIPRI) on military expenditure as a proportion of government spending to proxy for hawkishness. We use the number of years a country has been a democracy and the Physical Integrity Rights Index to indicate strength of democratic norms. Finally, we use the number of human rights treaties a country has ratified to represent the level of international legal obligation.

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 Brazil, Germany, India, Israel, and Japan respondents could choose between two options – the local national language (Brazilian Portuguese, German, Hindi, Hebrew, and Japanese, accordingly), and English. In the U.S. and in Nigeria, since the local national language is English, it was only possible to take the survey in English. 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:

- **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 [respondent's country]? (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 [respondent's country] 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 one of the authors' 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 [respondent's country] 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 [respondent's country].
- The country does not have high levels of trade with [respondent's country].
- The country's nonnuclear military forces are half as strong as [respondent's country] 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, [respondent's country] could prevent the country from producing any nuclear weapons. Would you favor or oppose using the [respondent's country] military to attack the country's nuclear development sites? (5-point scale)
- By joining a joint international military mission now, [respondent's country] could prevent the country from producing any nuclear weapons. Would you favor or oppose [respondent's country] 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 [respondent's country] considered preventing another country from producing nuclear weapons. How believable is this situation for [respondent's country]? (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 [respondent's country] 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 [respondent's country leader] announced that [respondent's country] 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 [respondent's country leader] announced that if the attacking country continued to invade, [respondent's country] military would immediately engage and attempt to push out the attacking country. The [respondent's country leader] sent troops to the region. The attacking country continued to invade and the [respondent's country leader] ordered the [respondent's country] military to engage. The [respondent's country] 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 [respondent's country leader] announced that if the attacking country continued to invade, the [respondent's country] military would immediately engage and attempt to push out the attacking country. The [respondent's country leader] sent troops to the region. The attacking country continued to invade. The [respondent's country leader] ordered the [respondent's country] 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 [respondent's country leader] 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 [respondent's country] considered intervening in a foreign dispute. How believable is this situation for [respondent's country]? (Very believable/Somewhat believable/Not very believable/Not at all believable)
- In the passage you just read, the [respondent's country leader]: (Announced the [respondent's country] would stay out of the conflict and stayed out. / Announced that the [respondent's country] military would engage in the conflict, but ultimately did not engage. / Announced that the [respondent's 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 [respondent's country] is involved in a military conflict with another country. The other country is not a democracy. We will describe some interrogation methods that [respondent's country] 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. [Respondent's country] 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: [Respondent's country] 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 [respondent's country] was considering using interrogation methods. How believable is this scenario for [respondent's country]? (Very believable/Somewhat believable/Not very believable/Not at all believable)
- Did the passage you read mention that [respondent's country] 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 [respondent's country] policies involving the ability of foreign companies to purchase [respondent's country] companies. The situation is hypothetical but may reflect something that has happened in the past or could happen in the future.

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

Outcomes and Diagnostics

- Should [respondent's country] make it easier or harder for companies from this country to buy [respondent's country] companies? ([Respondent's country] should make it much easier / [Respondent's country] should make it somewhat easier / [Respondent's country] should make no change / [Respondent's country] should make it somewhat harder / [Respondent's country] 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 [respondent's country] was considering changing its policies on the ability of foreign companies to purchase [respondent's country] companies. How believable is this situation for [respondent's country]? (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 [respondent's country] companies to buy companies in their country. / Much harder for [respondent's country] 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 (see Table A4), as well as country-specific descriptive statistics tables (See Tables A5-A11).

F Diagnostics

In this Section, we report several diagnostic tests relating to our four experiments. First, in Table A12 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 A13, 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

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	21,281	3.208	1.405	1	5
DP outcome 2	21,275	3.752	1.229	1	5
AC outcome	21,303	4.427	1.922	1	7
IL outcome	21,293	2.711	1.428	1	5
FDI outcome	21,433	2.998	1.326	1	5
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

Statistic	N	Mean	St. Dev.	Min	Max
DP outcome	3,032	3.010	1.441	1	5
DP outcome 2	3,030	3.523	1.366	1	5
AC outcome	3,030	4.290	1.894	1	7
IL outcome	3,027	2.079	1.310	1	5
FDI outcome	3,058	2.959	1.286	1	5
Manipulation DP	3,028	0.392	0.488	0	1
Manipulation AC	3,027	0.270	0.444	0	1
Manipulation IL	3,023	0.606	0.489	0	1
Manipulation FDI	3,055	0.462	0.499	0	1
Democratic norms	3,058	3.207	0.730	1.000	5.000
Hawkishness	3,058	2.746	0.952	1.000	5.000
Intl legal obligation	3,058	4.099	0.692	1.333	5.000
Gender	3,058	0.493	0.500	0	1
Education	3,058	4.276	1.201	1	7
Eligable to vote	3,058	0.992	0.088	0	1
Age	3,058	38.813	13.896	18	74

Table A5: Descriptive Statistics - Brazil

Statistic	N	Mean	St. Dev.	Min	Max
DP outcome	2,988	2.594	1.282	1	5
DP outcome 2	2,988	3.136	1.264	1	5
AC outcome	2,992	4.016	1.883	1	7
IL outcome	2,989	2.033	1.241	1	5
FDI outcome	3,014	3.400	1.188	1	5
Manipulation DP	2,986	0.422	0.494	0	1
Manipulation AC	2,987	0.311	0.463	0	1
Manipulation IL	2,988	0.494	0.500	0	1
Manipulation FDI	3,011	0.469	0.499	0	1
Democratic norms	3,014	3.293	0.608	1.000	5.000
Hawkishness	3,014	2.490	0.919	1.000	5.000
Intl legal obligation	3,014	4.189	0.703	1.000	5.000
Gender	3,014	0.487	0.500	0	1
Education	3,014	3.824	1.197	1	7
Eligable to vote	3,014	0.973	0.162	0	1
Age	3,014	46.252	15.439	18	74

Table A6: Descriptive Statistics - Germany

Statistic	N	Mean	St. Dev.	Min	Max
DP outcome	3,056	3.754	1.267	1	5
DP outcome 2	3,054	4.194	0.998	1	5
AC outcome	3,061	5.402	1.896	1	7
IL outcome	3,058	3.605	1.325	1	5
FDI outcome	3,073	2.352	1.376	1	5
Manipulation DP	3,053	0.651	0.477	0	1
Manipulation AC	3,059	0.203	0.402	0	1
Manipulation IL	3,057	0.707	0.455	0	1
Manipulation FDI	3,071	0.285	0.452	0	1
Democratic norms	3,073	2.832	0.517	1.000	5.000
Hawkishness	3,073	3.545	0.832	1.000	5.000
Intl legal obligation	3,073	3.954	0.749	1.000	5.000
Gender	3,073	0.535	0.499	0	1
Education	3,073	5.243	0.948	1	7
Eligable to vote	3,073	0.992	0.092	0	1
Age	3,073	36.214	13.003	18	74

Table A7: Descriptive Statistics - India

Statistic	N	Mean	St. Dev.	Min	Max
DP outcome	3,053	3.994	1.112	1	5
DP outcome 2	3,051	4.165	1.006	1	5
AC outcome	3,058	4.567	1.759	1	7
IL outcome	3,057	3.180	1.227	1	5
FDI outcome	3,070	2.963	1.169	1	5
Manipulation DP	3,051	0.398	0.490	0	1
Manipulation AC	3,058	0.311	0.463	0	1
Manipulation IL	3,056	0.527	0.499	0	1
Manipulation FDI	3,068	0.500	0.500	0	1
Democratic norms	3,070	3.434	0.624	1.250	5.000
Hawkishness	3,070	3.279	0.865	1.000	5.000
Intl legal obligation	3,070	3.681	0.864	1.000	5.000
Gender	3,070	0.501	0.500	0	1
Education	3,070	4.353	1.189	1	7
Eligable to vote	3,070	0.973	0.161	0	1
Age	3,070	41.516	15.455	18	74

Table A8: Descriptive Statistics - Israel

Statistic	N	Mean	St. Dev.	Min	Max
DP outcome	3,035	2.334	1.189	1	5
DP outcome 2	3,035	3.119	1.215	1	5
AC outcome	3,041	3.978	1.639	1	7
IL outcome	3,041	2.018	1.102	1	5
FDI outcome	3,063	3.476	0.992	1	5
Manipulation DP	3,035	0.352	0.478	0	1
Manipulation AC	3,041	0.344	0.475	0	1
Manipulation IL	3,039	0.489	0.500	0	1
Manipulation FDI	3,062	0.555	0.497	0	1
Democratic norms	3,063	3.319	0.541	1.000	5.000
Hawkishness	3,063	2.181	0.869	1.000	5.000
Intl legal obligation	3,063	3.924	0.731	1.000	5.000
Gender	3,063	0.489	0.500	0	1
Education	3,063	4.262	1.049	1	7
Eligable to vote	3,063	0.991	0.095	0	1
Age	3,063	47.255	15.048	18	74

Table A9: Descriptive Statistics - Japan

Statistic	N	Mean	St. Dev.	Min	Max
DP outcome	3,113	3.332	1.446	1	5
DP outcome 2	3,113	4.191	1.061	1	5
AC outcome	3,116	4.250	2.107	1	7
IL outcome	3,114	3.297	1.362	1	5
FDI outcome	3,137	2.672	1.491	1	5
Manipulation DP	3,111	0.487	0.500	0	1
Manipulation AC	3,115	0.278	0.448	0	1
Manipulation IL	3,113	0.473	0.499	0	1
Manipulation FDI	3,133	0.522	0.500	0	1
Democratic norms	3,137	2.959	0.521	1.000	5.000
Hawkishness	3,137	3.055	0.873	1.000	5.000
Intl legal obligation	3,137	3.940	0.740	1.000	5.000
Gender	3,137	0.513	0.500	0	1
Education	3,137	6.151	1.892	1	11
Eligable to vote	3,137	0.988	0.109	0	1
Age	3,137	32.741	11.228	18	73

Table A10: Descriptive Statistics - Nigeria

Statistic	N	Mean	St. Dev.	Min	Max
DP outcome	3,004	3.418	1.263	1	5
DP outcome 2	3,004	3.912	1.045	1	5
AC outcome	3,005	4.474	1.870	1	7
IL outcome	3,007	2.731	1.405	1	5
FDI outcome	3,018	3.182	1.345	1	5
Manipulation DP	3,002	0.489	0.500	0	1
Manipulation AC	3,003	0.312	0.464	0	1
Manipulation IL	3,006	0.564	0.496	0	1
Manipulation FDI	3,015	0.462	0.499	0	1
Democratic norms	3,018	3.224	0.625	1.000	5.000
Hawkishness	3,018	3.296	0.830	1.000	5.000
Intl legal obligation	3,018	3.850	0.870	1.000	5.000
Gender	3,018	0.484	0.500	0	1
Education	3,018	4.317	1.173	1	7
Eligable to vote	3,018	0.968	0.176	0	1
Age	3,018	45.627	15.333	18	74

Table A11: Descriptive Statistics - USA

	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 A12: Manipulation Test: Treatment Effects on Correct Recall

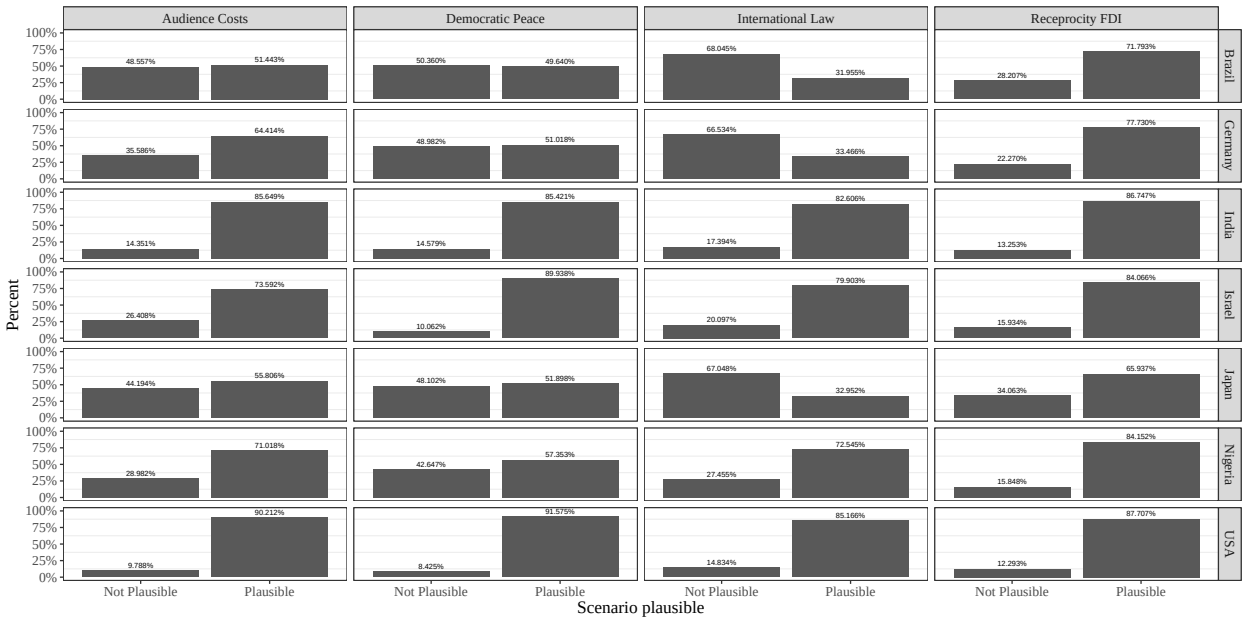

Figure A1: **Plausibility of Experimental Scenarios Across Countries.** This plot reports respondents' evaluation of how plausible an experimental vignette is, by experiment, per country.

	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 A13: Information Leakage: ATE on thinking of a specific country

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 Heterogeneity

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 A14.

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

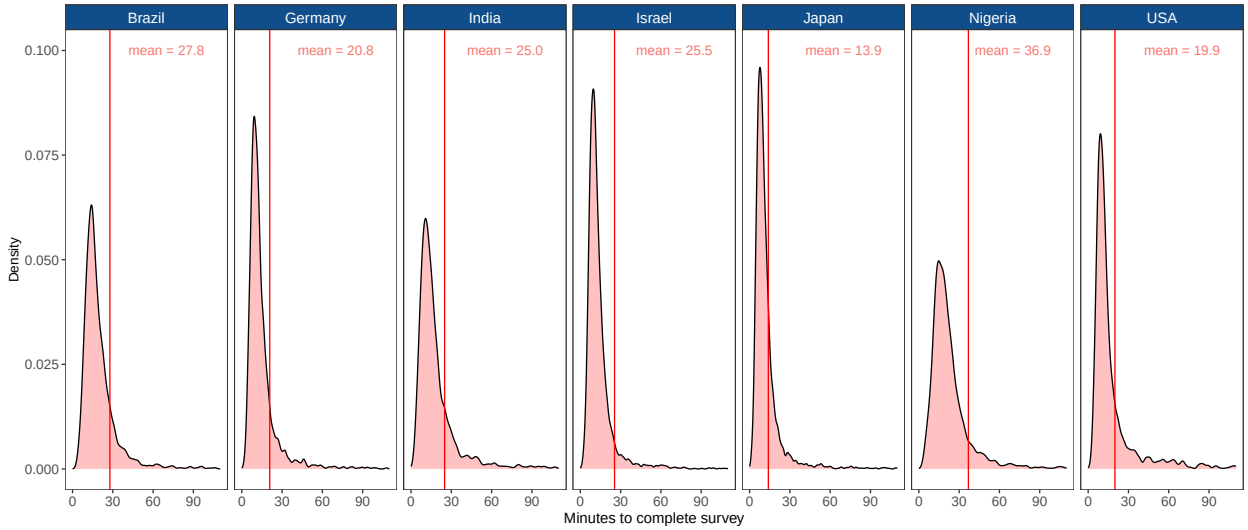

Figure A2: **Response Latency.** Figure reports the density of duration (in minutes) it took respondents to complete the survey. Averages and medians (in minutes) for each country are reported in red and blue, respectively, at the top of each figure.

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.

Next, report of I^2 statistics from our meta-analyses, calculated to be 83.4% for the democratic peace study, 85.3% for the audience costs study, 87.3% for the international law study, 97.6% for the reciprocity/FDI study, and 98.5% for the belligerence costs (audience costs extension) study. This implies that in each of our experiments, considerable heterogeneity *between* our country samples is present. Importantly, however, I^2 refers to heterogeneity across country-samples and not within them. As one may expect, in the audience costs extension where results do not replicate as well (the direction of the effect varies by different contexts), I^2 is highest.

In our manuscript, we report results of a test of treatment effect heterogeneity developed by Ding, Feller and Miratrix (2016). This test estimates the level of *unobserved* variation across individuals within the same country. A limitation of Ding, Feller and Miratrix's approach, however, is that like other tests of heterogeneity, it can be underpowered (Gerber and Green, 2012, p. 293). To address this concern more seriously, we follow Coppock (2019) and conduct a simulation analysis, varying the number of subjects per treatment arm and the degree of treatment effect heterogeneity (see Coppock, page 10). The results, presented in the figure below, show that we would be well powered to detect treatment heterogeneity on the scale of 0.15SD. Because we are relatively well-powered to detect small effects, and because, by contrast, we reject the null of homogeneity in 7/7 country samples of the audience costs extension study, we conclude that treatment effect homogeneity is a plausible explanation for our patterns of generalizability.

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. In effect: how different would population would have to be from our experimental sample in order to eliminate the treatment effect? External validity bias

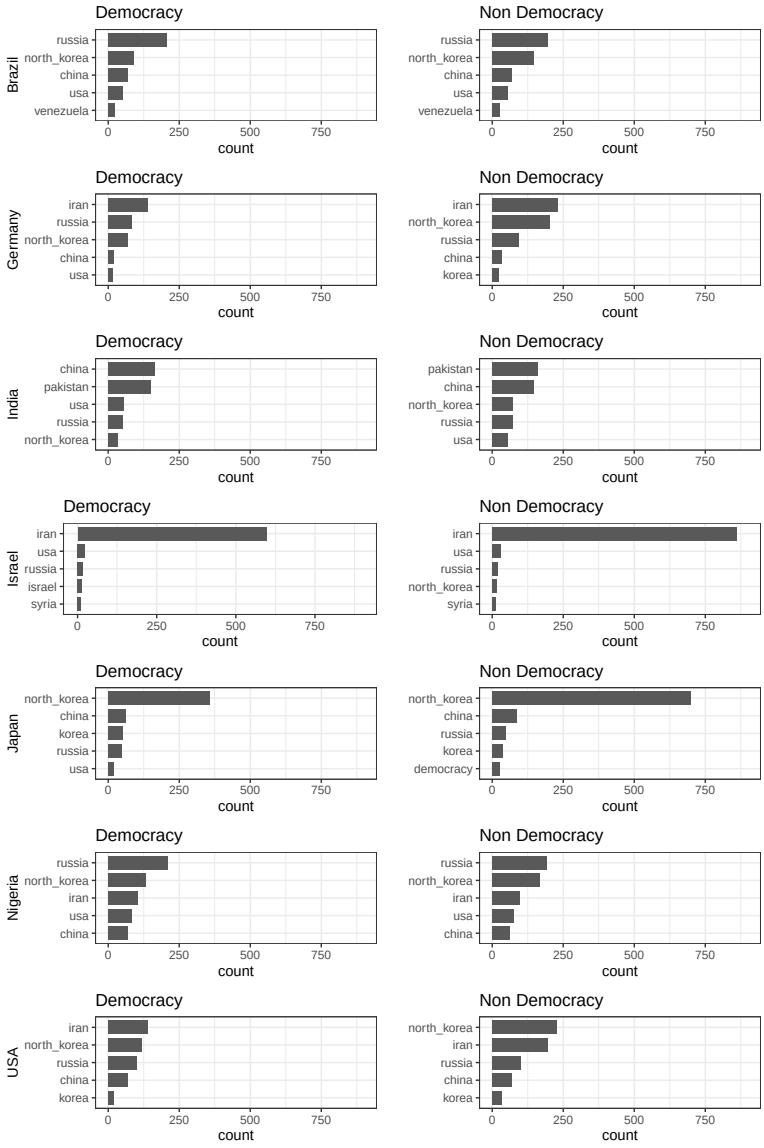

Figure A3: **Top 5 countries respondents thought of in the Democratic Peace study.** We report the top 5 most used words in an open-ended question asking respondents whether they thought of a specific country when reading the democratic peace scenario. X-axis orders the words from most to least mentioned.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

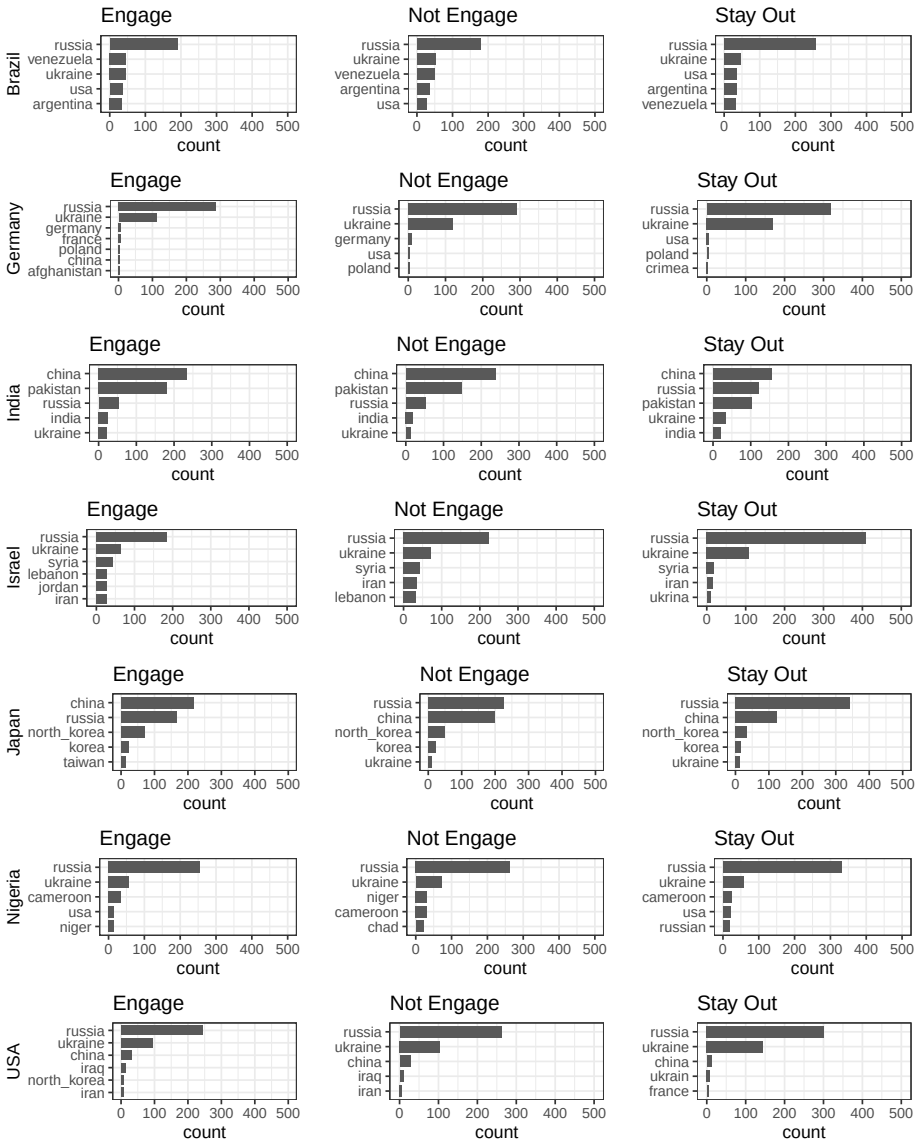

Figure A4: **Top 5 countries respondents thought of in the Audience Costs study.** We report the top 5 most used words in an open-ended question asking respondents whether they thought of a specific country when reading the audience costs scenario. X-axis orders the words from most to least mentioned.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

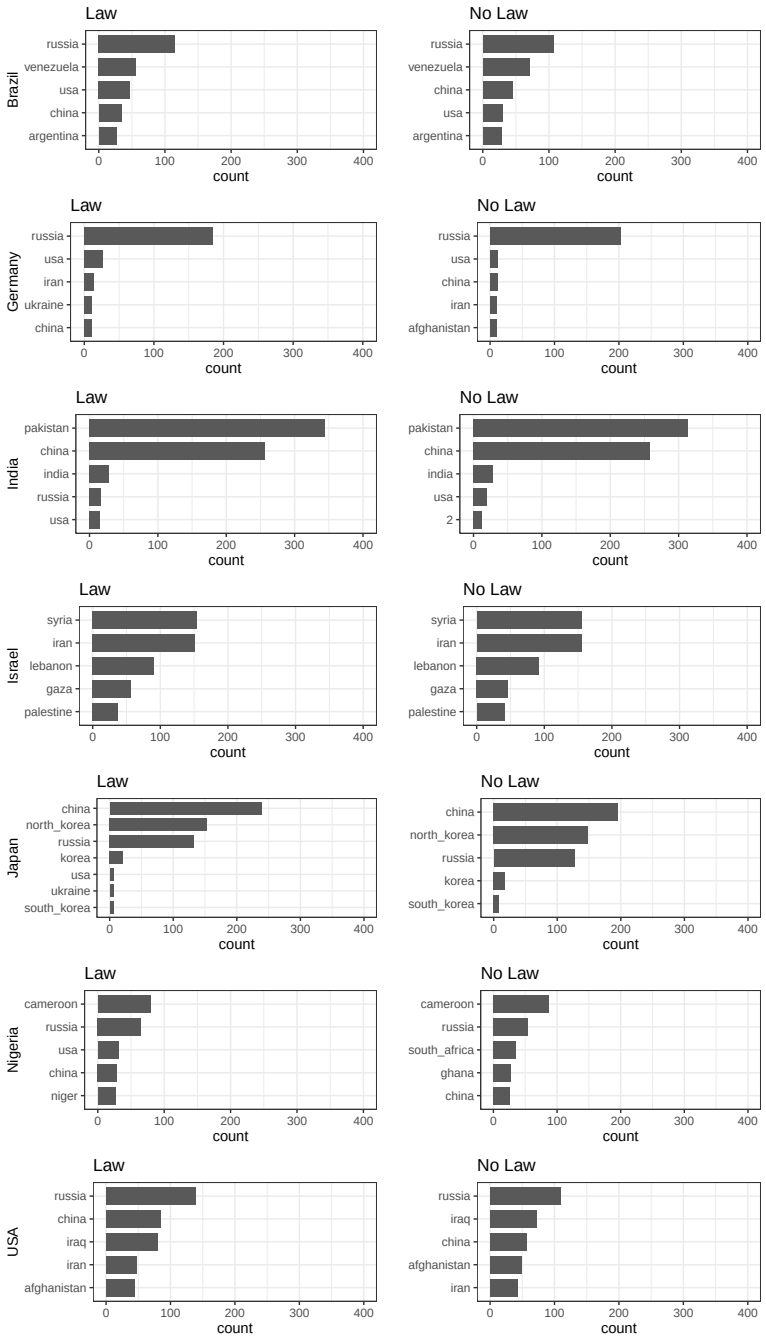

Figure A5: **Top 5 countries respondents thought of in the International Law study.** We report the top 5 most used words in an open-ended question asking respondents whether they thought of a specific country when reading the international law scenario. X-axis orders the words from most to least mentioned.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

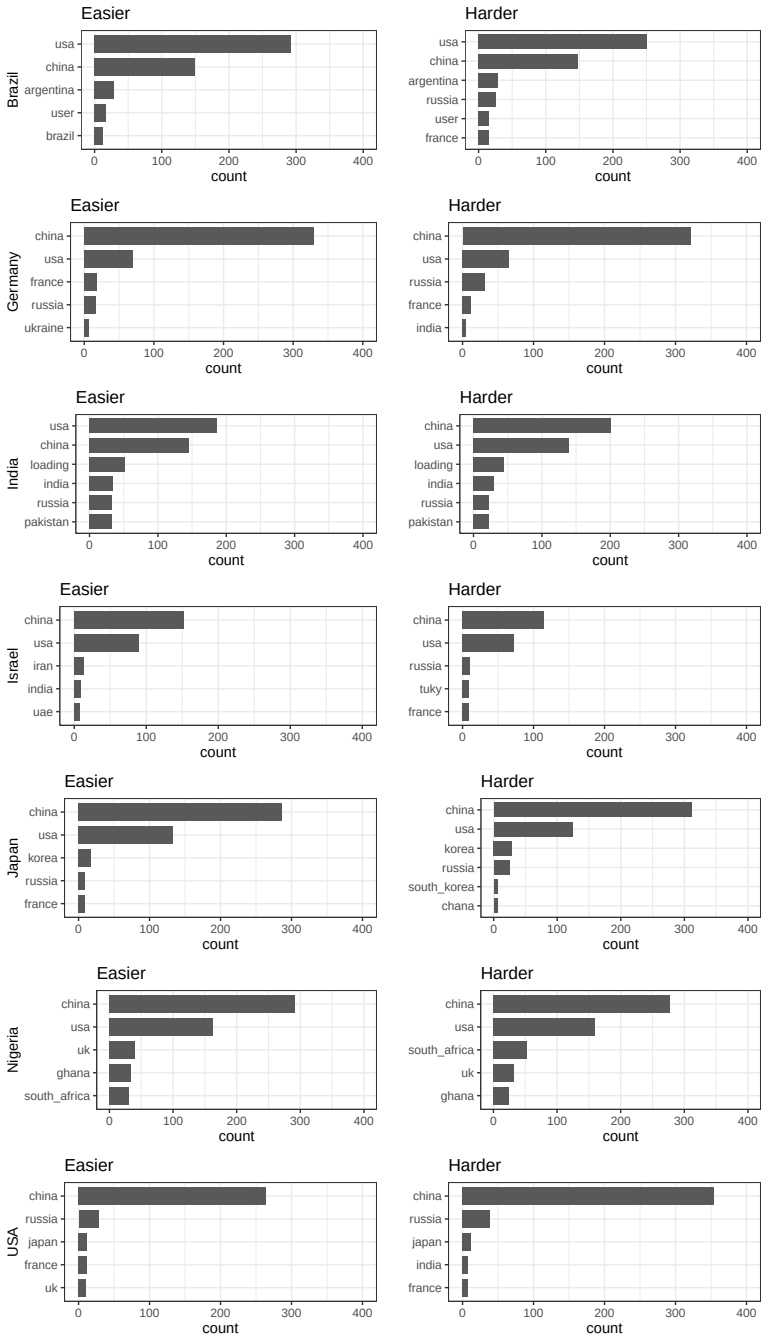

Figure A6: **Top 5 countries respondents thought of in the Reciprocity FDI study.** We report the top 5 most used words in an open ended question asking respondents whether they thought of a specific country when reading the reciprocity FDI scenario. X-axis orders the words from most- to least- mentioned.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

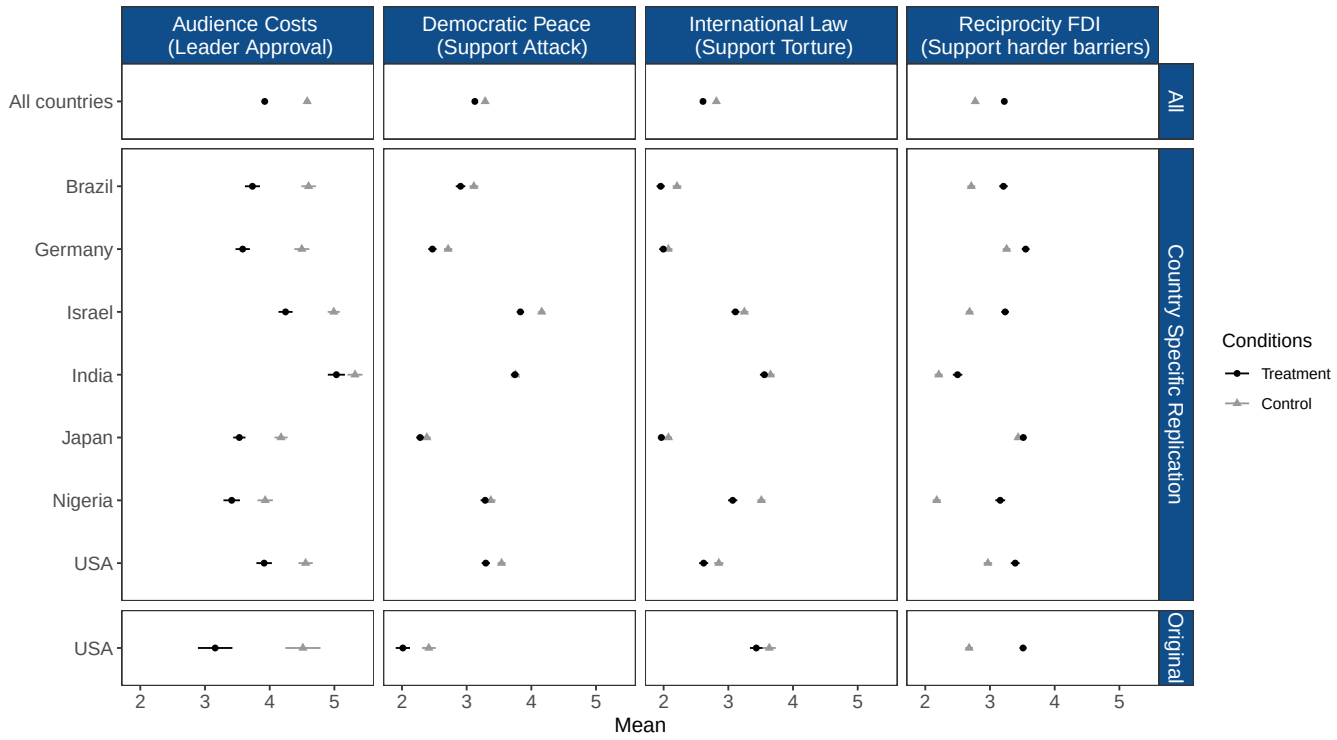

Figure A7: **Means by condition.** For each experiment, we report original means and confidence intervals of main treatment and control conditions from published studies, alongside means and confidence intervals from our country-specific replication and from the full multi-national sample ("Meta"). To 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.

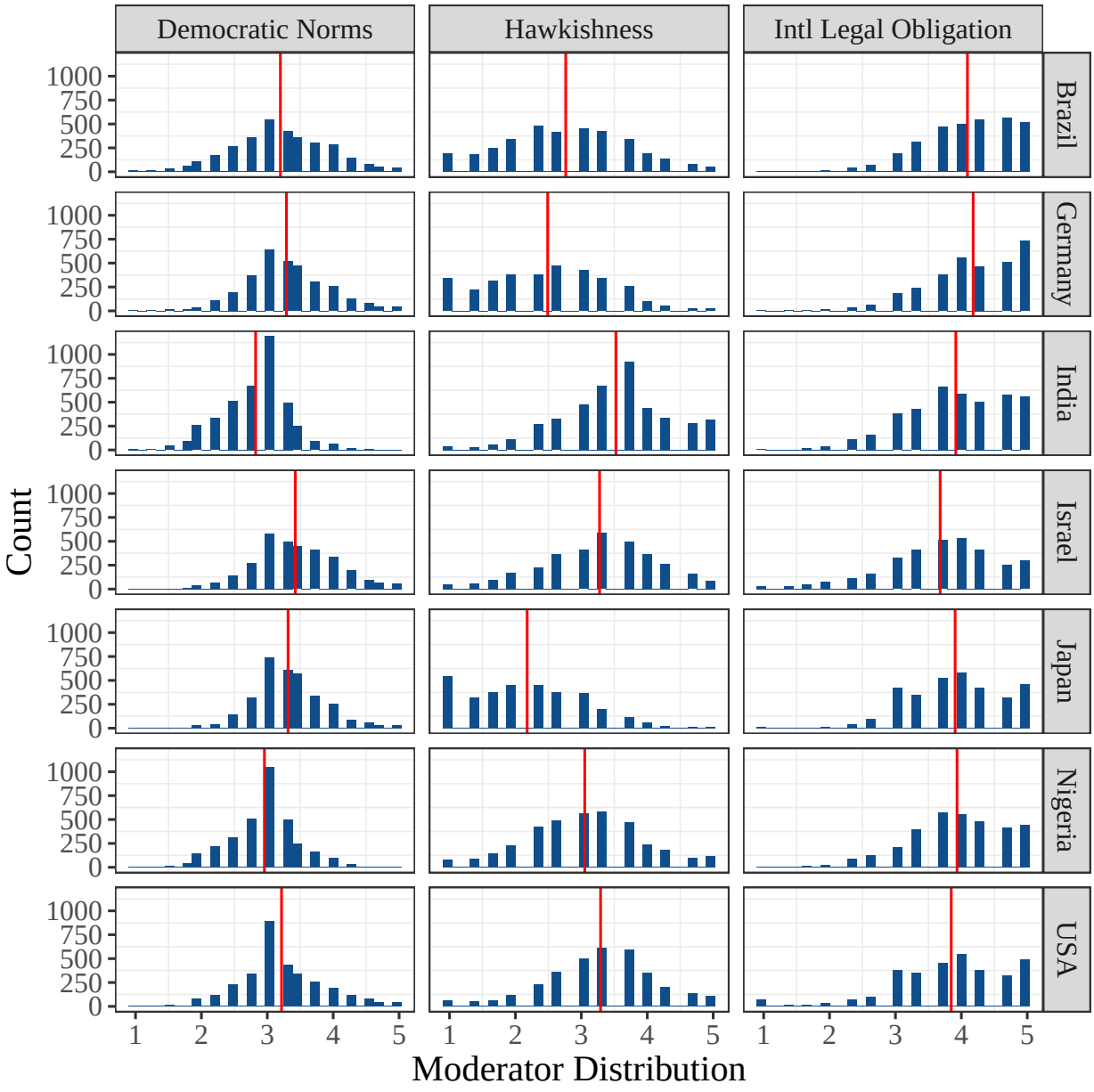

Figure A8: **Distribution of Moderators Across Countries.**

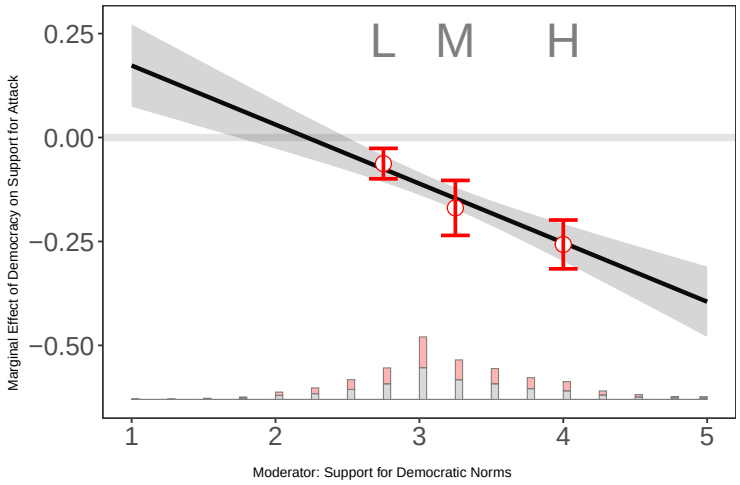

Figure A9: **Moderating Effect of Support for Democratic Norms Index in the Democratic Peace Experiment.** This figure demonstrates the negative moderation of support for democratic norms on the democracy treatment effects. That is, the effect of describing a country as a democracy reduces support for attacking the said country, and the effects are smaller (larger) for respondents with low (high) levels of support for democracy. This figure corresponds to Table A14.

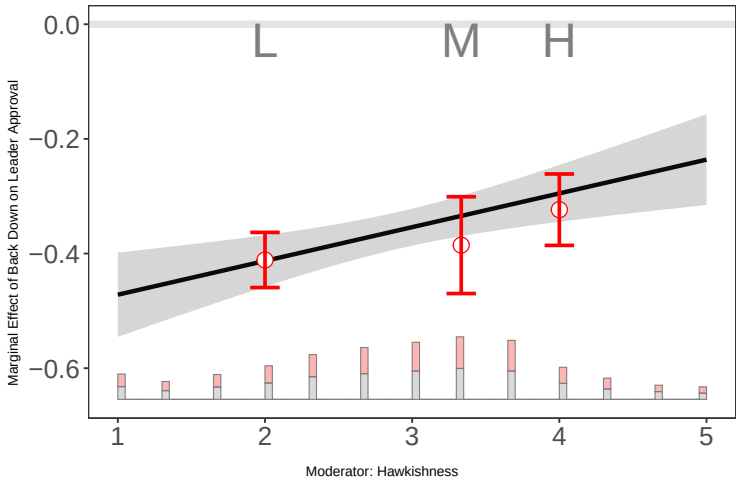

Figure A10: **Moderating Effect of Hawkishness Index in the Audience Costs Experiment.** This figure demonstrates there is no strong evidence for a moderation of hawkishness on the audience costs experiment. This figure corresponds to Table A14.

	Democratic Peace	Audience Costs	International Law
	Model 1	Model 2	Model 3
Democracy	−0.136*		
	(0.013)		
Dem Norms	−0.160*		
	(0.017)		
Dem*Norms	−0.106*		
	(0.024)		
Back Down		−0.357*	
		(0.016)	
Hawkish		0.039*	
		(0.014)	
BD*Hawk		0.028	
		(0.020)	
Intl Law			−0.149*
			(0.013)
Legal Obligation			−0.036*
			(0.013)
IL*Oblig			−0.061*
			(0.018)
Adj. R^2	0.033	0.050	0.037
Num. obs.	21426	14197	21445

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

Table A14: Moderation Tests

	Brazil	Germany	India	Israel	Japan	Nigeria	USA
Democracy	−0.142*	−0.169*	−0.013	−0.304*	−0.085*	−0.059	−0.197*
	(0.036)	(0.035)	(0.036)	(0.035)	(0.035)	(0.036)	(0.035)
Dem Norms	−0.063	−0.352*	−0.013	−0.087*	−0.390*	−0.066	−0.167*
	(0.036)	(0.042)	(0.054)	(0.042)	(0.047)	(0.052)	(0.044)
Dem*Norms	−0.095	−0.023	−0.138	−0.226*	−0.015	−0.043	−0.119
	(0.052)	(0.059)	(0.073)	(0.063)	(0.068)	(0.071)	(0.062)
Adj. R^2	0.029	0.110	0.017	0.057	0.064	0.002	0.083
Num. obs.	3062	3002	3077	3074	3058	3132	3021

*$p < 0.05$

Table A15: Moderation Test (Democratic Peace)

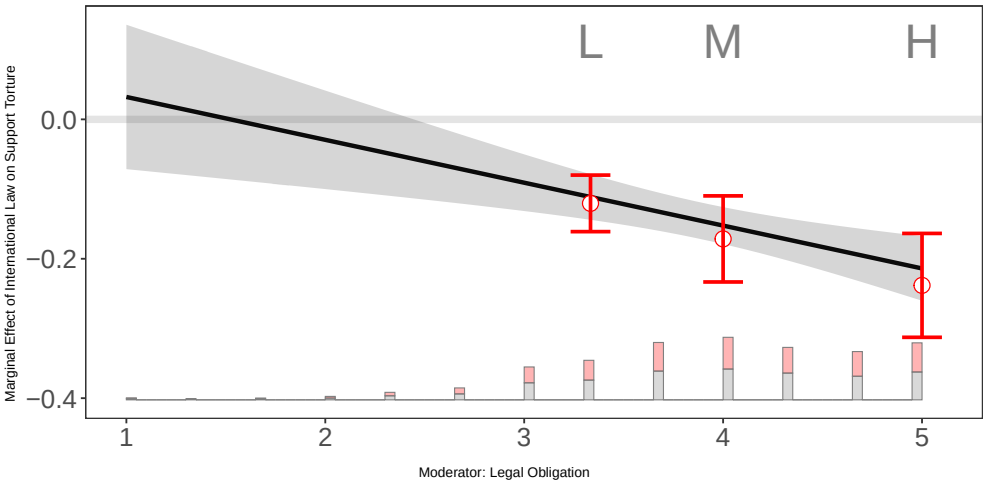

Figure A11: **Moderating Effect of International Legal Obligation Index in the International Law Experiment.** This figure demonstrates the negative moderation of legal obligation on the international law treatment effects. That is, mentioning that the respondent's country signed international law treaties 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. This figure corresponds to Table A14.

	Brazil	Germany	India	Israel	Japan	Nigeria	USA
Back Down	−0.460*	−0.492*	−0.153*	−0.424*	−0.380*	−0.247*	−0.345*
	(0.043)	(0.043)	(0.046)	(0.041)	(0.042)	(0.042)	(0.043)
Hawkish	0.008	−0.209*	0.219*	0.190*	−0.111*	0.135*	0.057
	(0.035)	(0.038)	(0.040)	(0.034)	(0.041)	(0.035)	(0.041)
BD*Hawk	0.071	0.252*	0.098	−0.213*	0.029	−0.040	0.014
	(0.049)	(0.053)	(0.058)	(0.053)	(0.056)	(0.050)	(0.059)
Adj. R^2	0.068	0.093	0.066	0.062	0.057	0.027	0.090
Num. obs.	2006	1953	2021	2091	2031	2081	2014

*$p < 0.05$

Table A16: Moderation Test (Audience Costs)

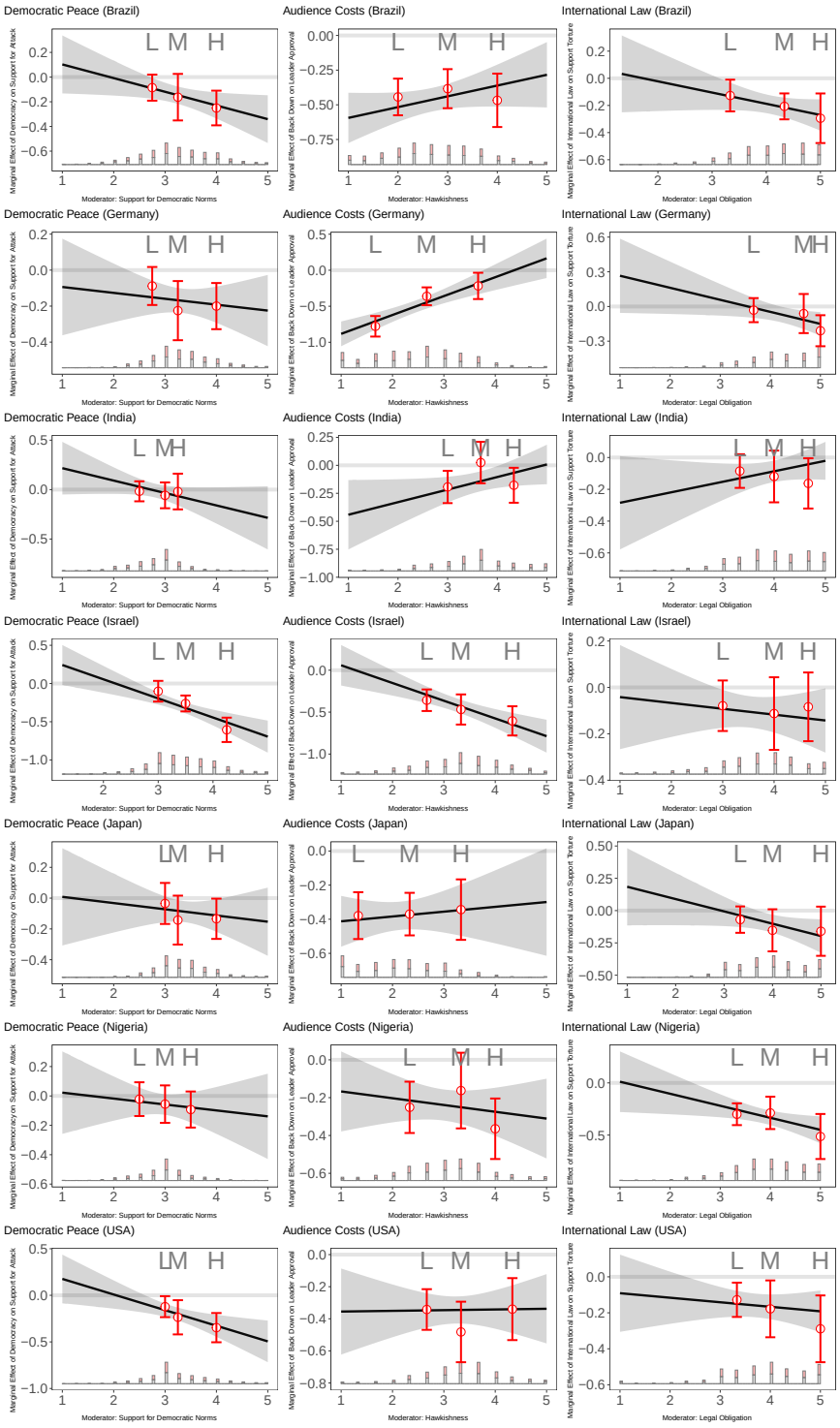

Figure A12: **Moderating effects in country-samples.** Individual figures of the moderating effects of democratic norms, hawkishness, and international legal obligation, in the DP, AC, and IL experiments, accordingly. Figures are broken down by country-samples. This figure corresponds to Tables A15-A17.

	Brazil	Germany	India	Israel	Japan	Nigeria	USA
Intl Law	−0.196*	−0.066	−0.090*	−0.109*	−0.094*	−0.330*	−0.162*
	(0.035)	(0.034)	(0.034)	(0.035)	(0.035)	(0.035)	(0.034)
Legal Obligation	−0.041	−0.214*	0.338*	−0.180*	−0.188*	0.151*	−0.139*
	(0.038)	(0.037)	(0.034)	(0.031)	(0.039)	(0.033)	(0.028)
IL*Oblig	−0.068	−0.053	0.049	−0.028	−0.102	−0.112*	−0.015
	(0.052)	(0.053)	(0.050)	(0.043)	(0.055)	(0.050)	(0.039)
Adj. R^2	0.070	0.115	0.098	0.072	0.049	0.038	0.147
Num. obs.	3055	3007	3072	3082	3065	3139	3025

* $p < 0.05$

Table A17: Moderation Test (International Law)

	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 A18: Estimating Differences Between Country Samples. Each model regresses relevant outcomes over country indicators compared to the US (which serves as a reference category).

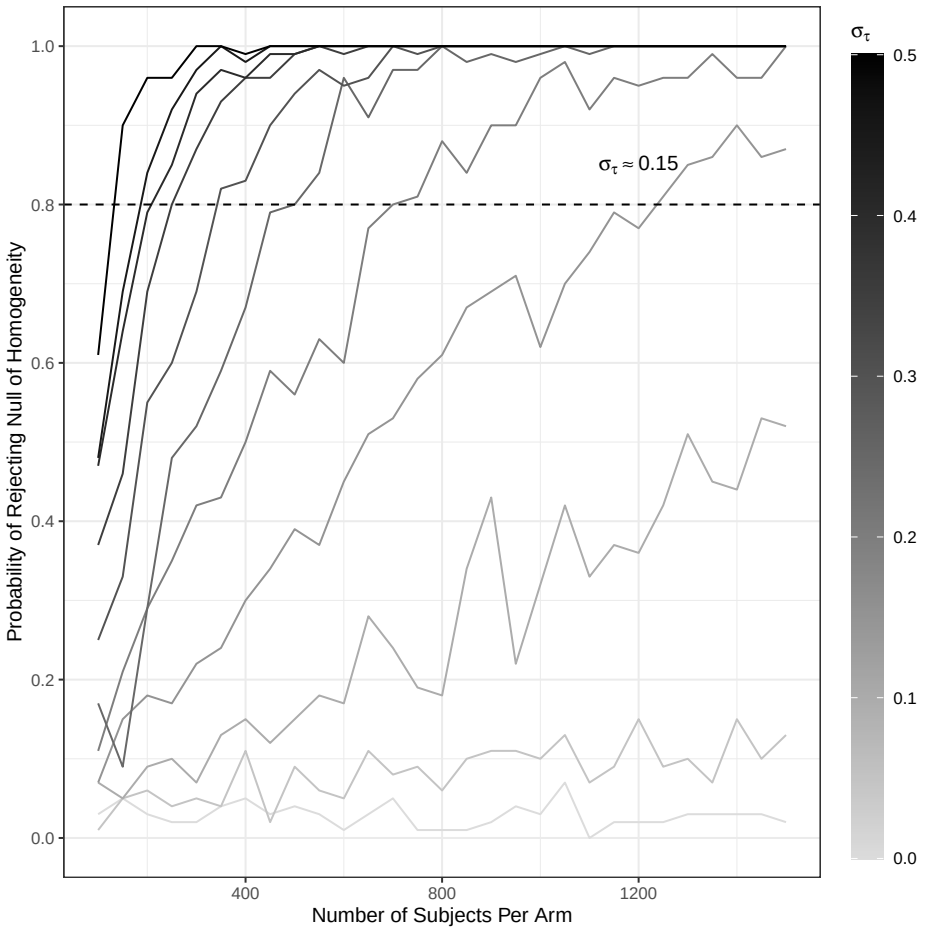

Figure A13: **Simulation study.** Power of Ding, Feller and Miratrix (2016) heterogeneity test, using R code from Coppock (2019).

	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 A19: Alternative Outcome (DP): Treatment Effect on Joining Intl Mission

depends on both the level of treatment effect heterogeneity and the size of the treatment effect (Devaux and Egami, 2022, 11).

Our results suggest that in most country-experiment pairs, causal conclusions would stay the same even in populations very different from our samples. In figure A14, we plot the estimated external robustness and the distribution of estimated CATEs for each country-study combination. 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). Specifically, in 21/28 country-experiment pairs, robustness to external validity bias is larger (> 0.57) than Devaux & Egami's more stringent benchmark for robustness.[22]

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 A19 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 A15, 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

[22] 0.57 is equal to the amount of reweighting required for MTurk samples to approximate nationally representative populations, which is relatively large. "This suggests that experimental findings have relatively high external robustness because causal estimates will be equal to zero only when the experimental sample is as different from a hypothetical population as the MTurk samples are from the U.S. general population" (Devaux and Egami, 2022, 18).

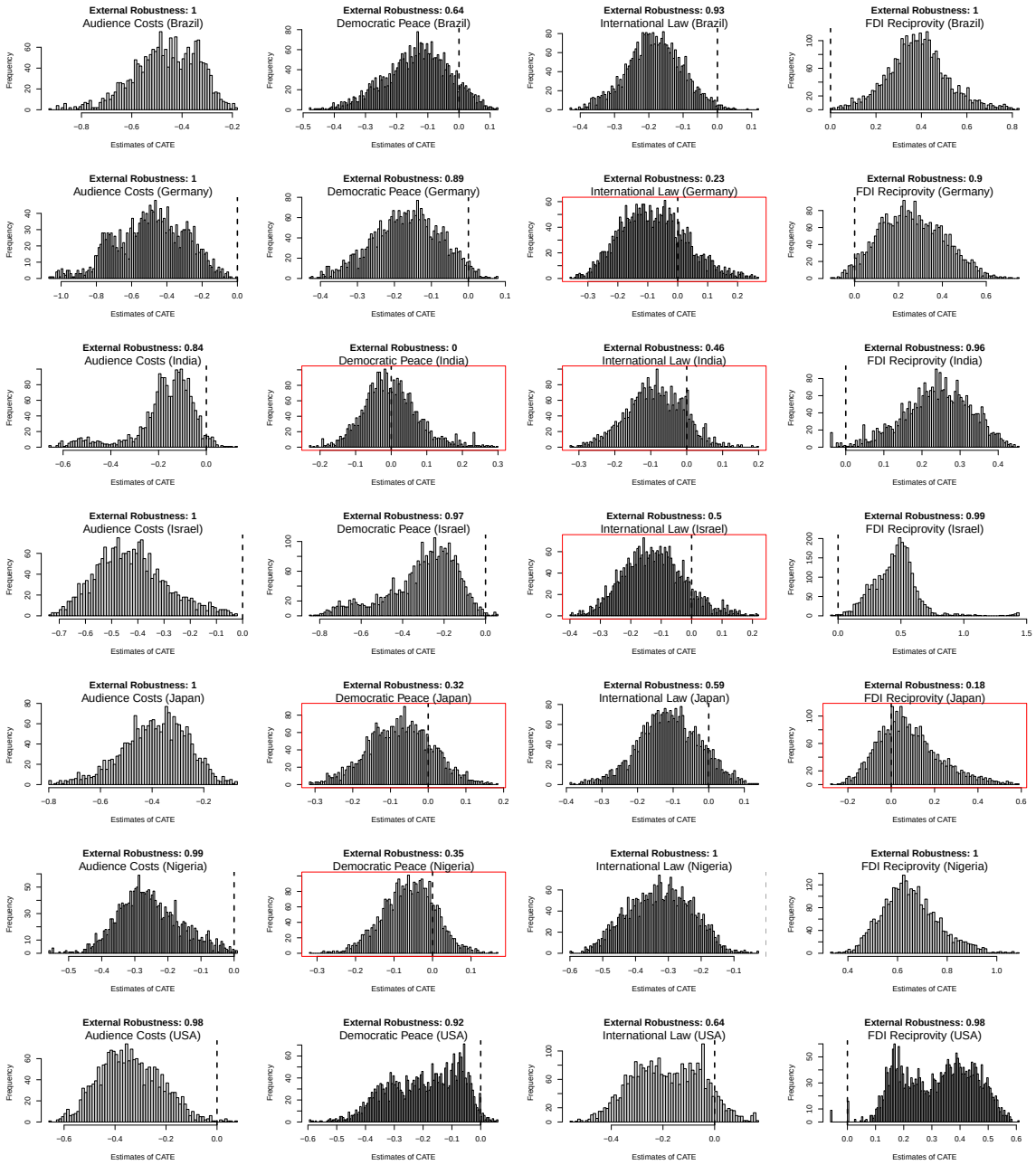

Figure A14: **External Validity Bias Test.**

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 A16. 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 A16, 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 A16 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.[23]

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 the 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.

[23] We suspect that at least two results from Figure A15 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.

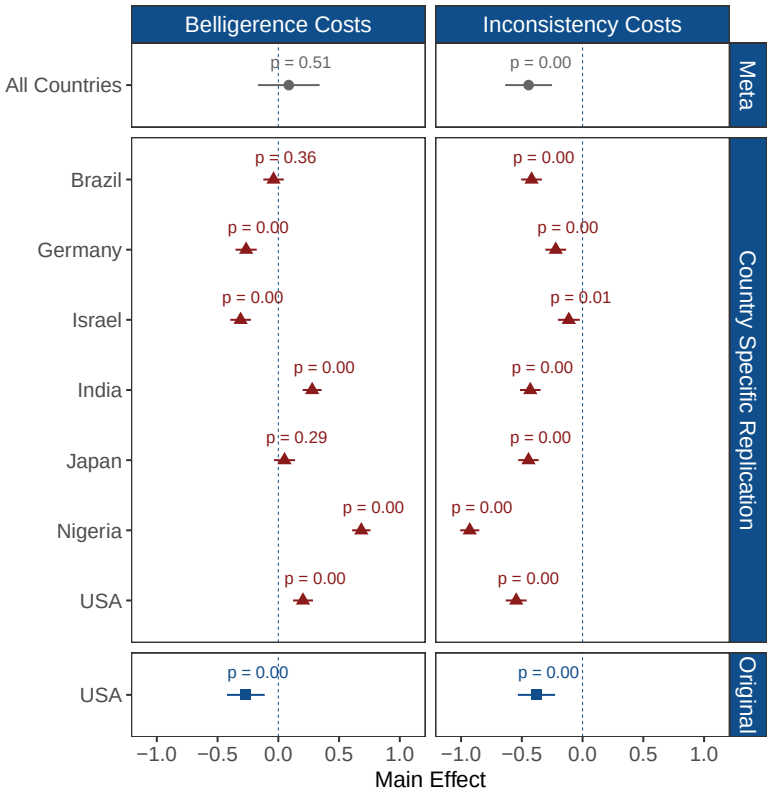

Figure A15: **Audience Costs extension.** We report the original estimates and p-values of the belligerence and inconsistency costs, as calculated in the original study. We further report the country-specific ATEs (and BH-adjusted p-values) from our replications, and a meta-analytic average treatment effect based on our harmonized studies.

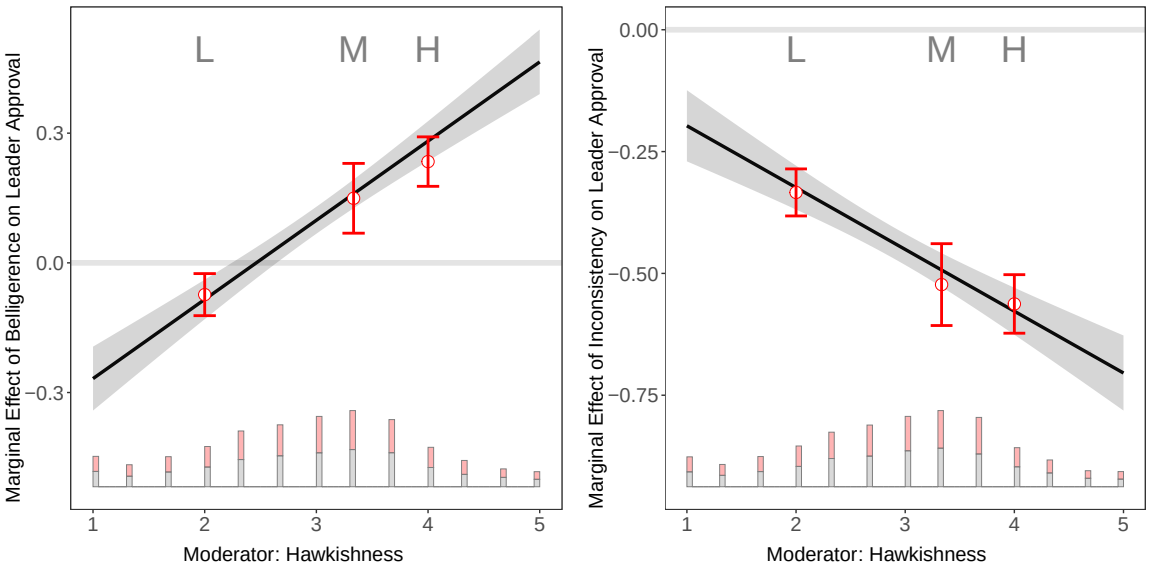

Figure A16: **Moderating Effect of Hawkishness Index in the extension of the Audience Costs Experiment.** This figure demonstrates the negative moderation of hawkishness on the inconsistency treatment effects, and the positive moderation of hawkishness on belligerence treatment effect. These results are consistent with findings from the original study.

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 A19).

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 A12, 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 A20). 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

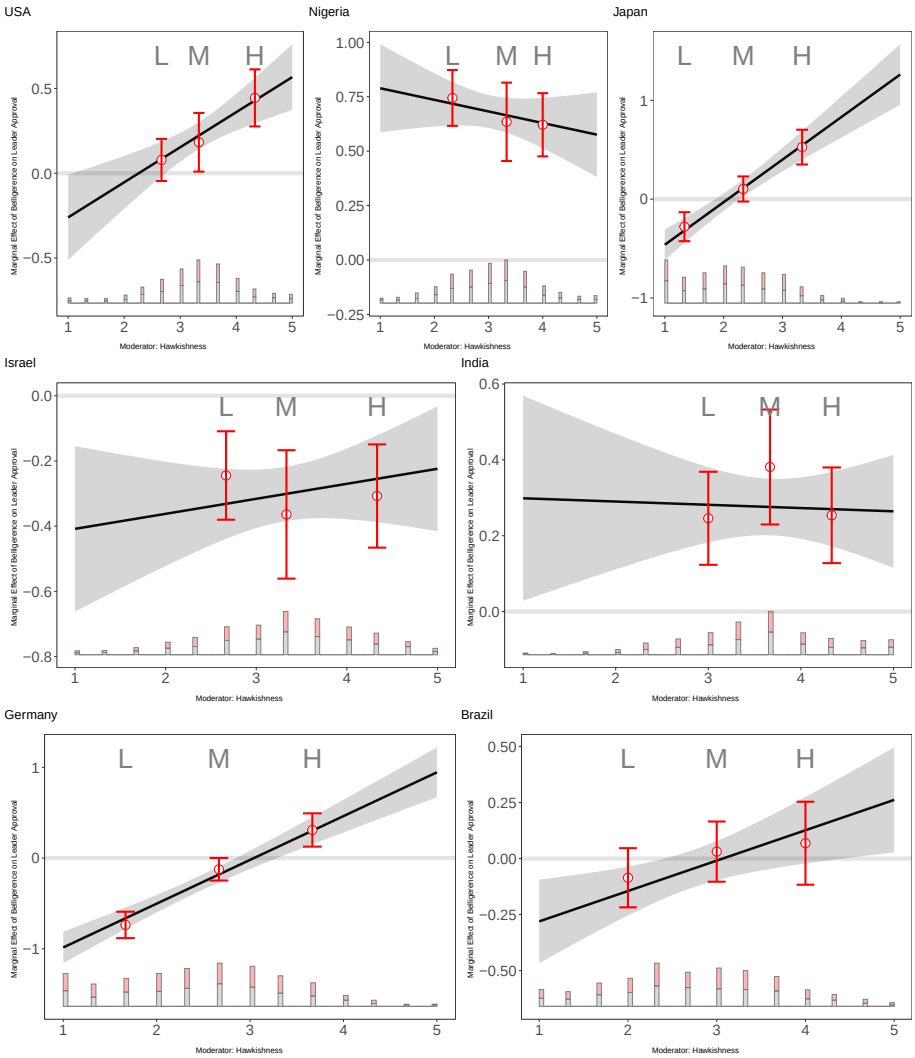

Figure A17: **Belligerence Costs Moderating effects in country-samples.** Individual figures of the moderating effects of hawkishness in the audience costs extension (belligerence costs). Figures are broken down by country-samples.

	Support attack
	Model 1
Democracy	−0.029
	(0.073)
Adj. R^2	−0.001
Num. obs.	824

$^{*}p < 0.05$.

Table A20: Screening out failed manipulation from other studies (India DP)

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 Robustness Checks

In this section we report additional robustness checks. First, we report in Figure A18 estimates and standard errors of models where the following pre-treatment covariates have been added as controls: Gender, Age, Ideology, Education, Voting, Democratic norms, Hawkishness, Legal obligation. Our results are largely robust to these model specifications. Next, we examine the role of the language in which respondents took the survey. If respondents were to overwhelmingly take the survey in a language that is not the country's main official language, this could be indicative of the sample's representativeness of the broader population within the country. In Figure A19 we report the proportion of respondents who used each language per country sample. As we demonstrate, the majority of respondents took the surveys in the national/local languages.

One exception is India, where a larger proportion of respondents (around 60%) took the survey in English. We note that this is somewhat expected, as English is an official language in India. Nonetheless, we examined whether our treatment effects in India were more pronounced for respondents who took the survey in English. Table A21 reports the main treatment effects, conditional on the survey language. We do not identify any statistically significant heterogeneous effects here.

M 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.

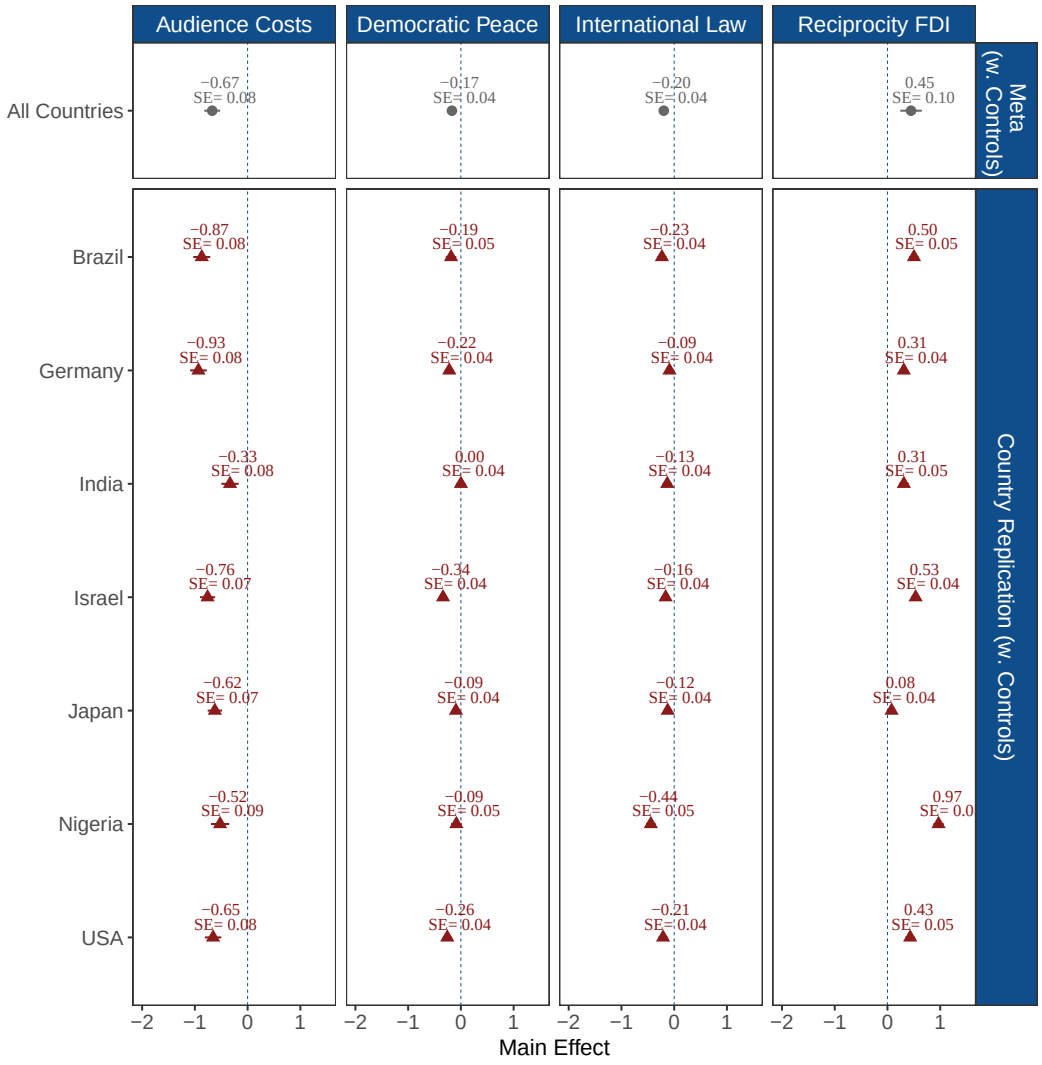

Figure A18: **Main analysis with demographic controls.** We report the estimates and standard errors of models where the following covariates have been added: Gender, Age, Ideology, Education, Voting, Democratic norms, Hawkishness, Legal obligation.

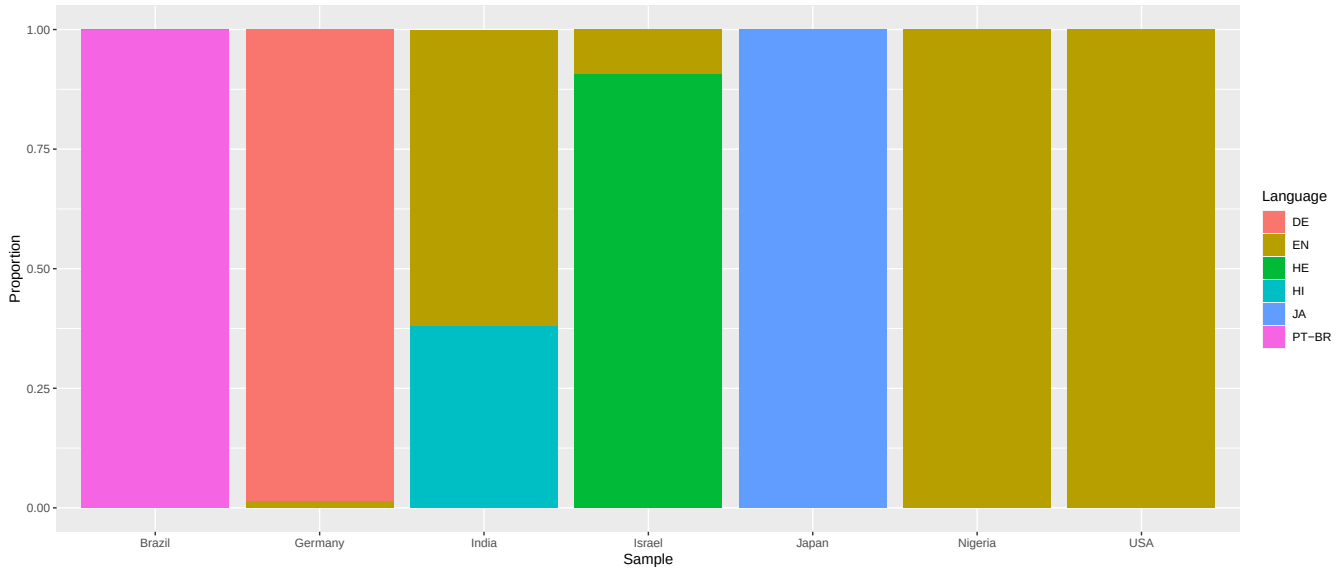

Figure A19: **Proportion of Use Languages per country sample.** Note that the majority of respondents took the surveys in the national/local languages.

	Dem Peace	Audience Costs	Int Law	Reciprocity
Intl Law			0.001	
			(0.116)	
English	−0.147	−0.007	−0.330*	−0.010
	(0.087)	(0.113)	(0.086)	(0.089)
IL*English			−0.080	
			(0.122)	
Back Down		−0.107		
		(0.150)		
BD*English		−0.050		
		(0.158)		
Democracy	−0.049			
	(0.117)			
Dem*English	0.044			
	(0.123)			
Harder barrier				0.284*
				(0.117)
Hard*English				−0.081
				(0.122)
Adj. R^2	0.000	0.004	0.012	0.011
Num. obs.	3077	2021	3072	3074

$^*p < 0.05$

Table A21: Treatment Effect*English in India Sample

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.

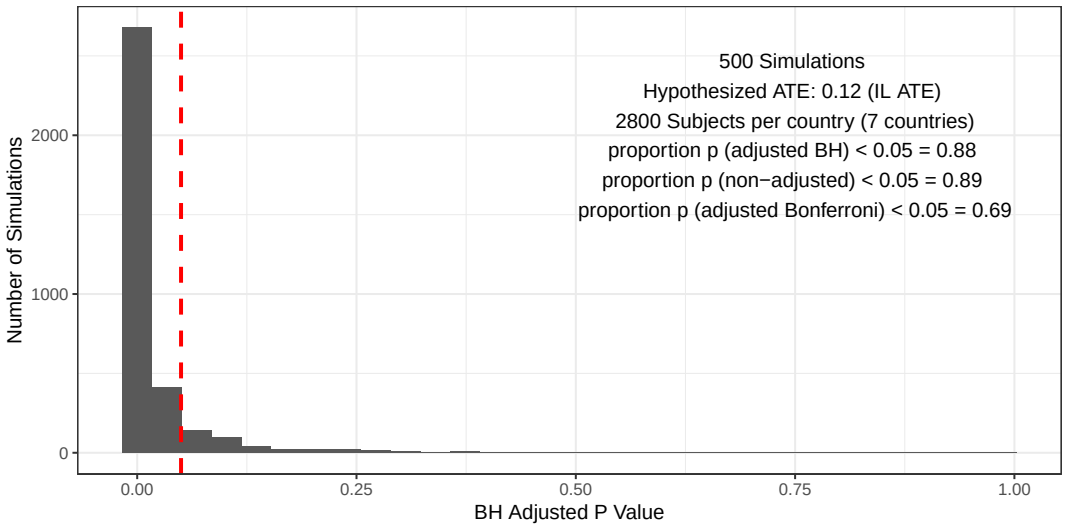

Figure A20: **Prospective power analysis.** Determining suitable N size to identify original point estimates within each country sample.

N Power analysis and pre-analysis plan

Our prospective power analysis (see figure A20), determined our sample size by ensuring that we are well powered ($> 80\%$) to identify original point estimates within each country ($\alpha = 0.05$) in the meta-analysis. In a retrospective power analysis (see figure A21), we show that we are also extremely powered ($> 99\%$) for our sign generalizability test.

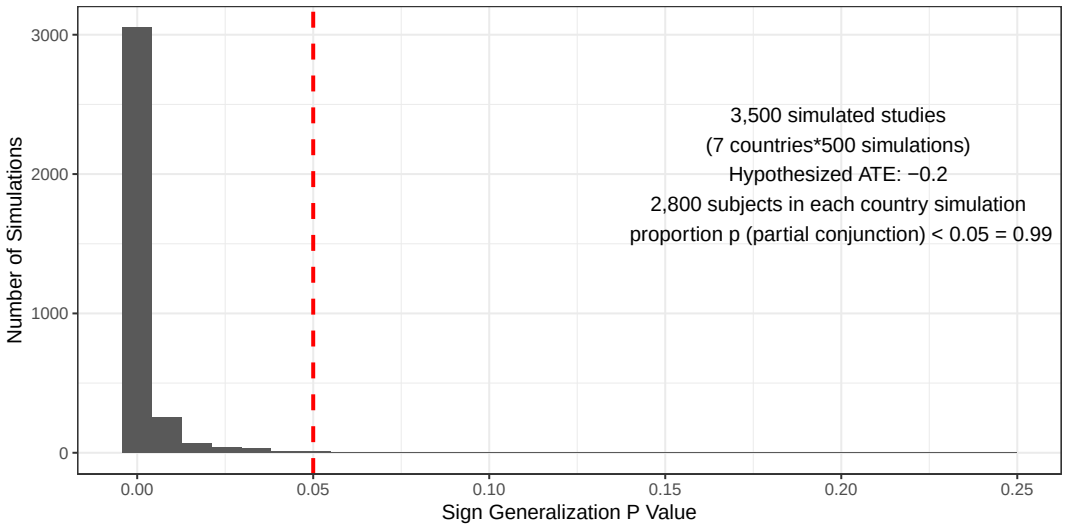

Figure A21: **Retrospective power analysis.** Ensuring we are powered for our sign generalizability test.

CONFIDENTIAL - FOR PEER-REVIEW ONLY

Generalizability of IR Experiments (#119402)

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

This is an anonymized copy (without author names) of the pre-registration. It was created by the author(s) to use during peer-review.
A non-anonymized version (containing author names) should be made available by the authors when the work it supports is made public.

---

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.

ASPREDICTED

Wharton CREDIBILITY LAB
University of Pennsylvania

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).

Version of AsPredicted Questions: 2.00

Available at https://aspredicted.org/6QK_X59