*American Political Science Review* (2024) 1–16
doi:10.1017/S0003055424001199

# The Generalizability of IR Experiments beyond the United States

LOTEM BASSAN-NYGATE *Harvard University, United States*
JONATHAN RENSHON *University of Wisconsin–Madison, United States*
JESSICA L. P. WEEKS *University of Wisconsin–Madison, United States*
CHAGAI M. WEISS *Stanford University, United States*

*Theories of international relations (IR) typically make predictions intended to hold across many countries, yet existing experimental evidence testing their micro-foundations relies overwhelmingly on studies fielded in the United States. We argue that the broad nature of many IR theories makes it especially important to evaluate the extent to which their predictions hold across countries. To examine the generalizability of IR experimental findings beyond the US, we implemented a preregistered and harmonized multisite replication study, fielding four prominent IR experiments across a diverse set of seven democracies: Brazil, Germany, India, Israel, Japan, Nigeria, and the US. We find high levels of generalizability across all four experiments, a pattern further analysis suggests is due to limited treatment effect heterogeneity. Our findings and approach offer important empirical and methodological insights for the design and interpretation of future experimental research in IR.*

## INTRODUCTION

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 2011b), 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]

At the same time, 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 2023). Recent work has provided theoretical foundations for these concepts (Egami and Hartman 2023; Findley, Kikuta, and Denly 2021; Humphreys and Scacco 2020; Slough and Tyson 2023) and empirically probed questions such as whether experimental findings hold across diverse country contexts (Coppock and Green 2015; Dunning et al. 2019a). Scholars of Comparative Politics have engaged in multisite replications (Dunning et al. 2019b), and recent research in American Politics has combined large-scale replication projects with meta-analyses (Blair, Coppock, and Moor 2020; Coppock, Hill, and Vavreck 2020; Schwarz and Coppock 2022).

The field of 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 (Renshon, Yarhi-Milo, and Kertzer 2023; Suong, Desposato, and Gartzke 2020; Tomz and Weeks 2013), often introducing design changes across countries and providing limited motivation for case selection. Existing multisite 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 findings—in this case, from a series of prominent IR papers—"apply to other sets of individuals, to other types of interventions, and in other contexts" (Blair and McClendon 2021, 411). We focus on a form of generalizability known as "*C*-validity" (Egami and Hartman 2023), which captures whether findings

Lotem Bassan-Nygate, Assistant Professor, John F. Kennedy School of Government, Harvard University, United States, lbassan@hks.harvard.edu

Corresponding author: Jonathan Renshon, Professor, Department of Political Science, University of Wisconsin–Madison, United States, renshon@wisc.edu

Jessica L. P. Weeks, Professor, Department of Political Science, University of Wisconsin–Madison, United States, jweeks@wisc.edu

Chagai M. Weiss, Postdoctoral Fellow, Conflict and Polarization Initiative and Polarization and Social Change Lab, Stanford University, United States, cmweiss@stanford.edu

Received: June 10, 2023; revised: March 06, 2024; accepted: September 19, 2024.

[1] For example, Keele, McConnaughy, and White (2012), Clifford and Jerit (2015), Dafoe, Zhang, and Caughey (2018), Offer-Westort, Coppock, and Green (2021), Blair, Coppock, and Moor (2020), Clifford, Sheagley, and Piston (2021), Mutz (2021), Chaudoin, Gaines, and Livny (2021), and Brutger et al. (2022; 2023).

extend to contexts ("*C*") in which theories have not yet been tested. More specifically, we examine the extent to which replications of IR experiments in new contexts produce statistically significant effects in the same direction as the original results.[2] Two notable aspects of our work are thus the focus on direction of effects (rather than magnitude) and our view of generalizability as a continuum rather than a binary property of a given finding.

Assessing the generalizability of IR findings is crucial for remedying the mismatch between the predictive scope of IR theories and the breadth of their underlying evidence. Although the broad predictions of IR theories 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 2010b). It is thus difficult to judge whether IR theories are truly *international*, or merely explain the foreign policy preferences of Americans. Assessing the generalizability of experimental results across countries also holds important implications for equity in the profession, including whether findings from sites outside the US, which may be more accessible for non-U.S.-based researchers, yield generalizable results.

To explore these issues, we implemented a preregistered and harmonized multisite replication study designed to sidestep challenges such as publication bias (i.e., selective reporting of positive results) and study comparability (Slough and Tyson 2023). 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 foreign direct investment* (Chilton, Milner, and Tingley 2020)—in a set of seven democracies (Brazil, Germany, India, Israel, Japan, Nigeria, and the US), which we selected using a strategy of "purposive variation" (Egami and Hartman 2023). 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 the plausibility 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. Our top-line findings indicate stability of treatment effects across experiments, country contexts and demographic profiles of respondents. Though we cannot say whether we would find such consistent results across all IR theories or countries of interest, our harmonized replications of important and well-known experiments from different substantive domains across a set of purposively varied countries suggest reasons for optimism.

Second, our findings indicate that the US is not an outlier in terms of experimental evidence on the microfoundations of general IR theories—nor are any of the countries we studied. Americans are different from other populations in many ways, but our results suggest that such differences do not dramatically shape experimental findings across countries for common IR theories. Rather, in line with recent studies in American politics (Coppock 2023), the theories we tested appear to exhibit low treatment effect heterogeneity (Coppock 2019): samples with considerable variation along a number of covariates responded similarly to our treatments.

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 US or elsewhere. This conclusion has important practical and normative implications, reducing barriers to entry for non-U.S.-based scholars and for correcting the impression that the US ought to be the default site for experimental research. However, our findings also emphasize the importance of theorizing *ex ante* about variables that could moderate treatment effects, incorporating measures of these moderators at the design stage, and probing whether treatment effects are heterogeneous within a given sample. Homogeneous treatment effects should increase confidence in cross-country generalizability. However, heterogeneous treatment effects, particularly changes in sign (rather than merely magnitude), should spur scholars to consider how samples in other contexts might differ and suggest caution in making more general claims.

At the same time, our study demonstrates the value of harmonized multisite replication studies when such efforts are possible and research programs are mature enough to warrant it. Future efforts may rely on our approach, which brings together an innovative suite of tools for choosing sites, analyzing experimental data—building on Egami and Hartman's (2023) framework of "purposive variation" and sign-generalization—and designing research to probe theoretically relevant moderators and investigate possible null results. While our findings provide reassuring insights regarding the generalizability of IR experimental research, they also allow us to identify an important context in which one of our experiments does not replicate and 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 preregistered harmonized multisite replication studies as an important component in the IR research cycle in which researchers establish the generalizability and scope of single-country findings.

[2] We sidestep the issue of the magnitude of effects, which is typically more relevant for theories about particular policy interventions (Egami and Hartman 2023 1080–6).

https://doi.org/10.1017/S0003055424001199 Published online by Cambridge University Press

## DEFINING EXTERNAL VALIDITY AND GENERALIZABILITY

Political scientists 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 2011a, 28), and is a quality specific to a particular study (McDermott 2011a, 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 repeated replications reveal the extent to which conclusions generalize (McDermott 2011a). 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 (Bisbee and Larson 2017; Hainmueller, Hall, and Snyder 2015; Kertzer 2022).

We define external validity, that is, generalizability, as the extent to which existing findings "apply to other sets of individuals, to other types of interventions, and in other contexts" (Blair and McClendon 2021, 411). More specifically, we build on Egami and Hartman (2023), who decompose external validity into four components, $X-$, $T-$, $Y-$, and $C-$validity, referring to populations, treatments, outcomes, and contexts/settings, respectively. We aim to assess $C-$validity: the extent to which experimental findings generalize from one context to others where no data currently exist (Egami and Hartman 2023). Our specific focus is on cross-country variation in contexts (as opposed to cities, counties, regions, or other geographic units).

We consider a particular finding *generalizable* to the degree that the sign of the effect generalizes across more country contexts that fall within the bounds of a theory's scope conditions. Our conception of generalizability emphasizes its continuous nature: findings are "more versus less generalizable" rather than "generalizable or not." Our focus on direction and significance (rather than magnitude of effect) is motivated by Egami and Hartman (2023, 1086), who recommend generalizability tests of direction/sign for synthesizing scientific findings, 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 feature predictions about effect sizes either implicitly or explicitly. 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 more generalizable to the extent that we find consistent experimental support for it across an array of democratic countries.

## GENERALIZABILITY IN IR

Foundational IR theories were usually intended to provide broad insights about interstate relations across a wide range of countries (see, e.g., Wolfers 1947, 26). Likewise, contemporary IR theories seek to explain international politics in "*general* causal terms" (emphasis added; Walt 2005, 26). In both "grand" frameworks such as realism and "middle-range" theories such as the democratic peace, IR scholars typically portray their theories as providing general explanations of interstate relations rather than insights into one specific country or region. For example, theories of reputation (Downs and Jones 2002; Wolford 2007) and resolve (Kertzer 2016) make general predictions about states, leaders, and perceptions, not restricted to any one state, leader, or specific empirical context. If a theory applied to 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. Many scholars have suggested, however, that IR research has tended to focus on the US (Colgan 2019a; Hoffmann 1977; Kristensen 2015; Levin and Trager 2019).[3] Per Hendrix and Vreede (2019, 311), the US "is not the eight-hundred-pound gorilla in the literature, but the three-hundred-thousand-pound blue whale."

To assess whether the microlevel experimental IR literature is similarly U.S.-centric, we conducted a quantitative literature review identifying all IR articles containing experimental studies (a total of $N = 216$ articles and $n = 369$ studies) published in the top political science journals (*APSR, AJPS, JOP*) and IR subfield journals (*IO, ISQ, JCR*) over the past two decades (2000–21). Figure 1 provides a heat map of these studies by country site (location). Strikingly, nearly 60% of the experiments utilized U.S. subjects. Moreover, the US was eight times more popular than the next most common site, Israel.[4] Evidently, experimental research on the micro-foundations of prominent IR theories relies predominantly on studies of U.S. foreign policy attitudes, behaviors, and perceptions.

Scholars could reasonably worry that conducting microlevel empirical tests of IR theories nearly exclusively on U.S.-based samples would provide little insight into broader empirical relationships. The US is wealthier, has longer-standing democratic institutions,[5] is more geographically protected, more conflict-prone, and

[3] U.S. centrism is prevalent not only in terms of empirical focus 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), and even the content of prominent cross-national datasets (Colgan 2019a). See also Aronow and Samii (2016) on problems of generalizability even when samples are representative.

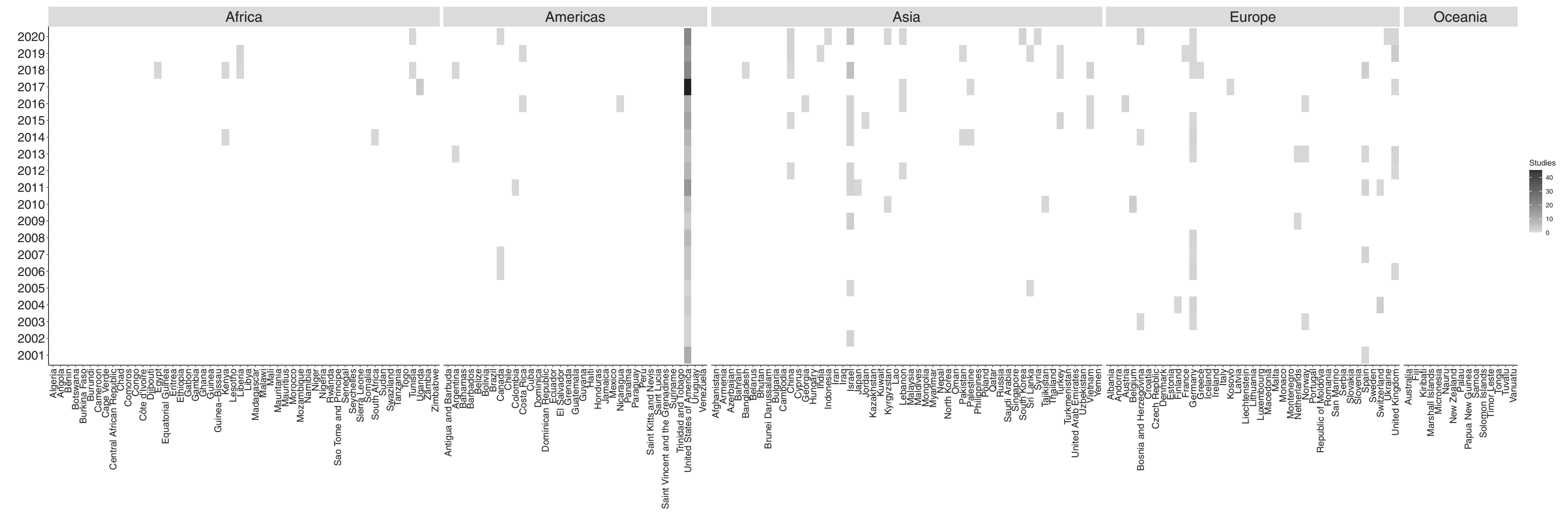
[4] This finding echoes Hendrix and Vreede (2019, 311), who point out that Israel and the US receive scholarly attention far out of proportion to their population, GDP, etc.

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

https://doi.org/10.1017/S0003055424001199 Published online by Cambridge University Press

https://doi.org/10.1017/S0003055424001199 Published online by Cambridge University Press

## FIGURE 1. Heatmap of IR Experiments Published in Six Top Journals between 2001 and 2020 by Country



*Note*: As is demonstrated by the single vertical line signifying U.S.-based studies, IR survey experiments are predominantly conducted in the US. Unit of analysis is study (*N*=369) rather than article.

more powerful and authoritative than most other countries. To the extent that such country-level factors affect ideologies, perceptions, or judgments, experimental findings from U.S.-based subjects might shed little light on whether particular theories apply to populations in other places. U.S. subjects might also be unusual at the individual level: Americans tend to be less knowledgeable than peers in other locations (Dimock and Popkin 1997; Levin and Trager 2019), and the US stands out demographically even from other large, powerful countries (Brooks et al. 2018), including its "psychologically unusual" WEIRD peers (Henrich, Heine, and Norenzayan 2010b; Jones 2010b, 29; see also Henrich, Heine, and Norenzayan, 2010a; Jones 2010).

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": that is, for many studies, treatment effects do not differ much across subgroups. One implication of this finding is that effects from IR experiments might not differ much across national contexts, either. In IR, at least some results have been found to be 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 US and the United Kingdom generalizes to Brazil; and Tomz, Weeks, and Bansak (2023) find that formal military alliances have robust causal effects across 13 North Atlantic Treaty Organization (NATO) countries. However, without systematic harmonized research assessing the generalizability of prominent IR theories, it is impossible to say whether the US focus of existing IR experiments represents an acceptable base for broader knowledge or an empirical crisis.

## RESEARCH DESIGN

### Overview

The conception of generalizability developed above informs the design of our harmonized multisite replications. We note four key features of our design. First, our study is specifically designed to probe generalizability across multiple studies and contexts, with clear criteria established *ex ante* for assessing findings. 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 two alternative sites to explore whether an effect identified in an initial context replicates there (e.g., Lupu and Wallace 2019; Tomz and Weeks 2013). In contrast, we focus on two broader 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 2023). It has the advantage of being both principled and empirically verifiable, while lending itself directly to the two analytical methods (sign-generalization tests and meta-analysis) that enable us to answer the questions outlined above. It also allows us to make inferences about countries outside of the sample, subject to certain assumptions discussed below.

Third, our design is harmonized, reducing the possibility that idiosyncrasies in timing, logistics, or design variations could render studies incomparable (Slough and Tyson 2023). We sought harmony in terms of treatments and 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 preregistered our study to reduce the risk of selective reporting, which could be 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 US, employ relatively simple designs, were found to produce robust effects in the US, and cross substantive boundaries within IR. This approach led us to include experiments on the democratic peace (Tomz and Weeks 2013), audience costs (Kertzer and Brutger 2016; Tomz 2007), international law (Wallace 2013), and reciprocity in foreign direct investment (Chilton, Milner, and Tingley 2020). More information on the four studies is provided in Section B of the Supplementary Material, 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.

### 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

https://doi.org/10.1017/S0003055424001199 Published online by Cambridge University Press

**TABLE 1. Theoretical Components of Our Studies**

| | 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 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 (+) |
| *Reciprocity FDI* | A foreign country's FDI policy | Concern for fairness | Support for FDI policy | NA |

*Note:* Sign in parentheses indicates the direction of the moderating effect.

process (Egami and Hartman 2023) we use to select seven country sites.

Approaches to case selection can generally be characterized as either random or nonrandom. Random approaches have obvious benefits but would provide little leverage here as sample of seven countries does not permit strong inferences about a broader population of interest (i.e., all countries within the scope of a theory). On the other hand, nonrandom 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 US 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.[6]

We opt for a different nonrandom approach, using "purposive variation" to select sites that ensure variation 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 2023, 11–2). 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.

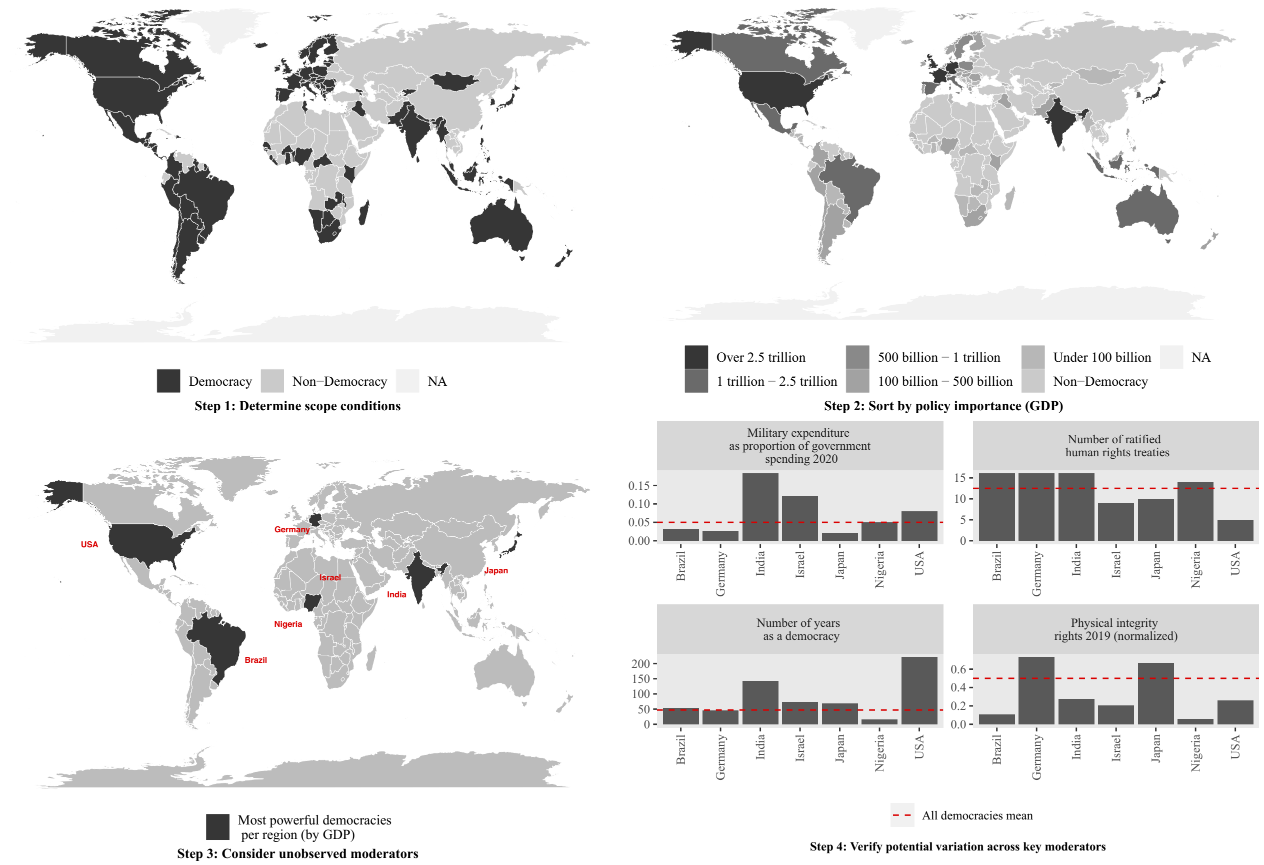
For three out of four studies, we identified theoretically relevant moderators—strength of democratic norms in *democratic peace*; hawkishness in *audience costs*; and international legal obligation in *international law*.[7] Table 1 summarizes the theoretical components of all four studies and specifies the expected direction of the moderating effect.

After parsing the theories, our country selection proceeded systematically through the process depicted in Figure 2 (detail in Section C of the Supplementary Material). 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

[6] 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).

[7] 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.

FIGURE 2. Steps in Selecting Countries for Replication



Step 1: Determine scope conditions

Step 2: Sort by policy importance (GDP)

Step 3: Consider unobserved moderators

Step 4: Verify potential variation across key moderators

*Note*: GDP data are from the World Bank (2022). Military expenditure data are from Stockholm International Peace Research Institute (SIPRI N.d.). Physical integrity rights data are from Fariss (2019).

likely plays a larger role in democracies, we focus on countries above a minimum threshold of democracy (Polity $\geq 6$). 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 Gross Domestic Product (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 latent) by selecting one country from each major region of the world.[8] 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).[9] 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 US).

## 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.

[8] 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.

[9] While theoretical moderators are often at the individual level, our site selection process used country-level proxies in deference to data availability.

https://doi.org/10.1017/S0003055424001199 Published online by Cambridge University Press

### *Sign-Generalization Test*

To assess the extent to which the direction of causal effects is generalizable, we use the sign-generalization procedure proposed by Egami and Hartman (2023). 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 Average Treatment Effects (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 (2023, 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-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 not empirically verifiable).

### *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 are 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 Ordinary Least Squares (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]

### *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 A8, Dataverse-only Appendix). Because this is a more demanding standard, our estimates ensured that our other empirical test (sign-generalization) was extremely well-powered ($> 90\%$, see Figure A9, Dataverse-only Appendix). Power is particularly important in the case of generalizability studies, as low power can lead to spurious estimates of "high generalizability" (Coppock 2019, 8). While our visualizations below can sometimes draw attention to differences in magnitudes of effects across various country or study combinations, we are not powered to detect such 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.

[10] This is subtly different from tallying up the number of significant ATEs 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 (2023, 1082).

[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 seven 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).

https://doi.org/10.1017/S0003055424001199 Published online by Cambridge University Press

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 preregistered secondary analyses related to attentiveness, respondents having a particular country in mind, the plausibility of the scenarios, and effect heterogeneity along theoretically relevant moderators.

## 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 three thousand 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 pretreatment measures of social and political dispositions. Attentive respondents proceeded to our four experiments, shown in randomized order. Section A of the Dataverse-only Appendix details our survey instruments whereas Section E of the Supplementary Material reports descriptive statistics of each sample.

### Strong Support for Sign-Generalization among IR Experiments

Figure 3 displays results from sign-generalization tests for each of the four experiments to evaluate in how many and which countries the sign of the result matches theoretical expectations. As indicated by the flags and associated $p$-values, *audience costs* (top-left panel) yields a high level of sign generalizability with $p$-values $< 0.05$ across all seven countries. We obtain similar findings for *reciprocity FDI* (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 are around $p = 0.05$.We thus construe the pattern of results for *international law* to imply relatively high levels of sign-generalizability across countries.

Turning to *democratic peace* in the upper-right panel of Figure 3, we find broad 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 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 below.

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 2023 and also below) is plausible.

### Strong Underlying Support for Generalizability Using Meta-Analysis

Figure 4 displays the meta-analyses for all four experiments, assessing the underlying support for each theory across the pooled sample of countries. The top panel displays the meta-analytic average treatment effect for each experiment. These are based on the country-specific average treatment effects, shown in the middle 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 US) 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 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 US—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 US are identical to those one would draw from experiments implemented in a diverse set of countries with varying institutional, cultural, and economic characteristics. Notably, the directional congruence between original

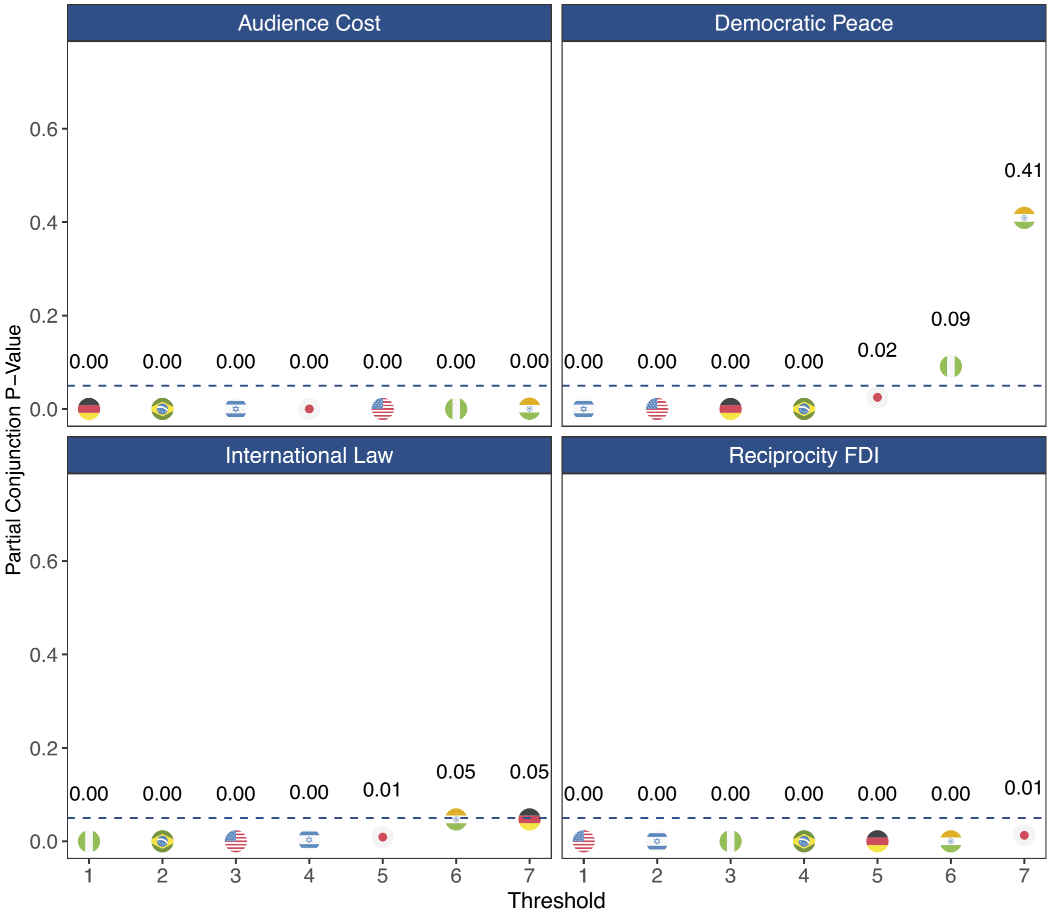
[13] Cint acquired Lucid in 2021.

[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 US and Nigeria. The Supplementary Material 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 Section D of the Supplementary Material.

[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.

https://doi.org/10.1017/S0003055424001199 Published online by Cambridge University Press

**FIGURE 3. Sign-Generalization Test**



*Note*: 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. Results are illustrated in Table A2 in the Supplementary Material.

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 28 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 (and the relevant theories lack clear predictions about effect sizes), readers might be interested in what our results suggest on that dimension. For *audience costs* and *democratic peace*, our meta-analytic ATEs are around half the size of the effects estimated in the original studies, while in *international law* the ATEs were similar, and in *reciprocity FDI*, our meta-analytic ATE appears to be about two-thirds as large as the originally estimated ATE. Future research might further investigate these potential patterns.

Together, the results in Figures 3 and 4 suggest optimism regarding the generalizability of IR experiments using two different approaches. We now turn to preregistered analyses designed to interrogate the one clear instance in which a study failed to replicate.

## Probing the Null: Why No Democratic Peace in India?

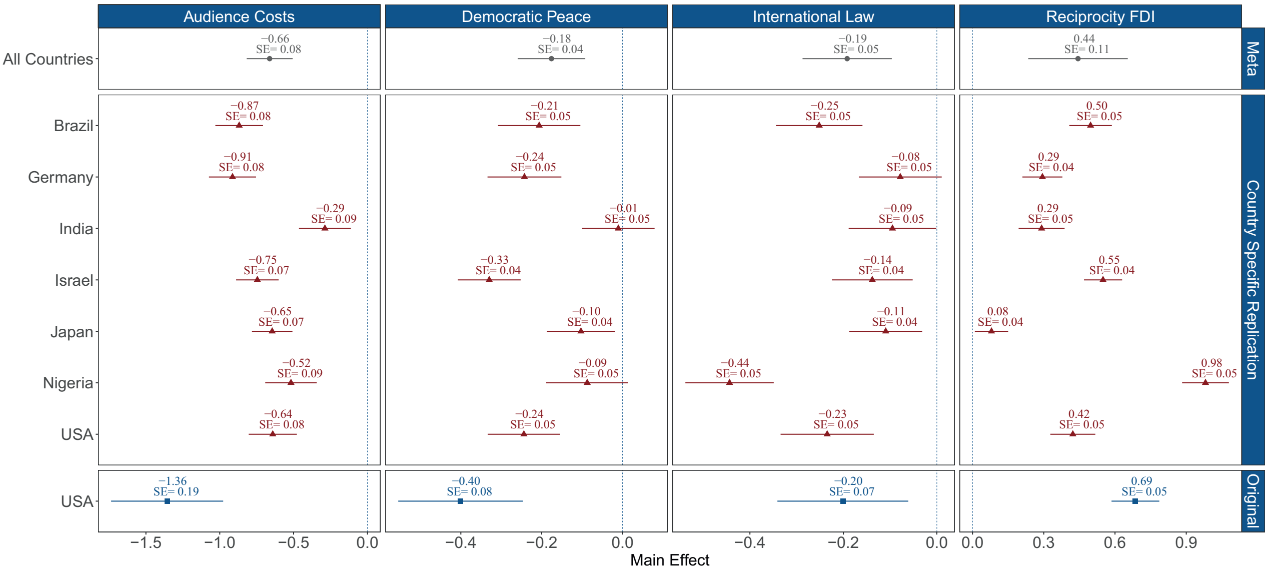
The one clear exception to the pattern of generalizable results we found was the *democratic peace* study in India, where the effect of democracy on support for an attack yielded a null finding ($\beta = -0.01$, $p = 0.818$, $CI = [-0.1, 0.08]$). Fortunately, we anticipated potential null results and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

Section I of the Supplementary Material details 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 prespecified moderators, three pieces of evidence suggest that weak democratic norms in India might help explain at least part of the null effect:

https://doi.org/10.1017/S0003055424001199 Published online by Cambridge University Press

**FIGURE 4. Meta-Analysis**



*Note*: 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. For results in table form, see Table A1 in the Supplementary Material.

1. Support for democratic norms significantly attenuates the effect of the democracy treatment across our full sample of all countries (see Figure A2 and Table A12 in the Supplementary Material, 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, see also Figure A1 in the Supplementary Material).
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 A13 and Figure A5 in the Supplementary Material; the interaction between norms and the democracy treatment in India is estimated at $\beta = -0.17, p = 0.058$).

We speculate that the remaining answer involves historical dynamics surrounding 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 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.

## Explaining Generalizability: Limited Treatment Effect Heterogeneity

What explains the strong degree of correspondence between estimates from the US and other countries? Below, we describe exploratory analyses designed to adjudicate between two possibilities:

1. *Similar sample characteristics* (i.e., low variation in the composition of the samples across countries).
2. *Limited treatment effect heterogeneity* (i.e., different individuals respond to a treatment in similar ways).

Overall, exploratory analyses indicate that (2) is considerably more plausible than (1). We find substantial variation in the composition of our samples across countries but little evidence of treatment effect heterogeneity. Although we cannot definitively provide evidence for the obverse (treatment effect homogeneity), we conclude it is a plausible explanation for the overall correspondence in results that we observe. We provide evidence for limited heterogeneity in three ways: by exploiting variation in our preregistered theoretical moderators, by utilizing a test of systematic heterogeneity proposed by Ding, Feller, and Miratrix (2019), and by contrasting results for our main studies with an extension of our *audience costs* study that was designed to have high levels of treatment effect heterogeneity.

### *Similar Sample Characteristics*

One possible explanation for our consistent results involves characteristics of the samples we collected.

https://doi.org/10.1017/S0003055424001199 Published online by Cambridge University Press

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 A1 in the Supplementary Material displays distributions of key covariates and demonstrates a meaningful degree of cross-country variation along hawkishness, international legal obligation, and support for democratic norms.[16] In Table A16 in the Supplementary Material, 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 of the US) then inadvertent cross-country similarity in samples is unlikely to explain our main pattern of results. Since 33/36 of these estimates are significantly different from the US, we conclude that our country samples do indeed vary along 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.

### *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: findings from one set of respondents can be generalized to other populations because even very different people react 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 thought to hold—that is, 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 in *democratic peace*; hawkishness in *audience costs*; and perceptions of international legal obligations in *international law*). Table A12 in the Supplementary Material reports our results for each experiment when pooling across country samples, demonstrating that: (1) there are statistically significant moderating effects in *democratic peace* and *international law* (but not *audience costs*) but that (2) even in those cases, the moderators never change the direction of the ATE. Instead they merely attenuate or amplify the treatment effect (Figures A2–A4 in the Supplementary Material).

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: for example, in *international law*, 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 (the same is true of *democratic peace*; see Figure A5 and Tables A13–A15 in the Supplementary Material). In *audience costs*, "hawkishness" significantly interacts with the treatment in only two out of seven countries and in only one of those countries (Germany) is the direction of the moderation counter to the theory's predictions (even there, the sign of the ATE does not flip except at the most extreme possible value of hawkishness). 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 Section G of the Supplementary Material. One reason that these various analyses are only suggestive, however, is that we may have limited power to examine variation within countries. Overall, the results suggest that the moderation effects are substantively somewhat small and do not shape the direction of the ATEs.

We complement this exploratory analysis with a formal procedure proposed by Ding, Feller, and Miratrix (2019), which tests the null hypothesis that a treatment effect is constant across all units, allowing us to estimate the presence of significant systematic variation within each country-study pair. Formally, the test leverages a Fisher Randomization Test (requiring minimal assumptions) 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 systematic variation in treatment effects. Out of 28 country-study pairs in the main preregistered analyses (above the horizontal line), only nine show evidence of systematic heterogeneity. Thus, in the majority of country-study pairs, we cannot reject the null of homogeneity. And, indeed, this accounting might overstate meaningful heterogeneity since the test does not distinguish between heterogeneity that shifts magnitudes of treatment effects and heterogeneity that flips the direction of effects for certain subgroups.

Though one might be tempted to declare this a case of "low heterogeneity," 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.[17]

[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 A16 in the Supplementary Material.

[17] Nevertheless, our ad hoc simulation study presented in Figure A6 in the Supplementary Material shows that given our large sample size, we should be well-powered to detect treatment effect heterogeneity on the scale of 0.15 SD.

https://doi.org/10.1017/S0003055424001199 Published online by Cambridge University Press

**TABLE 2. Results for Tests of Systematic Treatment Effect Heterogeneity Developed by Ding, Feller, and Miratrix (2019)**

| Study | *N* comparisons | *N* significant | *N* significant (BH adjustment) |
|---|---|---|---|
| *Audience Costs* | 7 | 3 | 2 |
| *Democratic Peace* | 7 | 3 | 0 |
| *International Law* | 7 | 3 | 3 |
| *Reciprocity FDI* | 7 | 4 | 4 |
| *Audience Costs Extension* | 7 | 7 | 7 |

*Note*: N Comparisons is 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 (third column) after adjusting for multiple comparisons. The top four rows denote our main studies, whereas the last row refers to *AC Extension*.

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 analyze an extension to our *audience costs* experiment. Recall that the preregistered *audience costs* 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 hawkishness moderator did not flip the sign of the ATE. We also fielded an extension to the main study based on Kertzer and Brutger (2016) that 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 *audience costs* to the Kertzer and Brutger (2016) version (*AC extension*), we can compare studies predicted to have varying levels of treatment effect heterogeneity. The results in *audience costs* compared the "back down" to "stay out" conditions, but in *AC extension* (described in Section I of the Supplementary Material), respondents were assigned to three experimental conditions, allowing us to decompose the general audience cost into belligerence and inconsistency costs.

Section I of the Supplementary Material shows that, consistent with our expectation of differences across groups and contexts, the belligerence treatment (the effect of "engaging" versus "staying out") yields null effects in two countries, negative effects in two countries, positive effects in three countries, and an overall null meta-analytic ATE. In Figure A3 in the Supplementary Material, we further shows that hawkishness not only moderates belligerence costs in *AC 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 A10 in the Supplementary Material). Furthermore, the treatment effect heterogeneity test proposed by Ding, Feller, and Miratrix (2019) shows that there is systematic treatment effect heterogeneity in 7/7 of country-pairs (see bottom row of Table 2).[18] In sum, comparing our general pattern of results discussed above, where treatment effects are largely homogeneous, with results from *AC extension*, where treatment effects are heterogeneous, further suggests that the generalizability of our main findings may be driven by limited treatment effect heterogeneity.

## CONCLUSION

This article was motivated by concerns that the breadth of experimental evidence in IR 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 microfoundations 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 preregistered and harmonized multisite 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—five out of seven countries for *democratic peace* and *international law* and seven out of seven countries for *audience costs* and *reciprocity FDI*. Our meta-analysis revealed statistically 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—*democratic peace* 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

[18] See also Section F of the Supplementary Material for comparison of $I^2$ statistics, testing for heterogeneity in treatment effect between country-samples.

https://doi.org/10.1017/S0003055424001199 Published online by Cambridge University Press

countries without producing a single example of contradictory treatment effects.

Consistent with other replication studies (Coppock 2019; Coppock, Leeper, and Mullinix 2018), 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 US does not appear to be an outlier when it comes to experimental results on the micro-foundations of IR theories. American respondents differ from respondents in other countries in terms of key demographic attributes (Henrich, Heine, and Norenzayan 2010b), and may have atypical foreign policy preferences (see Figure A1 and Table A16 in the Supplementary Material), but their responses to treatment 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 IR.

These findings have striking implications for future research in both IR and political science more broadly. On the one hand, our findings underscore the value of preregistered and harmonized multisite replication studies in the potentially limited contexts 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 limits the potential for selective reporting and file drawer problems, which ultimately result in publication bias. By allocating significant resources and coordinating multiple simultaneous replication studies across various countries, we were able to learn how specific findings generalize, pinpoint one 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 2011a; Samii 2016).

However, our findings also highlight the perhaps surprising potential value of single-country studies for testing the micro-foundations of general IR theories, whether such studies are fielded in the US 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 US, 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 decenter the US 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 treatment effect 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. Scholars could further distinguish between Heterogeneous Treatment Effects (HTEs) in which covariates shift the magnitude of a treatment effect between subgroups versus HTEs in which the sign of the treatment effect flips. However, when treatment effects show relatively low heterogeneity—as we find in our study—bolder claims may be warranted.

## SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S0003055424001199.

## DATA AVAILABILITY STATEMENT

All research documentation and data that support the findings of this study are openly available at the American Political Science Review Dataverse: https://doi.org/10.7910/DVN/9UXYCQ.

## ACKNOWLEDGMENTS

We thank Priyadarshi Amar, Ashika Amar, Lauren Barden-Hair, Saloni Boghale, Ryan Brutger, Alexander Coppock, Jamie Druckman, Naoki Egami, Kristian Vrede Skaaning Frederiksen, Erin Hartman, Josh Kertzer, Masanori Kikuchi, Yena Kim, Dillon Laaker, Oliver Lang, Ned Littlefield, Adeline Lo, Mitsuru Mukaigawara, Anton Peez-Zvetina, Yehzee Ryoo, Victória Santos, Lauren Sukin, Dustin Tingley, Mike Tomz, Geoff Wallace, Anna Wilke, and participants in the UW-Madison IR colloquium, Brown University's Watson Institute Security Seminar, the University College London Comparative Political Economy & Behavior Workshop, and UC Berkeley MIRTH for helpful comments and suggestions. Studies described were exempted by the IRB at UW-Madison (#2022-0748)

https://doi.org/10.1017/S0003055424001199 Published online by Cambridge University Press

and preregistered with As.predicted: https://aspredicted.org/zt39f.pdf.

## FUNDING STATEMENT

This research was funded by the University of Wisconsin–Madison.

## CONFLICT OF INTEREST

The authors declare no ethical issues or conflicts of interest in this research.

## ETHICAL STANDARDS

The authors declare the human subjects research in this article was reviewed by the IRB at the University of Wisconsin–Madison (#2022–0748) and was determined to be exempt. The authors affirm that this article adheres to the principles concerning research with human participants laid out in APSA's Principles and Guidance on Human Subject Research (2020). Further information including application number and determination letter are available in the Dataverse-only Appendix.

## REFERENCES

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *The Quarterly Journal of Economics* 130 (3): 1117–65.

Aronow, Peter M., and Cyrus Samii. 2016. "Does Regression Produce Representative Estimates of Causal Effects?" *American Journal of Political Science* 60 (1): 250–67.

Bassan-Nygate, Lotem, Jonathan Renshon, Jessica L. P. Weeks, and Chagai M. Weiss. 2024. "Replication Data for: The Generalizability of IR Experiments beyond the United States." Harvard Dataverse. Dataset. https://doi.org/10.7910/DVN/9UXYCQ.

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): 502–21.

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): 1297–315.

Blair, Graeme, and Gwyneth McClendon. 2021. "Conducting Experiments in Multiple Contexts." In *Advances in Experimental Political Science*, eds. James N. Druckman and Donald P. Green, 411–28. Cambridge: Cambridge University Press.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2021. *Introduction to Meta-Analysis*. Hoboken, NJ: 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, Dustin Tingley, and Chagai M. Weiss. 2023. "Abstraction and Detail in Experimental Design." *American Journal of Political Science* 67 (4): 979–95.

Brutger, Ryan, Joshua D. Kertzer, Jonathan Renshon, and Chagai M. Weiss. 2022. *Abstraction in Experimental Design: Testing the Tradeoffs*. New York: Cambridge University Press.

Chaudoin, Stephen. 2014. "Promises or Policies? An Experimental Analysis of International Agreements and Audience Reactions." *International Organization* 68 (1): 235–56.

Chaudoin, Stephen, Brian J. Gaines, and Avital Livny. 2021. "Survey Design, Order Effects, and Causal Mediation Analysis." *The Journal of Politics* 83 (4): 1851–6.

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): 129–53.

Clifford, Scott, and Jennifer Jerit. 2015. "Do Attempts to Improve Respondent Attention Increase Social Desirability Bias?" *Public Opinion Quarterly* 79 (3): 790–802.

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): 1048–65.

Colgan, Jeff D. 2019a. "American Bias in Global Security Studies Data." *Journal of Global Security Studies* 4 (3): 358–71.

Colgan, Jeff D. 2019b. "American Perspectives and Blind Spots on World Politics." *Journal of Global Security Studies* 4 (3): 300–9.

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

Coppock, Alexander. 2023. *Persuasion in Parallel: How Information Changes Minds about Politics*. Chicago, IL: 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): 113–31.

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): 12441–6.

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

Dimock, Michael, and Samuel L. Popkin. 1997. "Political Knowledge in Comparative Perspective." In *Do the Media Govern*, eds. Shanto Iyenger and Richard Reevs, 217–24. London: Sage.

Ding, Peng, Avi Feller, and Luke Miratrix. 2019. "Decomposing Treatment Effect Variation." *Journal of the American Statistical Association* 114 (525): 304–17.

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 2019a. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*. Cambridge: Cambridge University Press.

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, Craig McIntosh, and Gareth Nellis 2019b. "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. 2023. "Elements of External Validity: Framework, Design, and Analysis." *American Political Science Review* 117 (3): 1070–88.

Fariss, Christopher J. 2019. "Yes, Human Rights Practices are Improving over Time." *American Political Science Review* 113 (3): 868–81.

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

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton.

https://doi.org/10.1017/S0003055424001199 Published online by Cambridge University Press

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): 707–20.

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): 310–20.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010a. “Beyond WEIRD: Towards a Broad-Based Behavioral Science.” *Behavioral and Brain Sciences* 33 (2–3): 111.

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

Hoffmann, Stanley. 1977. “An American Social Science: International Relations.” In *International Relations—Still an American Social Science? Toward Diversity in International Thought*, eds. Robert M. A. Crawford and Daryl S. L. Jarvis, 212–41. Albany: State University of New York Press.

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: 403–24.

Jones, Dan. 2010. *A WEIRD View of Human Nature Skews Psychologists’ Studies*. Washington, DC: 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): 484–99.

Kertzer, Joshua D. 2016. *Resolve in International Politics*. Princeton, NJ: 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): 539–53.

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): 234–49.

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): 246–69.

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): 345–57.

Levitsky, Steven, and Daniel Ziblatt. 2019. *How Democracies Die*. New York: Crown.

Lupu, Yonatan, and Geoffrey P. R. Wallace. 2019. “Violence, Nonviolence, and the Effects of International Human Rights Law.” *American Journal of Political Science* 63 (2): 411–26.

Marshall, Monty G., and Ted Robert Gurr. 2020. “Polity5: Political Regime Characteristics and Transitions, 1800–2018.” Center for Systemic Peace.

McDermott, Rose. 2011a. “Internal and External Validity.” In *Cambridge Handbook of Experimental Political Science*, eds. James N. Druckman, Donald P. Greene, James H. Kuklinski, and Arthur Lupia, 27–40. New York: Cambridge University Press.

McDermott, Rose. 2011b. “New Directions for Experimental Work in International Relations.” *International Studies Quarterly* 55 (2): 503–20.

Mutz, Diana C. 2021. “Improving Experimental Treatments in Political Science.” In *Advances in Experimental Political Science*, eds. James N. Druckman and Donald P. Green, 219–38. Cambridge: Cambridge University Press.

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

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): 826–44.

Powers, Kathleen E. 2022. *Nationalisms in International Politics*. Princeton, NJ: 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): 659–95.

Renshon, Jonathan, Keren Yarhi-Milo, and Joshua D. Kertzer. 2023. “Democratic Reputations in Crises and War.” *The Journal of Politics* 85 (1): 1–18.

Samii, Cyrus. 2016. “Causal Empiricism in Quantitative Research.” *The Journal of Politics* 78 (3): 941–55.

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): 655–68.

Seawright, Jason. 2016. *Multi-Method Social Science: Combining Qualitative and Quantitative Tools*. Cambridge: Cambridge University Press.

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

SIPRI. N.d. Stockholm International Peace Research Institute, Military Expenditure Database. https://doi.org/10.55163/CQGC9685.

Slough, Tara, and Scott A. Tyson. 2023. “External Validity and Meta-Analysis.” *American Journal of Political Science* 67 (2): 440–55.

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 (1): 1–33. https://doi.org/10.1590/1981-3821202000010002.

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

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

Tomz, Michael, Jessica L. P. Weeks, and Kirk Bansak. 2023. “How Membership in the North Atlantic Treaty Organization Transforms Public Support for War.” *PNAS Nexus* 2 (7): pgad206.

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 P. R. 2013. “International Law and Public Attitudes toward Torture: An Experimental Study.” *International Organization* 67 (1): 105–40.

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): 24–6.

Wolford, Scott. 2007. “The Turnover Trap: New leaders, Reputation, and International Conflict.” *American Journal of Political Science* 51 (4): 772–88.

World Bank. 2022. “World Development Indicators.” https://data.worldbank.org/indicator/NY.GDP.PCAP.CD. Accessed 04/5/2022.