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 I, 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.[19] 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).[20]

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,

[19] 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.

[20] 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.

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 5 treatment conditions, varying how easy it is for a foreign firm to buy a domestic firm, ranging from ”much harder to ”much easier”. 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

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 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).

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

F Heterogeneity

In Figure A1, 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 A2-A4, as well as in Table A12.

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 I. Finally, we find evidence in support of moderation when focusing on the legal obligation index. In other words, respondents with high levels of legal obligations are more opposed to the use of torture when assigned to the information treatment regarding government commitment to a treaty banning torture.

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)

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

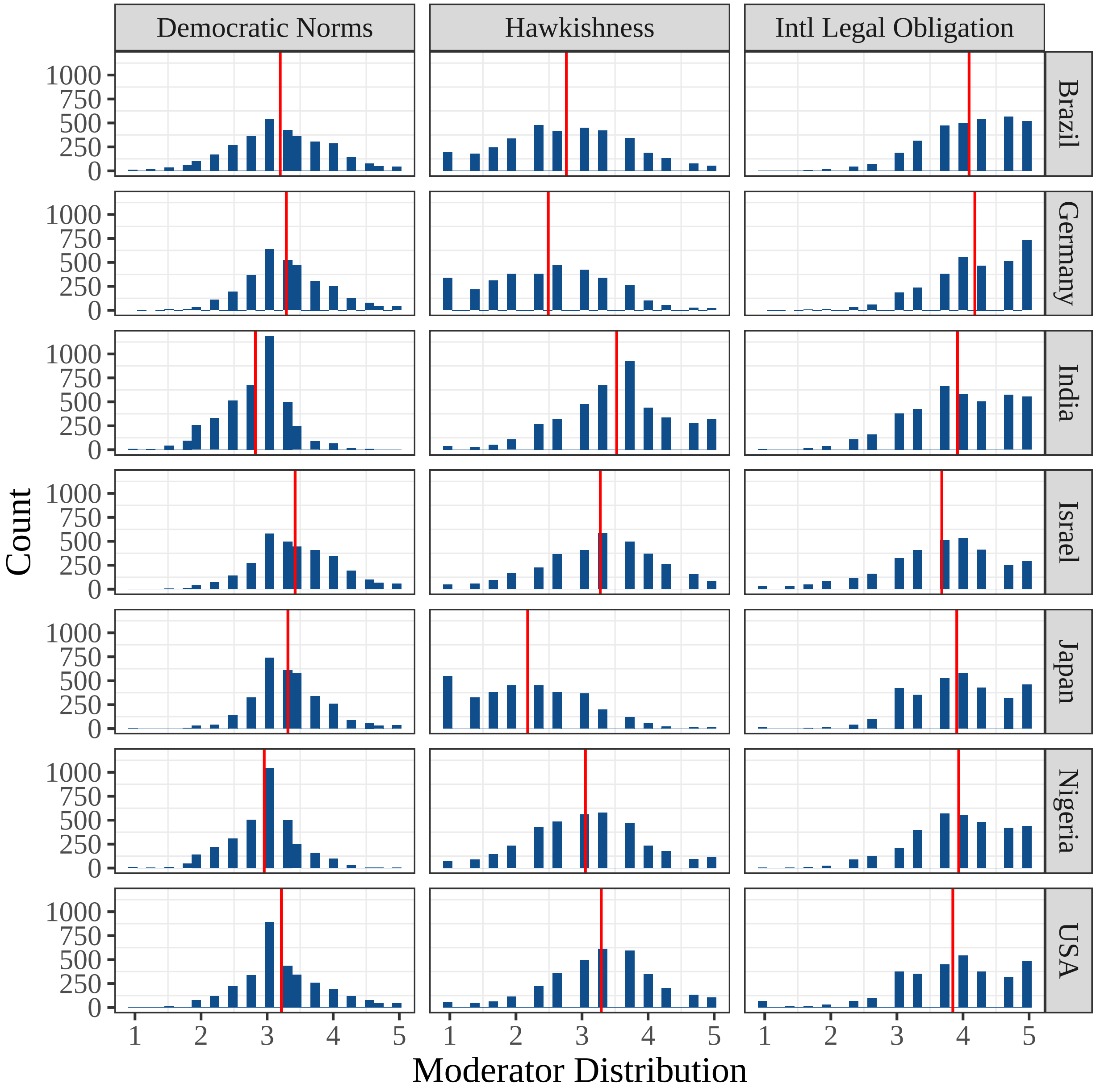

Figure A1: **Distribution of Moderators Across Countries.**

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

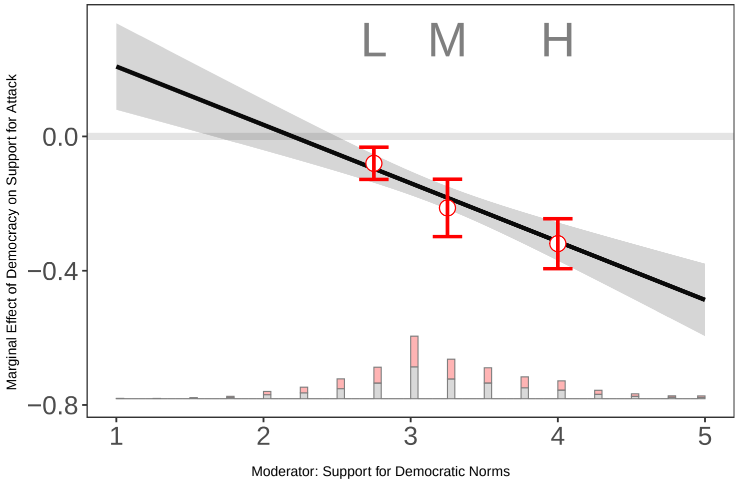

Figure A2: **Moderating Effect of Support for Democratic Norms Index in the Democratic Peace Experiment.** This figure demonstrates the negative moderation of support for democratic norms on the democracy treatment effects. That is, the effect of describing a country as a democracy reduces support for 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 A12.

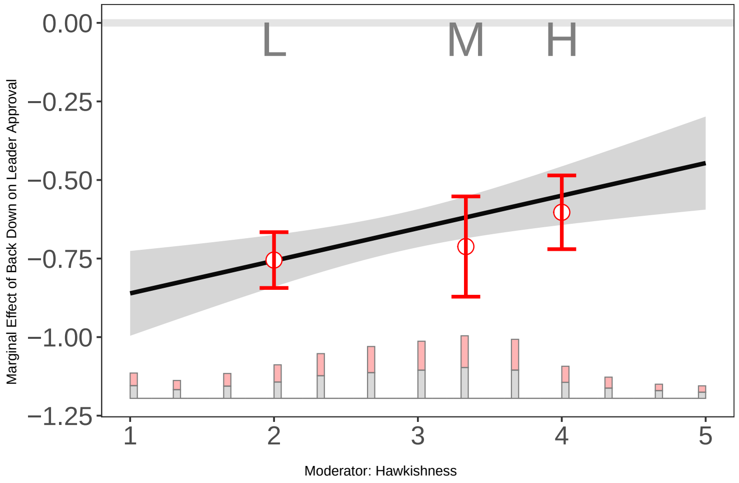

Figure A3: **Moderating Effect of Hawkishness Index in the Audience Costs Experiment.** This figure demonstrates there is no strong evidence for a moderation of hawkishness on the audience costs experiment. This figure corresponds to Table A12.

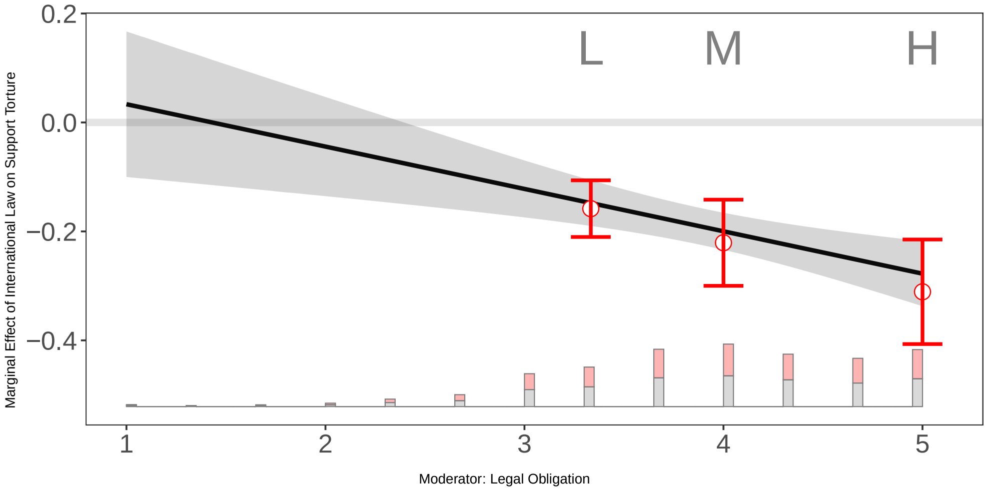

Figure A4: **Moderating Effect of International Legal Obligation Index in the International Law Experiment.** This figure demonstrates the negative moderation of legal obligation on the international law treatment effects. That is, mentioning that the respondent's country signed international law treaties 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 A12.

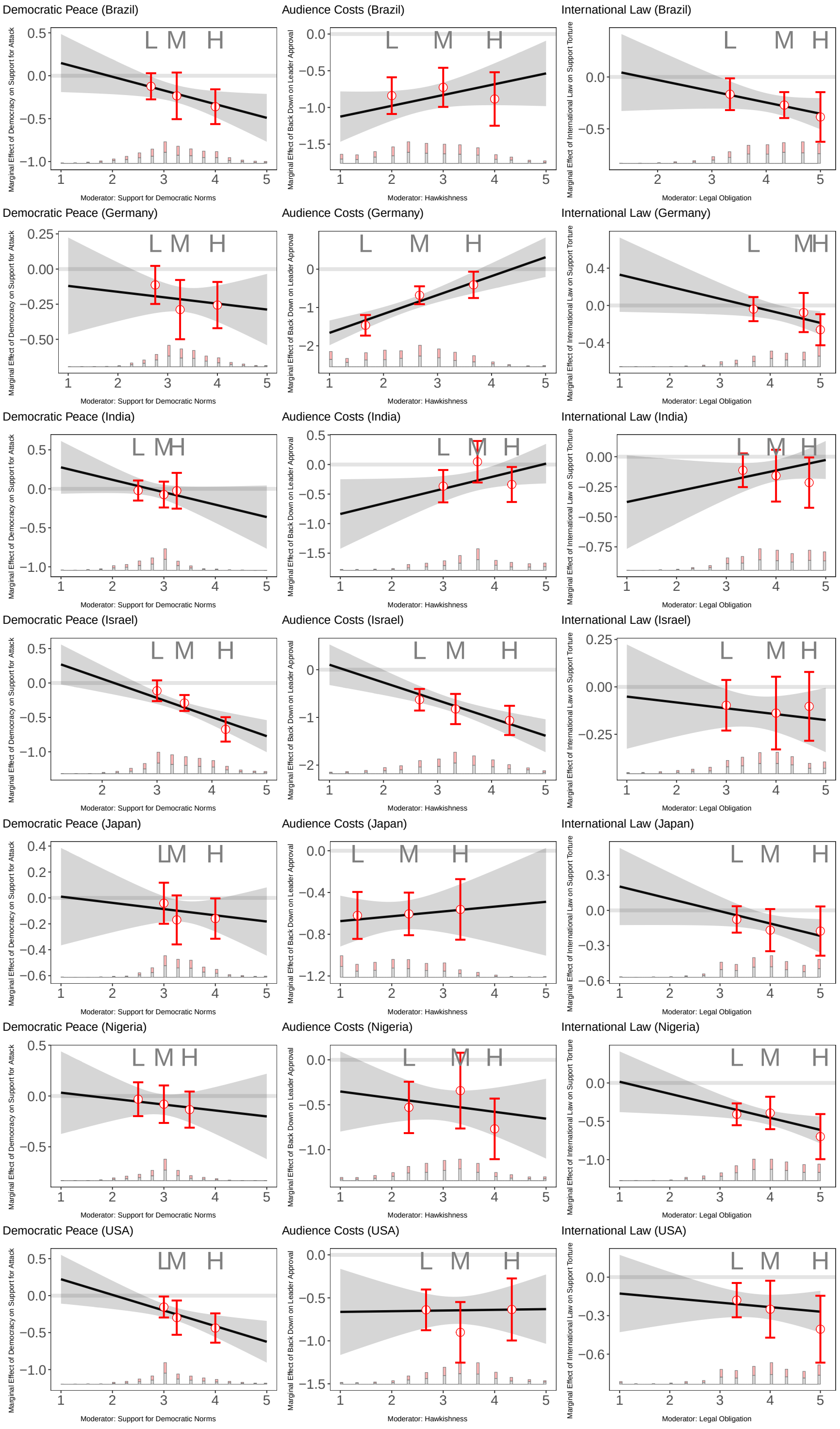

Figure A5: **Moderating effects in country-samples.** Individual figures of the moderating effects of democratic norms, hawkishness, and international legal obligation, in the DP, AC, and IL experiments, accordingly. Figures are broken down by country-samples. This figure corresponds to Tables A13-A15.

	Democratic Peace	Audience Costs	International Law
	Model 1	Model 2	Model 3
Democracy	−0.171*		
	(0.017)		
Dem Norms	−0.200*		
	(0.022)		
Dem*Norms	−0.133*		
	(0.031)		
Back Down		−0.659*	
		(0.031)	
Hawkish		0.078*	
		(0.027)	
BD*Hawk		0.056	
		(0.038)	
Intl Law			−0.196*
			(0.017)
Legal Obligation			−0.037*
			(0.017)
IL*Oblig			−0.076*
			(0.024)
Adj. R^2	0.181	0.098	0.221
Num. obs.	21426	14197	21445

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

Table A12: Moderation Tests

	Brazil	Germany	India	Israel	Japan	Nigeria	USA
Democracy	−0.205*	−0.217*	−0.016	−0.338*	−0.101*	−0.086	−0.248*
	(0.051)	(0.044)	(0.045)	(0.039)	(0.042)	(0.052)	(0.044)
Dem Norms	−0.091	−0.451*	−0.017	−0.097*	−0.464*	−0.095	−0.211*
	(0.052)	(0.053)	(0.069)	(0.047)	(0.056)	(0.075)	(0.056)
Dem*Norms	−0.137	−0.029	−0.175	−0.251*	−0.018	−0.062	−0.150
	(0.075)	(0.075)	(0.092)	(0.070)	(0.081)	(0.103)	(0.079)
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 A13: Moderation Test (Democratic Peace)

	Brazil	Germany	India	Israel	Japan	Nigeria	USA
Back Down	−0.870*	−0.926*	−0.290*	−0.747*	−0.622*	−0.520*	−0.645*
	(0.082)	(0.080)	(0.087)	(0.073)	(0.069)	(0.089)	(0.080)
Hawkish	0.014	−0.394*	0.416*	0.334*	−0.181*	0.285*	0.106
	(0.065)	(0.072)	(0.075)	(0.060)	(0.067)	(0.073)	(0.076)
BD*Hawk	0.135	0.474*	0.186	−0.375*	0.048	−0.085	0.025
	(0.092)	(0.099)	(0.110)	(0.093)	(0.092)	(0.106)	(0.110)
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 A14: Moderation Test (Audience Costs)

	Brazil	Germany	India	Israel	Japan	Nigeria	USA
Intl Law	−0.257*	−0.082	−0.119*	−0.134*	−0.104*	−0.449*	−0.227*
	(0.046)	(0.043)	(0.046)	(0.043)	(0.039)	(0.048)	(0.047)
Legal Obligation	−0.054	−0.265*	0.448*	−0.220*	−0.208*	0.205*	−0.195*
	(0.050)	(0.046)	(0.044)	(0.038)	(0.043)	(0.045)	(0.040)
IL*Oblig	−0.088	−0.066	0.065	−0.035	−0.112	−0.153*	−0.021
	(0.068)	(0.065)	(0.067)	(0.052)	(0.061)	(0.068)	(0.055)
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 A15: 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 A16: Estimating Differences Between Country Samples. Each model regresses relevant outcomes over country indicators compared to the US (which serves as a reference category).

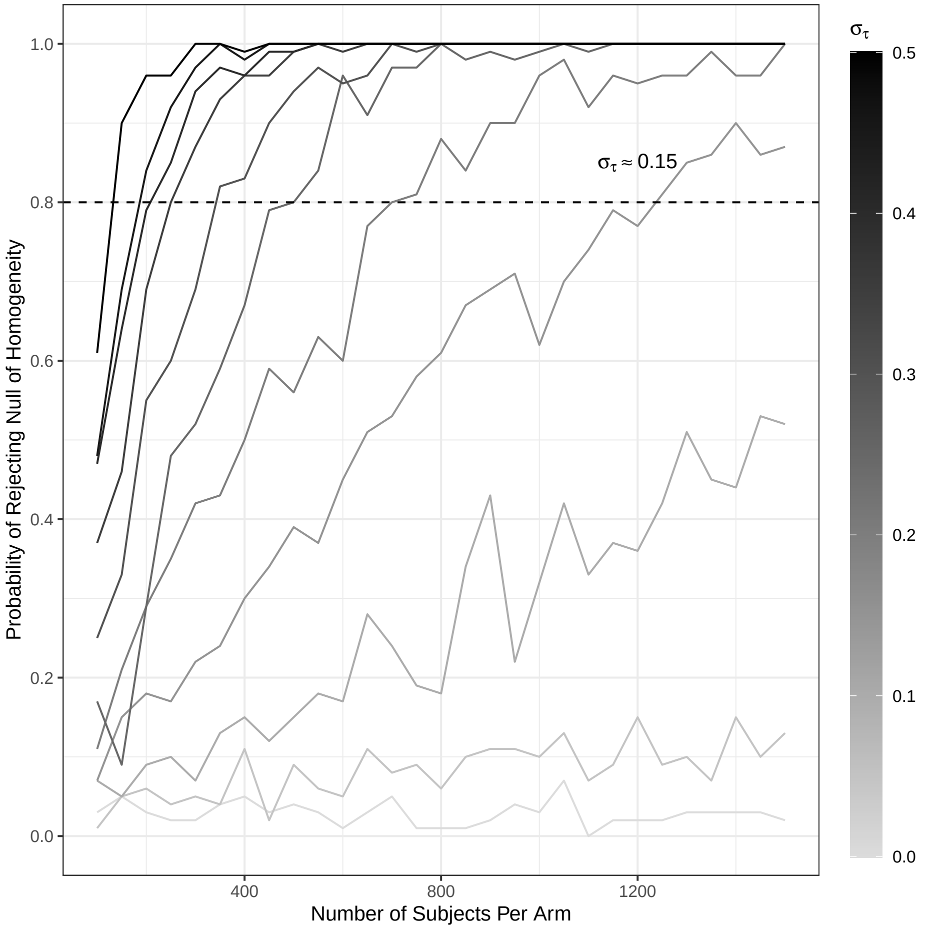

Figure A6: **Simulation study.** Power of Ding, Feller and Miratrix (2016) heterogeneity test, using R code from Coppock (2019).

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.

G 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 a population would have to be from our experimental sample in order to eliminate the treatment effect? External validity bias depends on both the level of treatment effect heterogeneity and the size of the treatment effect (Devaux and Egami, 2022, 11).

	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 A17: Alternative Outcome (DP): Treatment Effect on Joining Intl Mission

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 A7, 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.[21]

H 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 A17 using this alternative outcome measure are largely consistent with the results presented in the main text.

I 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 A8, 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 the main text, treatment effect heterogeneity likely explains why the belligerence treatment yields diverging effects across countries. Indeed, in their theory, Kertzer and Brutger (2016) argue that the ATE of belligerence — support for using force versus support for remaining out of the conflict altogether — should vary across subjects depending on their level of hawkishness. More hawkish subjects should be more likely to reward leaders who use force, while more dovish subjects should be more likely to punish belligerent leaders. We confirm this prediction

[21] 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).

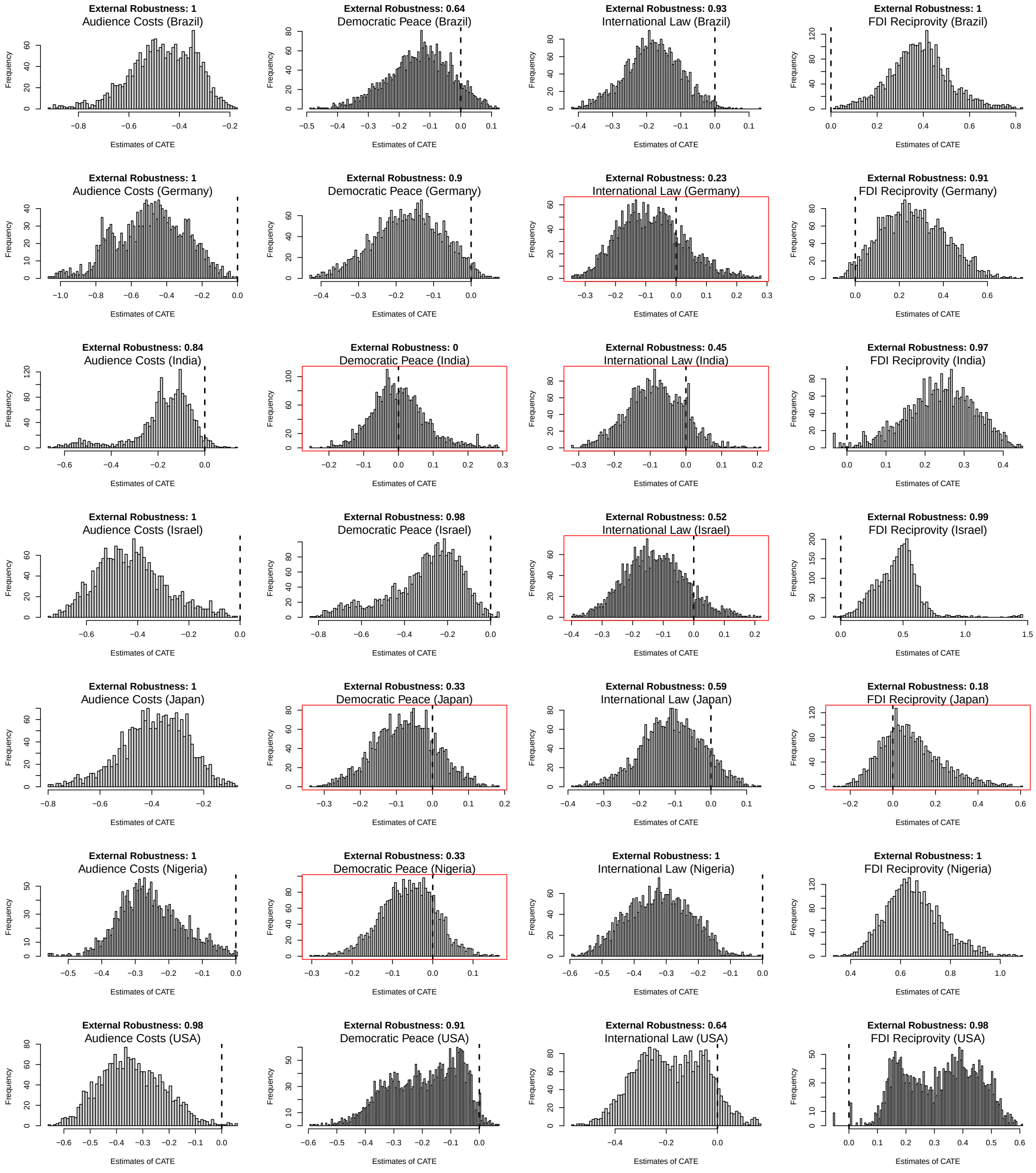

Figure A7: **External Validity Bias Test.**

	Beligerence Costs				Inconsistency Costs			
Samples	Estimate (Belligerence)	SE (Belligerence)	P value (Belligerence)	N (DP)	Estimate (Inconsistency)	SE (Inconsistency)	P value (Inconsistency)	N (Inconsistency)
Brazil	-0.07	0.08	0.36	2040	-0.79	0.08	0.00	2064
Germany	-0.50	0.08	0.00	2002	-0.42	0.08	0.00	2057
India	0.53	0.08	0.00	2037	-0.82	0.08	0.00	2094
Israel	-0.55	0.08	0.00	2040	-0.20	0.08	0.01	2013
Japan	0.08	0.07	0.29	1996	-0.73	0.07	0.00	2105
Nigeria	1.44	0.08	0.00	2118	-1.96	0.08	0.00	2073
USA	0.38	0.08	0.00	2023	-1.02	0.08	0.00	2013
All Countries	0.19	0.26	0.47	14256	-0.85	0.21	0.00	14419
Original (USA)	-0.56	0.17	0.00	711	-0.80	0.16	0.00	716

Table A18: Audience Costs Extension (Figure A8) in table form.

in Figure A9. 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 A9, 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 the main text, homogeneity tests proposed by Ding, Feller and Miratrix (2019) 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 A9 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.[22]

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 our manuscript when discussing our results, 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 of our Dataverse-only appendix, 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

[22] We suspect that at least two results from Figure A8 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.

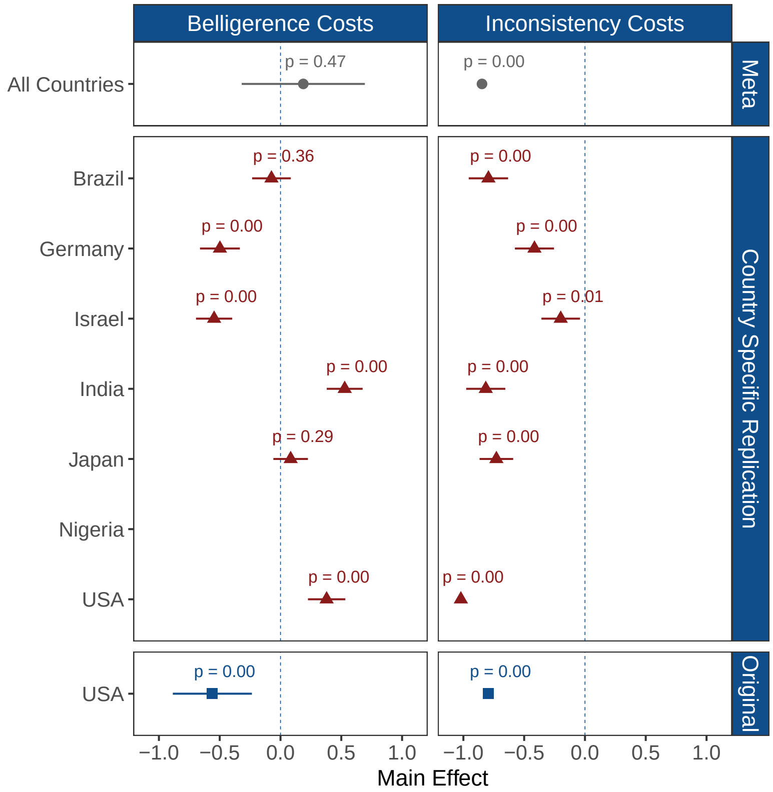

Figure A8: **Audience Costs extension.** We report the original estimates and p-values of the belligerence and inconsistency costs, as calculated in the original study. We further report the country-specific ATEs (and BH-adjusted p-values) from our replications, and a meta-analytic average treatment effect based on our harmonized studies. This Figure corresponds to Table A18.

	Belligerence Costs	Inconsistency Costs
Belligerence*Hawk	0.34^{*}	
	(0.03)	
Belligerence	-0.80^{*}	
	(0.09)	
Hawk	0.06^{*}	0.37^{*}
	(0.02)	(0.02)
Inconsistency*Hawk		-0.24^{*}
		(0.03)
Inconsistency		-0.15
		(0.10)
Adj. R^2	0.08	0.14
Num. obs.	14270	14433

$^{*}p < 0.05$

Table A19: Moderating effect of Hawkishness in AC extension (Figure A9) in table form.

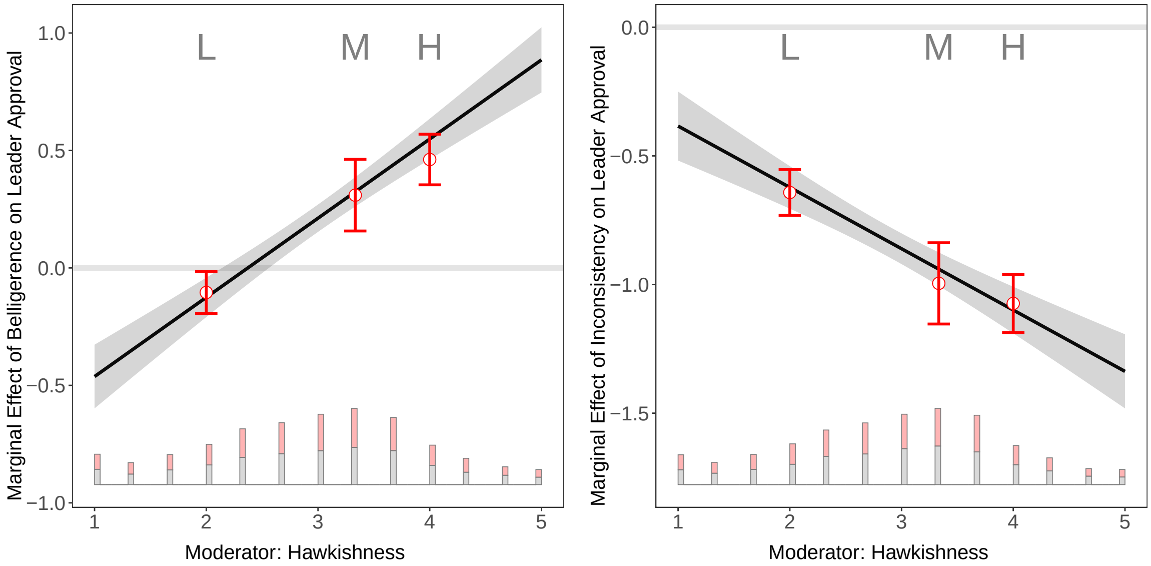

Figure A9: **Moderating Effect of Hawkishness Index in the extension of the Audience Costs Experiment.** This figure demonstrates the negative moderation of hawkishness on the inconsistency treatment effects, and the positive moderation of hawkishness on belligerence treatment effect. These results are consistent with findings from the original study. This Figure corresponds to Table A19.

	BRZ	GRM	IND	ISL	JPN	NGR	USA
Belligerence*Hawk	0.26^{*}	0.91^{*}	-0.02	0.08	0.70^{*}	-0.11	0.39^{*}
	(0.09)	(0.09)	(0.09)	(0.09)	(0.08)	(0.09)	(0.09)
Belligerence	-0.79^{*}	-2.76^{*}	0.58	-0.80^{*}	-1.46^{*}	1.78^{*}	-0.87^{*}
	(0.25)	(0.23)	(0.32)	(0.29)	(0.19)	(0.30)	(0.30)
Hawk	0.01	-0.41^{*}	0.41^{*}	0.28^{*}	-0.19^{*}	0.28^{*}	0.10
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)	(0.06)
Adj. R^2	0.02	0.08	0.08	0.06	0.05	0.14	0.08
Num. obs.	2042	2004	2039	2042	1998	2120	2025

$^{*}p < 0.05$

Table A20: Belligerence costs moderating effects by country (Figure A10) in table form.

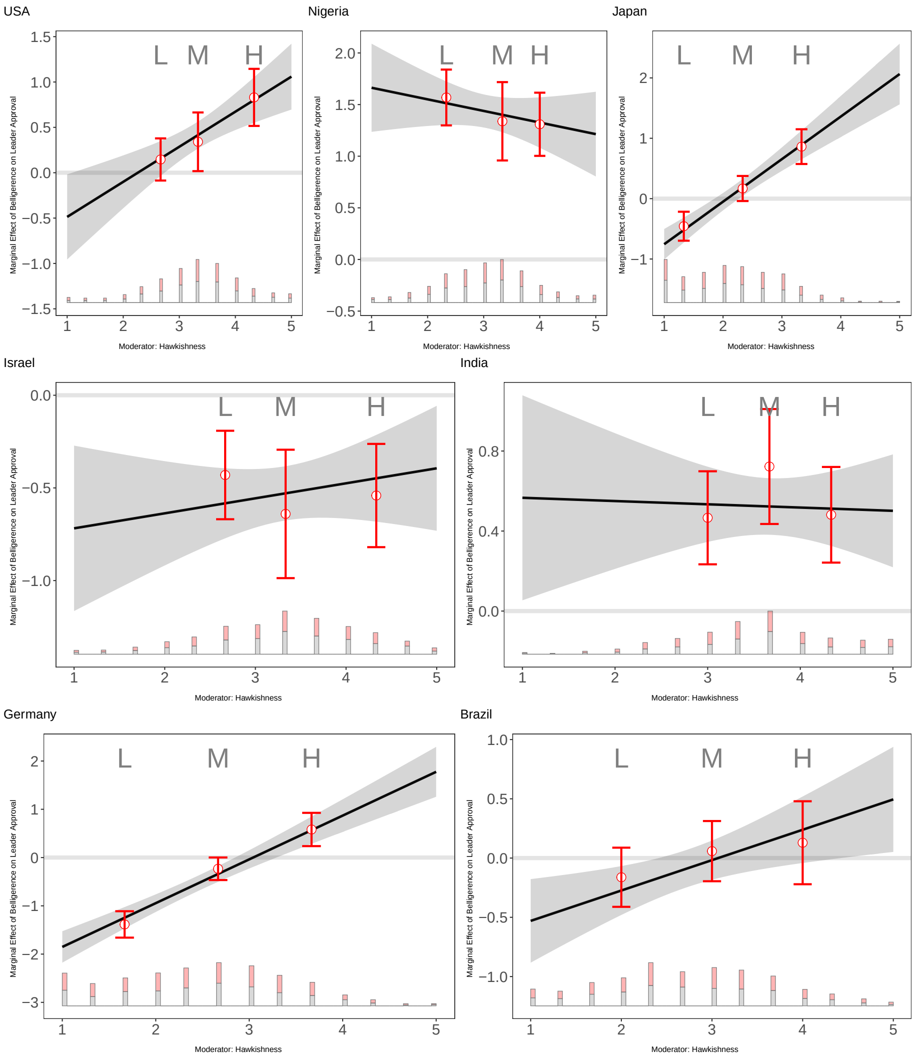

Figure A10: **Belligerence Costs Moderating effects in country-samples.** Individual figures of the moderating effects of hawkishness in the audience costs extension (belligerence costs). Figures are broken down by country-samples. This Figure corresponds to Table A20.

Figure A3 of our Dataverse-only appendix we demonstrate that respondents in India thought of similar countries across both conditions, with most respondents thinking of Pakistan and China. However, it is possible that if respondents in India always thought of an adversary like Pakistan, then perhaps they were prone to strike in both experimental conditions, muting the treatment effect. We note that in other countries – Israel and Japan – the proportion of respondents who name the same country (Iran and North Korea, respectively) was much higher in comparison to India, making them more obvious candidates for muted effects due to confounding. Nonetheless, it is possible that the 'true effect' of democracy in Israel and Japan is much larger than in India, allowing us to identify the effect regardless of information leakage. We are thus unable to fully rule out information leakage as a potential explanation.

3. *Floor or ceiling effects:* We examine whether our sample in India is prone to floor or ceiling effects due to particularly high or low levels on our outcome of interest – support for attacking the other country's nuclear facilities. We determine that this is an improbable explanation for two reasons. First, while the mean of the India sample on our main outcome in the democratic peace experiment was relatively high (3.75 on a scale of 1 to 5) it is not as high as the mean in the Israel sample (3.99) or as low as the mean in the Japan sample (2.33) which would be more obvious candidates for ceiling and floor effects, respectively (see Figure A7 of our Dataverse-only appendix where we plot the means by condition for each country-study pair). 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 A17).

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 A1 in our Dataverse-only appendix, 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 A21). Hence, while we cannot rule it out completely, we conclude that inattentiveness cannot serve as the sole explanation for the null effect in India.

5. *Ineffective mechanisms:* Finally, it is possible that the mechanisms outlined in the original democratic peace experiment by (Tomz and Weeks, 2013) do not generalize to India. Perhaps due to the ongoing conflict with Pakistan, a country which is occasionally labeled as a democracy, subjects in India have learned that democracies are not less threatening or costlier to attack, and that it is not normatively 'wrong' to attack a democracy. Our current design does not allow us to evaluate this explanation, but future research may wish to survey respondents in India about their beliefs about democracies with respect to threats, morality or cost of war.

J Robustness Checks

In this section we report additional robustness checks. First, we report in Figure A11 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 A12 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.

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

$^{*}p < 0.05$.

Table A21: Screening out failed manipulation from other studies (India DP)

	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	-0.19	0.05	0.00	3052	-0.87	0.08	0	1996	-0.23	0.04	0.00	3045	0.50	0.05	0.00	3051
Germany	-0.22	0.04	0.00	2992	-0.93	0.08	0	1943	-0.09	0.04	0.04	2997	0.31	0.04	0.00	3005
India	0.00	0.04	0.96	3067	-0.33	0.08	0	2011	-0.13	0.04	0.01	3062	0.31	0.05	0.00	3064
Israel	-0.34	0.04	0.00	3064	-0.76	0.07	0	2081	-0.16	0.04	0.00	3072	0.53	0.04	0.00	3060
Japan	-0.09	0.04	0.03	3048	-0.62	0.07	0	2021	-0.12	0.04	0.00	3055	0.08	0.04	0.04	3056
Nigeria	-0.09	0.05	0.13	3122	-0.52	0.09	0	2071	-0.44	0.05	0.00	3129	0.97	0.05	0.00	3129
USA	-0.26	0.04	0.00	3011	-0.66	0.08	0	2004	-0.21	0.04	0.00	3015	0.43	0.05	0.00	3011
All Countries	-0.17	0.04	0.00	21356	-0.67	0.08	0	14127	-0.20	0.04	0.00	21375	0.45	0.10	0.00	21376

Table A22: Meta analysis with controls (Figure A11) in table form.

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 A23 reports the main treatment effects, conditional on the survey language. We do not identify any statistically significant heterogeneous effects here.

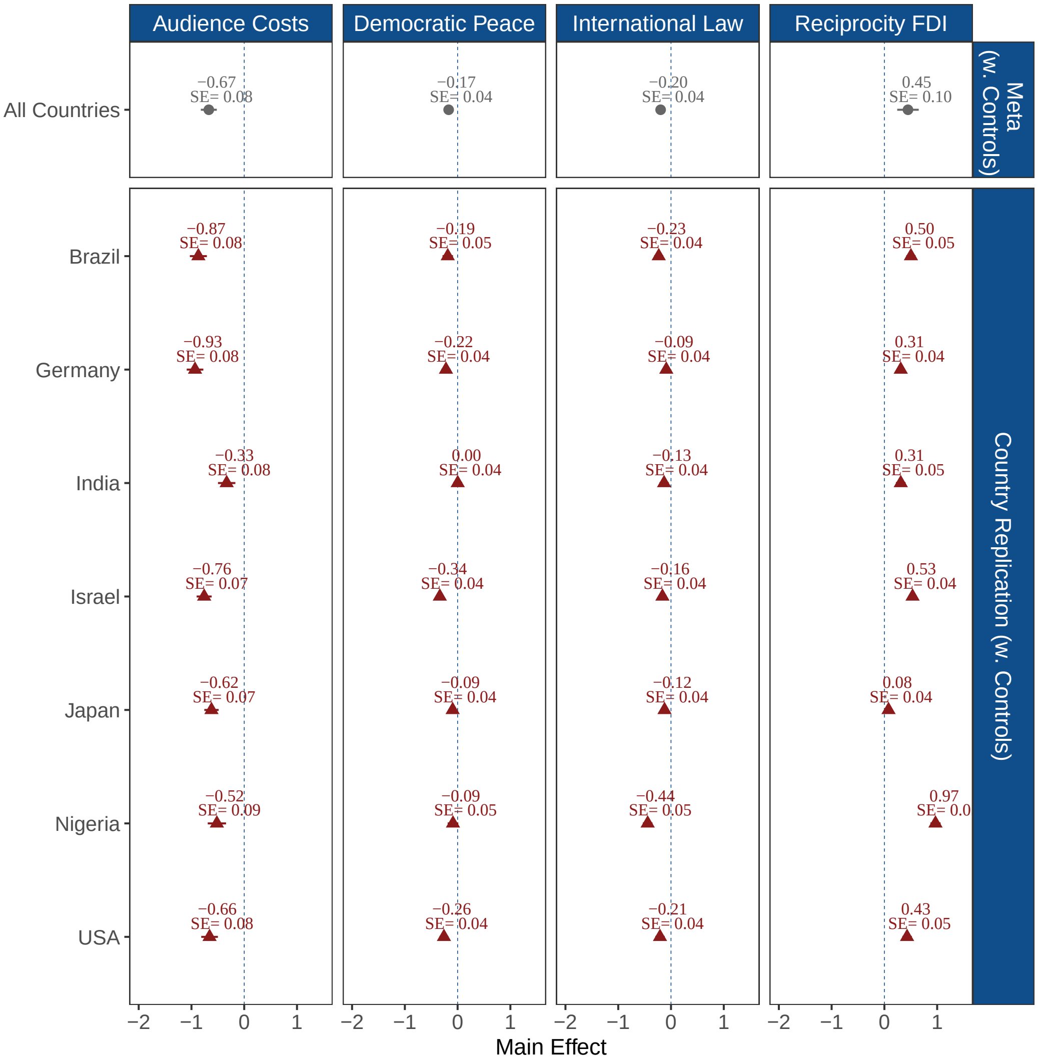

Figure A11: **Main analysis with demographic controls.** We report the estimates and standard errors of models where the following covariates have been added: Gender, Age, Ideology, Education, Voting, Democratic norms, Hawkishness, Legal obligation. This Figure corresponds to Table A22.

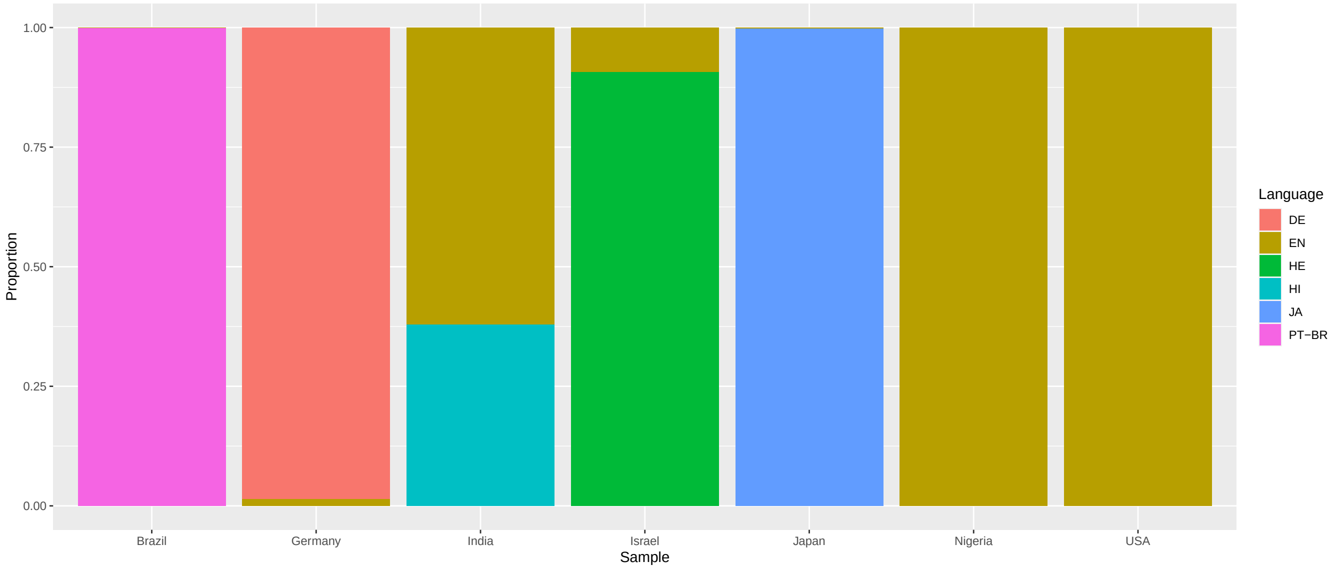

Figure A12: **Proportion of Use Languages per country sample.** Note that the majority of respondents took the surveys in the national/local languages.

	Dem Peace	Audience Costs	Int Law	Reciprocity
Intl Law			0.002	
			(0.154)	
English	−0.186	−0.013	−0.437*	−0.014
	(0.110)	(0.214)	(0.113)	(0.123)
IL*English			−0.106	
			(0.162)	
Back Down		−0.203		
		(0.285)		
BD*English		−0.095		
		(0.300)		
Democracy	−0.062			
	(0.148)			
Dem*English	0.056			
	(0.156)			
Harder barrier				0.391*
				(0.160)
Hard*English				−0.112
				(0.169)
Adj. R^2	0.000	0.004	0.012	0.011
Num. obs.	3077	2021	3072	3074

$^*p < 0.05$

Table A23: Treatment Effect*English in India Sample