# CONFIDENTIAL - FOR PEER-REVIEW ONLY

## Generalizability of IR Experiments (#119402)

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

---

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

---

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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