Do Experimental Results Generalize Across Countries?
Lessons from Twenty-Five Multi-Country Experiments in Political Science

Martin Devaux (Columbia), Naoki Egami (MIT)

Motivation

Countries around the world exhibit a wide diversity of economies, political institutions, and cultural norms. Influential work claims that social scientific theories should therefore be context-specific and not universal (Markus and Kitayama, 1991; Henrich et al., 2010).

Yet, recent studies have found high levels of homogeneity across countries:

- In international relations (Bassan-Nygate et al., 2024)
- In domestic politics (Banerjee et al., 2015; Dunning et al., 2019)

Do political science theories work differently in different contexts?

We analyze the universe of multi-country experiments to understand the state of cross-country heterogeneity in experimental political science

- 70 study-treatment pairs from 25 papers (with at least four countries)
- APSR, AJPS, JOP, Science, Nature, and PNAS
- Published between 2017 and 2024

Summary of the findings

Country ATEs are heterogeneous:

- Standard deviation of country-specific ATEs = 0.67 x Average of country-specific ATEs
- 43% of experiments ran in only one typical country (e.g., US, UK) would significantly overestimate ATE compared to average
- But sign of country-specific ATE is the same as the average of country-specific ATEs in 87% of cases

Heterogeneity is hard to explain:

- Country-level moderators and study designs do not predict it well

Heterogeneity we measure is a lower bound due to site selection

Implications for future researchers

1. Country-specific ATEs are heterogeneous, justifying external validity concerns
   - More multi-country experiments are needed to test it
2. Theoretical developments are needed:
   - How and why are country effects expected to differ?
3. Site selection should be diversified and follow theoretical considerations
   - Currently few countries and often little overlap between world regions
4. As the number of countries grows, MCEs will be able to answer more:
   - E.g., preregistered meta-regressions on heterogeneity

Methodology

- In each paper, we interpret different treatments as separate studies
- For country *j* in study *s*, estimate country-specific ATE:

$$\hat{\tau}_j = \frac{1}{N_{j1}} \sum_{i=1}^{N_j} T_i Y_i - \frac{1}{N_{j0}} \sum_{i=1}^{N_j} (1 - T_i) Y_i$$

- Cross-country heterogeneity defined as standard deviation of $\hat{\tau}_j$
- Report heterogeneity standardized by average of country-specific ATEs

$$\tilde{\sigma}_s = \frac{\sqrt{\mathrm{Var}(\hat{\tau}_j)}}{\frac{1}{J}\sum_{j=1}^{J} \hat{\tau}_j}$$

Bootstrap confidence intervals (resampling within site/treatment)

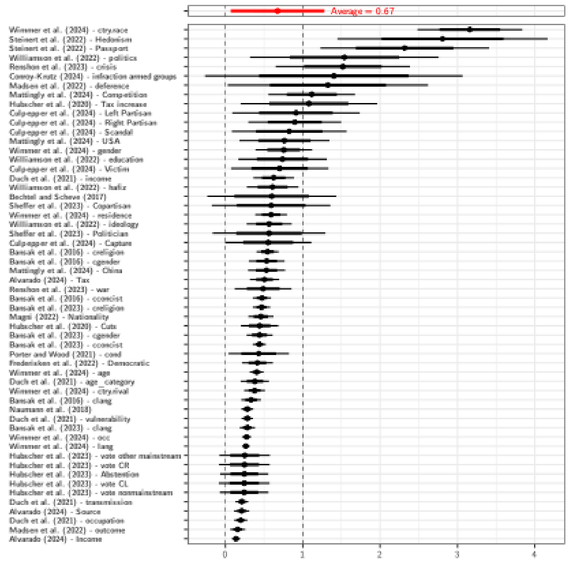
Population differences accounted for via entropy balancing

1. ATEs are heterogeneous across countries

We find evidence of heterogeneity across a large majority of studies:

- Standard deviation of country-specific ATEs = 0.67 x average of country-specific ATEs

Figure 1. Heterogeneity as a proportion of average country ATE



95% confidence intervals displayed. Larger confidence intervals are Bonferroni-corrected.
The average country ATE is the average of the country-specific ATEs in each study.
Null studies are excluded for clarity (variance much larger than ATE).

2. Single-country experiments may overestimate the ATE

We test what the conclusions would be with single-country studies:

- 43% of experiments ran in only one typical country (e.g., US, UK) would significantly overestimate ATE compared to average
- But sign of country-specific ATE matches the average of country-specific ATEs in 87% of cases

3. Typical country-level moderators do not explain heterogeneity

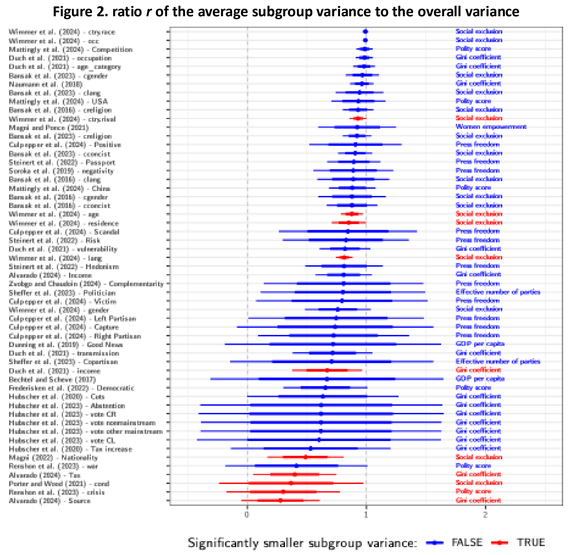
Do relevant country-level moderators explain heterogeneity?

- Measure variance reduction ratio from splitting countries by variable X:

$$r = \frac{\frac{n_{above}}{n}\widehat{\mathrm{Var}}\left(\hat{\tau}_j | X_j \geq X_{med}\right) + \frac{n_{below}}{n}\widehat{\mathrm{Var}}\left(\hat{\tau}_j | X_j < X_{med}\right)}{\widehat{\mathrm{Var}}\left(\hat{\tau}_j\right)}$$

- The variance is reduced significantly in only a dozen cases.

Figure 2. ratio *r* of the average subgroup variance to the overall variance



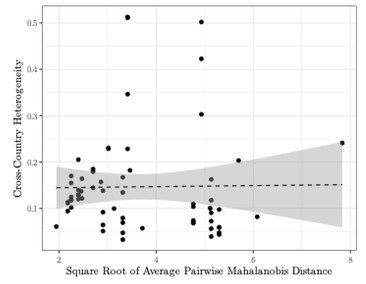
95% confidence intervals displayed. Larger confidence intervals are Bonferroni-corrected.

4. Study characteristics do not predict heterogeneity

Methodological and substantive study characteristics do not predict the amount of heterogeneity:

- Experiments with more diverse country sets are not more heterogeneous

Figure 3. Heterogeneity as a function of selected country diversity



- Topic and type of experiment are not clear predictors

Figure 4. Heterogeneity by topic

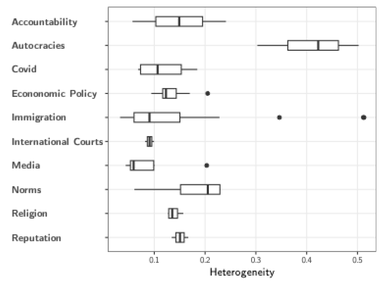
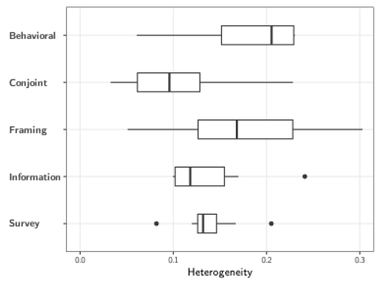


Figure 5. Heterogeneity by study type

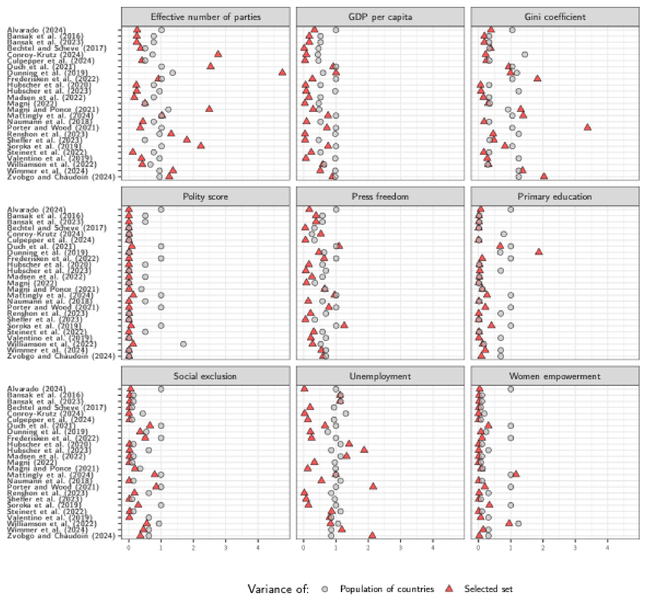


5. Heterogeneity measured is a lower bound due to site selection

Sampling remains limited by researchers' constraints:

- 13 out of 25 of the studies include five countries or fewer
- 70% of selected countries located in Europe and North America
- Countries show less variation than broader population of countries they are intended to represent (e.g., European countries, democracies)

Figure 6. Variance of typical country-level moderators



References

Banerjee, A., Karlan, D., & Zinman, J. (2015). Six Randomized Evaluations of Microcredit: Introduction and Further Steps. *American Economic Journal: Applied Economics*, *7*(1), 1–21.

Bassan-Nygate, L., Renshon, J., Weeks, J. L. P., & Weiss, C. M. (2024). The Generalizability of IR Experiments beyond the United States. *American Political Science Review*, 1–16.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29–29.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*(2), 224–253.