The Aggregation Challenge

Essay written for *World Development* Symposium on RCTs for Development and Poverty Alleviation in recognition of the 2019 Nobel Prize awarded to Banerjee, Duflo, and Kremer.

Macartan Humphreys and Alexandra Scacco

November 2019

Abstract. Banerjee, Duflo, and Kremer have had an enormous impact on scholarship on the political economy of development. But as RCTs have become more central in this field, political scientists have struggled to draw implications from proliferating micro-level studies for longstanding macro-level problems. We describe these challenges and point to recent innovations to help address them.

Introduction

Banerjee, Duflo, and Kremer have had an enormous impact on work on the political economy of development. Although there were parallel trends in political science (e.g., Green and Gerber 2002), much of the early expansion of RCTs into topics in political economy was not just inspired by but led by the Nobel laureates and their close associates. Key studies on leadership, aid politics, and accountability include Chattopadhyay and Duflo (2004), Gugerty and Kremer (2008), and Banerjee et al. (2011). Olken (2004) and Miguel et al. (2004) used experiments and natural experiments to address complex questions about corruption and political violence. Today, hundreds of researchers are implementing experiments examining the political economy of development. There are well over a thousand registered experimental studies on the Evidence in Governance and Politics (EGAP) registry alone. Experiments have taken root.

At the same time, there are broad concerns the experimental turn has diverted researchers from core questions in our field. Indeed, there has been an awareness of this risk from the outset. In 2006, Robert Bates wrote that "Banerjee's approach might teach us more about impact but at the expense of larger matters," warning political scientists against transforming the field "from a search for the underlying forces of development into a form of policy analysis." Yet despite consciousness of this concern, attempts to aggregate the lessons learned from rigorous micro-level experimental work in order to shed light on larger puzzles have been at best casual.

Micro-Macro Disconnects

The aggregation problem is distinct from the problem of external validity of experimental results, though that's a part of it. Rather, the problem is like knowing how pieces of a jigsaw puzzle fit together when many of the pieces are missing.

Three examples of aggregation problems:

- You are interested in whether freedom of the press fosters better government. What can you learn from an experiment that shows that voters are more likely to vote against politicians when they are told that they are underperforming?

- You are interested in whether natural resources weaken state-society linkages. What can you learn from an experiment that shows that voters that are told that revenues are derived from natural resources – rather than taxes – exhibit less concern about government expenditures?
- You are interested in whether inter-ethnic violence is caused by residential segregation. What can you learn from an experiment that shows that prejudice decreases among individuals exposed to higher levels of contact with out-group members?

In all three cases, a macro-level question motivates the research, but the researcher has micro-level experimental evidence at hand. The micro level experiments *seem* to provide relevant evidence, but it is unclear what inferences to draw from this micro evidence for the macro questions.

Let's take the third example and use it to flesh out three distinct parts of the aggregation problem. We are interested in whether ethnic conflict is exacerbated by segregated settlement patterns. This is an attribution question. A macro-level hypothesis (drawn from several rich largely non-experimental literatures in political science) might be that certain residential patterns – such as ethnically mixed but highly segregated cities – can heighten feelings of insecurity which, when exploited by opportunistic political elites, plays an important role in explaining the incidence of communal violence.

Figure 1 illustrates with a simple graphical model, linking ethnic demography and conflict at the macro level. Different parts of the model find support in different bodies of research. Kasara (2017) used observational data at the locality level to document the relationship between segregation and levels of intergroup violence in Kenya. Allport's (1954) work on the negative relationship between contact and prejudice sparked decades of work on this link. Horowitz (1985) has explored connections between prejudice and ethnic conflict in multiple settings, and a rich literature has linked the strategic behavior of political elites to communal violence.

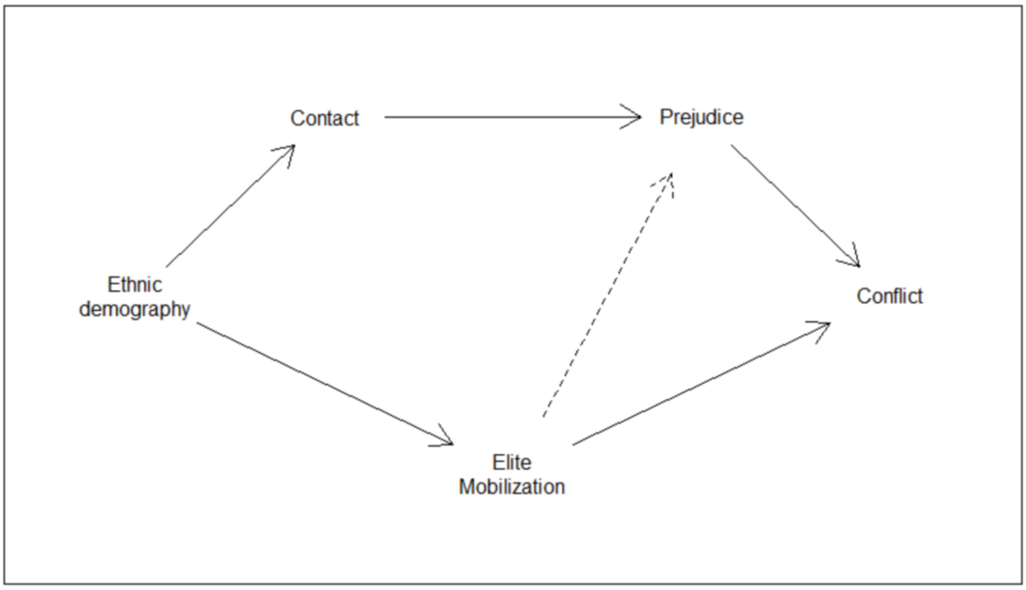

Figure 1: Simple macro-micro model of ethnic demography and conflict linkages. We are interested in the effect of demography on conflict. What can we learn from experimental evidence on the effect of contact on prejudice?

There are obvious reasons to be worried about multiple forms of confounding for the model shown in Figure 1. But there is little scope, for practical and ethical reasons, for experimentation at the macro level on demographic structures. Conscious that experimentalists have repeatedly shown that randomization can be used for a much wider set of problems than skeptics expected, we take it as given that major "treatments" such as conflict histories and demographic structures, are out of reach. Can (individual) micro-level interventions help?

Fortunately, the theoretical accounts in macro-level studies often specify micro-level logics (Kasara interprets her macro findings through the lens of a micro-level hypothesis about interpersonal mistrust in settings of segregation) and recent experimental studies have examined some of these micro hypotheses directly. In one example, Scacco and Warren (2018) conducted an education-based field experiment in which 850 randomly sampled Christian and Muslim young men in a riot-prone city in Nigeria were randomly assigned to religiously mixed or homogeneous computer training courses. After four months of intergroup contact, they found significant declines in discrimination among subjects assigned to mixed classes.

The question is whether and how inferences from such a contact study can help us understand the macro-level relationship between segregation and prejudice, or segregation and violence.

Making direct inferences in this example is difficult for a number of reasons. We highlight three.

Misunderstood selection. A great advantage of randomization is that it can overcome selection biases. This achievement creates a problem, however, if the macro processes you want to study involve self-selection. Following the example illustrated in Figure 1, we can imagine situations where, even if values on the macro-level node (say, "the share of individuals engaging with out-group members") are randomly assigned, the values of micro-level nodes (whether a given individual encounters out-group members) might not be.

Imagine, for example, that for a given level of societal segregation, the set of individuals that encounters out-group members self-select from among those for whom social contact has the weakest effect. Say an experimentalist randomly selects individuals that have rarely been exposed to out-group members and experimentally induces exposure. In this case, their estimate of the average effect would be larger than the average effect for the population and would be a poor estimate for the effect of segregation. The question the researcher needs to answer is what would be the average, or overall, effect of treatment for those that self-select into it, given overall levels of segregation.

We need to understand selection logics in order to map from micro estimates to macro estimands, but, by design, experimental approaches often prevent us from learning about them.

Micro averages, macro nonlinearities. Experimental approaches often measure the average effect of a binary treatment, and average effects are treated like linear coefficients when making inferences regarding aggregate effects. If outgroup exposure results in one fewer interpersonal conflict, on average, then treating one person will reduce conflicts by 1 and exposing 1000 will reduce conflicts by 1000. Yet, while experiments typically measure average effects *given an overall level of exposure*, the overall level of exposure may also matter for the average effect, and variation in the level of exposure (such as the degree of segregation) may be precisely the variation of interest at the macro level. We need to understand the nonlinearities for a host of case level counterfactual estimands (what would be the level of prejudice in a given country if exposure was at level *x* rather than level *y*?). Effects may be different at different exposure levels because of spillover or general equilibrium effects. But there are other reasons this might be the case. For instance, the kind of self-selection into contact we described above produces a nonlinear relationship between the share of individuals exposed to out-groups and the share that is prejudiced.[1] One implication of this is that studies in different sites can produce different answers, even if the underlying causal processes are the same everywhere.

The average effect of a micro-level treatment can depend on the overall (macro-level) exposure level which can produce nonlinearities in effects of exposure levels. Understanding these nonlinearities can be important for macro level attribution questions and may require multisite studies.

Aggregation requires understanding rival pathways. If our goal is to draw inferences about the effects of ethnic demography on intergroup conflict from evidence about the operation of one step in a causal chain – the effects of interpersonal contact on prejudice – then we need to know something about how alternative paths operate. If the relationship between ethnic demography and intergroup contact is well understood (for example, residential segregation can reasonably be understood to contact), and if there are no other channels through which demography affects prejudice, then an argument for aggregation might invoke "the front door criterion" (Pearl 2009) – where the effect of demography on prejudice is the effect of demography on contact multiplied by the effect of contact on prejudice. But if there are rival pathways, this argument becomes much harder since in that case it is possible that segregation could increase prejudice (see the dotted line in Figure 1; an example of an alternative path might be via a mechanism of mobilization by opportunistic political elites), even if there is a conditional negative effect running through individual contact.

You cannot draw implications for the effects of X on Y from evidence of effects along links on a causal chain between X and Y unless you understand alternative paths from X to Y.

Pointers toward solutions

These challenges are present even in *auspicious* settings, where the macro nodes in Figure 1 are simple aggregates of micro-level measures, where macro variables are as-if-randomly assigned, and where we assume that experimental estimates correctly estimate effects induced by observational variation.

To make progress addressing the aggregation challenge, we need to grapple with a number of problems that do not figure prominently enough in current research.

---

[1] A more formal illustration: Imagine a polity with a unit mass of individuals. Say that in the case that share *k* individuals encounters out-group members these, under natural assignment, will be individuals in [0, *k*]. Say everyone is prejudiced in the absence of contact, and encountering out-group members eliminates prejudice for individuals in [*m*, 1]. Then there is a nonlinear relationship between the share exposed to out-groups and the share that is prejudiced, with no effect for $k < m$ and a unit effect for $k > m$. The effect of moving from no contact to full contact is (1-m). However an experimentalist implementing a contact experiment with individuals randomly selected from among those that have not been exposed to out-groups will find an effect of 1 If $m<k$ and of $(1-m)/(1-k)$, otherwise. In both cases the estimate is too large. Note also that the answer depends upon *k*. If they implemented the experiment on already exposed they would face a compliance problem.

First, we need to better understand processes of selection. In order to translate from the individual-level marginal effect to the macro-level marginal effect we need to *understand* self-selection rather than simply remove it. One way to do this might be for researchers to randomly partition their study sample into two groups. In one, one assignment to treatment is experimentally controlled. In the other, study subjects self-select into treatment. Combining data from these sub-samples would allow for updating on treatment effects, on selection propensities, and on how these relate to each other. Another promising approach is provided by Chassang et al. 2012, in which an experiment simultaneously studies incentives to enter treatment and the effects of treatment given different propensities to enter treatment.

Second, we need to better understand nonlinearities in relevant relationships at the micro and macro levels. To understand nonlinearities in the macro effect (for example, the relationship between segregation and conflict), we might build on innovations in multi-site experimental studies, as exemplified by the "Metaketa" approach (Dunning et al. 2019). For the contact example, one would want to understand how effects of contact differ in locations with greater or lower levels of baseline exposure to out-group members.

Third, we need to better understand rival pathways. To justify mappings from micro-level evidence to macro-level claims, we need to articulate a theory that provides fuller mappings between macro treatments, micro treatments, and outcomes. Even if the link from segregation to contact is well understood, understanding the marginal effect of contact on prejudice alone may tell us little about what effects are due to segregation if segregation also operates through other channels (as illustrated by the dotted line in Figure 1). In this case, we need to know how to best put all of the pieces together. We need tools to combine inferences from multiple sources in a principled way. Structural approaches provide pointers here, for example, as demonstrated by Pearl and Bareinboim (2014) on transportability and data fusion.

Finally, we need to start actually doing it. The norm in research in the political economy of development is to generate tight micro-level inferences and then gesture towards macro-level implications. Doing more to figure out which inferences for larger questions are justifiable will likely require a commitment to articulating how macro conclusions can be justified from micro data, greater re-coordination of research around core substantive agendas, a greater openness to learning from data even when we only enjoy partial identification, and a greater tolerance for deploying models to aid inference – or at least to make explicit the model we already have when we gesture to broader implications of experimental findings.

References

1. Allport, G.W., 1954. *The Nature of Prejudice*. Addison-Wesley.
2. Bates, R. 2006. "Banerjee's approach might teach us more about impact but at the expense of larger matters." *Boston Review of Books*, July/Aug. 31(4).
3. Banerjee, A., Kumar, S., Pande, R. and Su, F., 2011. "Do informed voters make better choices? Experimental evidence from urban India." *Unpublished manuscript*.
4. Chassang, S., Miquel, P.I. and Snowberg, E., 2012. "Selective trials: A principal-agent approach to randomized controlled experiments." *American Economic Review*, 102(4), pp. 1279-1309.
5. Chattopadhyay R, Duflo E. 2004. "Women as policy makers: evidence from a randomized policy experiment in India." *Econometrica* 72 (Sep.), pp. 1409–1443
6. Dunning, T., Grossman, G., Humphreys, M., Hyde, S.D., McIntosh, C. and Nellis, G. eds., 2019. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*. Cambridge University Press.
7. Green D, Gerber A. 2002. "Reclaiming the experimental tradition in political science," in *Political Science: State of the Discipline*, eds. I Katznelson, H Milner, pp. 805–32. New York: W.W. Norton.
8. Gugerty, MK, Kremer M. 2008. "Outside funding and the dynamics of participation in community associations." *American Journal of Political Science* (July), pp. 585–602.
9. Horowitz, D.L. 1985. *Ethnic Groups in Conflict.* University of California Press.
10. Kasara, Kimuli. 2017. "Does Local Ethnic Segregation Lead to Violence? Evidence from Kenya," *Quarterly Journal of Political Science* 11 (4), pp. 441-470.
11. Miguel, E., Satyanath, S. and Sergenti, E. 2004. "Economic shocks and civil conflict: An instrumental variables approach." *Journal of political Economy*, *112*(4), pp.725-753.
12. Olken BA. 2007. "Monitoring corruption: evidence from a field experiment in Indonesia." *Journal of Political Economy*, 115, pp. 200–49.
13. Pearl, J., 2009. *Causality*. Cambridge university press.
14. Pearl, J. and Bareinboim, E., 2014. "External validity: From do-calculus to transportability across populations." *Statistical Science* 29 (4), pp. 579-595.
15. Scacco, A. and Warren, S., 2018. Can social contact reduce prejudice and discrimination? Evidence from a field experiment in Nigeria. *American Political Science Review*, *112*(3), pp.654-677.