Conducting Experiments in Multiple Contexts*

Graeme Blair, UCLA
Gwyneth McClendon, NYU

This version: December 20, 2019

Abstract

In an effort to assess the generalizability of treatment effects across contexts, scholars (or teams of scholars) are increasingly conducting experiments around the same research questions in multiple country and subnational contexts. In this chapter, we categorize recent and ongoing efforts to conduct cross-context experiments into three types: "uncoordinated," "coordinated, sequential," and "coordinated, simultaneous." We discuss some practical trade-offs across these types, arguing that coordinated cross-context designs offer the most promise for meta-analyses. We then draw attention to four areas in which the current approaches arguably *all* fall short in facilitating cumulative learning about treatment effects and treatment effect heterogeneity across contexts. We conclude by proposing some ways forward to continue improving our approach to learning about generalizability across contexts.

*We thank Jamie Druckman and Don Green for inviting us to develop this chapter and for constructive feedback. We are also grateful to Brigitte Seim, our discussant, and Erin Hartman for constructive comments that improved the chapter.

Many researchers conducting experiments care both about "descriptive causal inferences" (the effect of a given intervention on a given outcome in a given context) and about generalizability: whether those same effects apply to other sets of individuals, to other types of interventions and in other contexts.[1] In particular, researchers interested in informing and aiding policymakers typically see generalizability as a crucial goal. As Duflo and Kremer (2005) explain: "The benefits of knowing which programs work and which do not extends far beyond any program or agency, and credible impact evaluations are global public goods in the sense that they can offer reliable guidance to international organizations, governments, donors, and nongovernmental organizations (NGOs) beyond national borders" (pp. 1-2). Whether one particular intervention affected one particular outcome in one particular context among one particular set of individuals may not be relevant to the decisions practitioners need to make. Instead, practitioners are often interested in whether treatment effects in a given study would also "travel:" that is, carry over to other contexts and to other sets of individuals more relevant to policymakers' own decisions.

Practical learning to improve policymaking is not the only motivation for seeking generalizable causal inferences: the tight links between theory and generalizability are an increasing focus of discussion among experimentalists (see Humphreys and Wilke, 2019, for a review). Descriptive causal inferences may be used to build new theory, and theory may be used to guide how (and whether) generalizable claims beyond the sample can be made from an experiment (see Hartman in this volume, for a discussion of this role for theory). In the interest of providing guidance to policymakers, as well as in pursuit of improving intellectual understandings of causal relationships and their scope, many researchers conducting experiments want to answer questions about the generalizability of causal effect estimates.

Recent advances in the social and medical sciences have focused on identifying feasible, data-driven methods for approaching questions of generalizability using data from a single experiment. For instance, Hartman in this volume, drawing on Egami and Hartman (2018), outlines a way to estimate the average treatment effect in a target population (PATE), using only data collected from the experimental sample. Hartman et al. (2015) describe ways to estimate the average treatment effect among the treated in a target population (PATT) by combining data from observational studies of the target population with data from the experimental sample. Ratkovic in this volume provides guidance for data-driven ways to identify sources of treatment heterogeneity in experimental samples.

In this chapter, we review another set of research strategies that individual scientists and scientific communities have adopted to address issues of generalizability of findings. This set of meta-analytic approaches involves combining data from multiple experiments, some conducted in different institutional

[1] For a detailed discussion, see Hartman's chapter in this volume.

contexts,[2] and then conducting a meta-analysis of the pooled results. Meta-analysis is a method for formally summarizing effects across studies, for example by calculating average effects. One type involves *post-study* efforts to identify and pool data from similar experiments for meta-analysis. This type is the most common approach to date. Another approach involves purposefully *creating* a dataset for meta-analysis over time by sequentially repeating experiments in new places. Replications in new contexts of previously published studies fall in this second category. A third type — exemplified by the recent Metaketa initiative study organized by the Evidence in Governance and Politics (EGAP) network — involves creating a dataset for meta-analysis by coordinating research teams around similar interventions and measures that are then implemented in multiple institutional contexts quasi-simultaneously.

We use this chapter to describe and discuss these approaches to assessing and addressing the generalizability of experimental results. First, we categorize recent efforts to compile and/or conduct experiments in multiple contexts in order to address generalizability into the three types mentioned above: (1) an uncoordinated, (2) a coordinated-sequential, and (3) a coordinated-simultaneous approach.

Second, we identify practical and professional trade-offs across the three types that researchers and research communities should consider. In our view, the three current research strategies are complementary and are each likely to lead to different types of learning in the search for what works where. Coordinated-simultaneous research is likely to lead to the most generalizable insights but only on the effects of one or a small set of interventions. Coordinated-sequential research is less likely than coordinated-simultaneous research to facilitate meta-analysis because, for instance, the set of outcomes and treatments are less likely to be standardized. However, the approach may lead to improvements in designs over time and may lead to learning about what dimensions of the treatment are important given that the intervention will vary. Uncoordinated research is least likely to facilitate meta-analysis but can yield new interventions that affect new outcomes because of publishing and professional incentives that reward innovation. Each approach is likely to contribute to the accumulation of knowledge.

However, third, we then draw attention to four dimensions of cumulative learning along which *all three* current approaches fall short. Ultimately, we suggest that, despite some advantages to current approaches of multi-context experimentation, research going forward might achieve greater knowledge accumulation by doing more to leverage individual-level heterogeneity (rather than context heterogeneity), in conjunction with theory, to assess generalizability.

We have in mind three audiences for the following discussion. The first is senior researchers who are already at the forefront of efforts to coordinate experiments in multiple country contexts, or who are key

[2]We use "institution" in the way North (1991) does: as a set of humanly devised constraints that structure political, economic and social interactions. Multiple institutional contexts could mean multiple countries, districts or cities, or in the same context over different points in time as institutional features change.

players in compiling uncoordinated studies for meta-analysis and in evaluating experimental research for publication and promotion. We hope that the discussion in this chapter gives them greater insight into the trade-offs across multi-context experimental designs and into the shortcomings of current approaches. The second audience is practitioners and donors deciding which research efforts to incentivize and fund in order to further their goals of understanding whether and where interventions are likely to have particular effects. We hope that the following discussion generates ideas for incentivizing improved multi-context experimentation. The third audience is early-career researchers (graduate students, post-doctoral fellows, junior faculty) who are deciding whether and how to address concerns of generalizability in their own research. We hope that the following discussion leaves them optimistic about participating in multi-context experimentation. However, we emphasize that senior scholars and donors need to take the lead in making sure that early-career scholars are rewarded for these efforts.

Three Types of Research Programs

Meta-analysis is the formal synthesis of the findings from multiple studies on the same topic in order to characterize consensus (or lack thereof) in a literature. Studies are combined by summarizing study effect estimates, for example by taking the average, or by stacking datasets from multiple studies and analyzing them together. In the last decade, meta-analyses of experiments on common research questions in multiple contexts have grown increasingly common. These studies have varied in their levels of ex ante coordination: some have involved no ex ante coordination across research teams and simply involved later reanalysis of studies' data. Others have involved fairly extensive ex ante coordination that aimed to standardize interventions and measures. Among studies that have involved some form of ex ante coordination, there has also been variation in whether that coordination has involved a simultaneous, or a sequential, effort to coordinate. We describe and give examples of each of type of study below.

We begin by categorizing multi-context experiments into three predominant types:

1. *Uncoordinated experiments* ("let 1,000 flowers bloom"): This approach, the most common to-date we think, relies on researchers to congregate informally around particular research questions. Researchers may occasionally borrow aspects of others' designs and measures in order to make their work more comparable to existing studies in the field, but the approach does not typically involve standardization across studies or involve researchers' agreeing to common principles and approaches *ex ante*. Among our examples below, the contact hypothesis literature falls under this first type of cross-context experimentation (Paluck, Green and Green, 2018).

2. *Coordinated, sequential experiments*: This approach involves re-use of measurement or interventions (or both) across studies in different contexts *over time*, either by the same researcher or team of researchers or by different research teams seeking to replicate previously publicized findings in new contexts. Researchers move to a new context after the results of previous studies are known. For instance, a single researcher might build on her own experimental work over time, sequentially testing the generalizability of her previous results in new (country) contexts. A funder might ask researcher teams to incorporate measures and interventions from previously published studies. Among our examples below, efforts to replicate Butler and Broockman (2011)'s audit study in new contexts would fall under this type of cross-context experimentation (Costa, 2017).

3. *Coordinated, simultaneous experiments*: This approach involves different researchers (or teams of researchers) coordinating measurement or interventions (or both) *ex ante* before conducting experiments in different institutional contexts at roughly the same time. In other words, in contrast to the coordinated, sequential approach, this coordinated, simultaneous approach means not waiting for each team to analyze results before designing the follow-up experiments but instead involves attempting to design and run experiments with common measures and/or interventions in multiple institutional contexts before knowing what the results of any given study will be. The ex ante coordination in this approach may focus on standardizing the interventions, the outcomes, or contexts. In some cases the team may not coordinate the set of outcomes per se but rather coordinate the measurement strategies for those outcomes. When the outcome is a latent construct such as an attitude, often different *measurement strategies* might be used in different contexts to best represent that construct.[3] Among our examples below, the six-country ultra-poor graduation study (Banerjee et al., 2015) and the EGAP metaketa studies fall under this third type of cross-context experimentation.

Examples

In what follows, we describe applications of each of these three types of research programs.

An early example of a coordinated, simultaneous study in the social sciences, though not a randomized experiment, is Henrich et al. (2006). Researchers in 15 sites around the world conducted behavioral games with local populations.[4] Following the results of one study in one site (Machinguega), research teams gathered together at a three-day workshop to coordinate the structure of the games and the outcome

[3]This is a frequent occurrence when outcomes are measured through surveys in translation: direct, naive translations may lead to different concepts being measured. Varying translations that tap the concept as closely as possible may be preferred in these cases even though measurement differences will be confounded with context.

[4]See Candelo and Eckel in this volume for a longer discussion of lab-in-the-field studies.

measures before implementing studies in 14 other locations. Although there was some variation in how teams ultimately implemented the protocol in order to make sure each study was comprehensible to local populations, the studies overall implemented the same treatments and outcome measures. Two years later, the researchers reconvened to compare findings; they then conducted a meta-analysis of the ultimatum game, which had been conducted in all sites and in the same way. The meta-analysis served to highlight consistent deviations from "homoeconomicus" behavior around the world. The authors also argued that the meta-analysis shed light on some of the contextual variables (e.g., a society's degree of market integration) that might produce heterogeneity in ultimatum game behavior.

In a similar vein, the Evidence in Governance and Politics (EGAP) network recently conducted coordinated, quasi-simultaneous experiments in six different countries in order to estimate the effect of politician performance information on voting behavior.[5] Teams of researchers, overseen by a core group of principal investigators, gathered together to coordinate a standard informational intervention (a "common arm") as well as a common set of outcome measures of voter attitudes and behaviors across the countries. After the studies were completed, the principal investigators produced a meta-analysis of the results of the common arm (Dunning et al., 2019a). The authors used the meta-analysis to underscore the limited impact of simple non-partisan, credible information on voter behavior even across a wide variety of contexts (Dunning, 2019b). EGAP is currently conducting similar coordinated-simultaneous experiments to assess the generalizability of treatment effects for other topics (on taxation, natural resource governance, community policing, and women's participation in authoritarian regimes).

In another example of a coordinated-simultaneous study, Banerjee et al. (2015) coordinated RCTs in six countries (Ethiopia, Ghana, Honduras, India, Pakistan, and Peru) around a common anti-poverty intervention. Based on a program that had been designed and run previously in Bangladesh, the experiments selected similar study samples in each country (the poorest of the poor in a set of the poorest villages) and assigned households and/or villages to treatment or control. The treatment involved giving poor households a grant, training and coaching, consumption support and often health care and access to savings accounts; it aimed to increase participant welfare and lead to sustainable self-employment. The authors used their meta-analysis to underscore the effectiveness of the program in improving welfare outcomes, even across a wide variety of institutional contexts.

Other examples of cross-context experiments have also involved explicit ex ante coordination of interventions and measures but through *over-time* efforts at replication in new contexts rather than through a large-scale simultaneous set of projects. We call these coordinated, sequential experiments. For instance,

[5] A seventh experiment was planned for but not conducted due to implementation challenges in the study site. Enumerators were kidnapped and held by politicians after the first part of the information treatment was distributed in one community.

following the publication of Butler and Broockman (2011)'s audit study of United States state representatives' responsiveness to inquiries from constituents of different race groups, several other researchers (and research teams) borrowed elements of Butler and Broockman's design in order to test whether the same treatment effects would travel to other countries, or to other bureaucratic institutions in the United States.[6] In many of these studies, emails are sent to representatives and the key outcome is whether the representative responds to the email. After identifying and compiling data from these experiments, Costa (2017) conducted a meta-analysis of audit experiments of elected officials' responsiveness to constituents, which included 41 experiments conducted between 2011 and 2017 in four countries (the United States, Germany, South Africa and China) and the European Union, all borrowing elements of the Butler and Broockman design. There are also instances in which the same researcher, or team of researchers, has conducted similar experiments in different institutional contexts during the course of his or her career, often reusing elements of the same interventions and similar outcome measures across studies over time. For instance, in her research program on how perceptions of social norms shape behavior (Tankard and Paluck, 2016), Paluck has conducted a series of experiments in diverse contexts (Rwanda, the Democratic Republic of the Congo, Nigeria, South Sudan, the United States) in which study participants were exposed to media programs targeted at shifting, among other factors, perceptions of social norms (Paluck and Green, 2009; Paluck, 2010; Paluck, Blair and Vexler, 2010; Trujillo and Paluck, 2012; Blair, Littman and Paluck, 2019).

Far more commonly, at least to date, researchers pursue their own research agendas more or less independently, typically introducing novel interventions and measures rather than explicitly repeating interventions and measures used by others or coordinating with others, even when working on the same research questions. *Ex post*, scholars then attempt to identify a set of studies that are similar enough for meta-analysis. For instance, building on Pettigrew and Tropp (2006), Paluck, Green and Green (2018) compiled results from 27 independent experiments conducted in 8 countries (the US, Norway, India, South Africa, Germany, Nigeria and Israel) aimed at testing "the contact hypothesis," the idea that members of different social identity groups will exhibit lower levels of inter-group prejudice if they spend time with one another. The meta-analysis looked at sets of experiments that were conducted in multiple contexts but that researchers did not coordinate ex ante or conduct in such a way as to try to replicate the interventions and measures of previously published studies. The analysis revealed mixed results.

[6]See Butler and Crabtree's chapter and Nathan and White's chapter in this volume for discussions of these and other audit studies.

Trade-offs

In our view, choosing among the three approaches above involves practical and ethical trade-offs.[7] In particular, because they differ in levels of coordination, cost, and timing constraints, the three approaches to generalizable learning each present different practical challenges. These include (1) ensuring sufficient standardization of outcomes and interventions; (2) allowing data-sharing for later meta-analysis; (3) reconciling the goal of knowledge accumulation with the need to publish research in time for professional advancement and to influence policymaking; (4) reconciling the goal of knowledge accumulation with professional incentives; (5) allowing for cost-effectiveness assessments; (6) creating economies of scale in the research itself; and (7) ensuring equity within research communities. Table 1 summarizes our assessment of these trade-offs.

Table 1: Trade-offs Across the Three Types of Multi-Context Approaches

	Uncoordinated	Coordinated, Sequential	Coordinated, Simultaneous
Enables Standardization of Treatments/Measures?	No	Maybe	**Yes**
Ensures Data Available for Meta-Analysis?	No	Maybe	**Yes**
Yields Answers Quickly?	Maybe	No	**Yes**
Ensures Incentive-Compatibility for Individual Researchers?	**Yes**	Maybe	Maybe
Allows for Interim Cost-Effectiveness Evaluations?	**Yes**	**Yes**	No
Creates Economies of Scale?	No	Maybe	**Yes**
Ensures Some Degree of Equity Among Researchers?	**Yes**	Maybe	Maybe

Let's discuss the challenges for meta-analyses first. Meta-analyses demand (1) some consistency in treatments, measures, and inferential targets[8] across the studies they pool, as well as (2) access to the data from those studies. Researchers undertaking coordinated, simultaneous multi-context experimentation typically design the studies with these demands of meta-analysis in mind *ex ante*. Indeed, the focus of much of the early stages of coordinated, simultaneous experimentation is on corralling research teams to use as close to identical treatments and measures (or measurement strategies) as possible, even as individual research teams seek to ensure that their designs involve contextually appropriate, meaningful, and ethical interventions and measurements. Coordinated, simultaneous multi-context experiments rarely, and perhaps never, achieve perfect consistency across studies, and the task of negotiating comparability across interventions and measures takes time and effort. Yet, coordinated, simultaneous multi-context experiments go further in ensuring standardization and data availability than the other two current approaches. In Henrich et al. (2006), all research teams agreed to conduct the ultimatum game anony-

[7] We assume that researchers *want* to assess the generalizability of a treatment effect across contexts, thus setting aside questions about whether the pursuit of this form of generalizability is advisable or worthwhile.

[8] Even if treatments and measures are comparable, if one study only reports effects for one group and the second reports effects for another, they cannot be usefully combined in a meta-analysis unless no heterogeneous effects across groups are expected.

mously with comparably high stakes, such that treatments were very similar at that level of generality even as other words used to frame the game differed slightly from context to context. This approach facilitated a relatively straightforward protocol for coding treatments and outcomes and pooling results for the meta-analysis. In Dunning et al. (2019a), research teams implemented a common treatment, in which information about politicians past performance was distributed to voters, and also pre-registered a common set of estimands and outcome measures. The aim was to facilitate a later meta-analysis on the consequences of that common arm. In all of the cases of coordinated, simultaneous multi-context experiments of which we are aware, researchers have also been required to share their data publicly as a condition of participating in the coordinated effort.

Coordinated, sequential multi-context experimentation can also facilitate meta-analysis *if* later studies mimic earlier designs closely and also target the same estimands as the first studies. For instance, because all the audit studies in Costa (2017) used similar types of constituent-elite messages and measured whether elites responded to the messages they received (as a binary measure), Costa (2017) could straightforwardly conduct a meta-analysis of the main results of those papers. By contrast, other types of responsiveness (e.g., the quality/content of the responses) were not measured or were not reported in all studies, so she could not conduct a meta-analysis of those outcomes with the same geographic breadth or statistical power. Further, the coordinated, sequential approach does not necessarily guarantee that data will be made available by all researchers. If one researcher (or research team) conducts the studies over time, then she or they will presumably have access to all relevant data, thus facilitating a meta-analysis. But if many researchers and research teams conduct the replications in different contexts, gathering the data may require the meta-analyst to reach out to researchers, who may or may not be responsive to her requests. See Boudreau in this volume for a discussion of other reasons to encourage data sharing as part of research transparency.

Finally, meta-analysts of studies from the uncoordinated experiments approach typically have the most challenging work to do. In describing their meta-analysis of uncoordinated contact hypotheses experiments, Paluck, Green and Green (2018) note that the participants, interventions, outcome measures and estimators all varied considerably across the studies they identified. The authors had to make difficult decisions about what to pool and analyze, and ended up dropping studies that did not provide enough information or were not similar enough to the others to facilitate the meta-analysis. Data and code sharing issues may also be most pronounced for this type of research.

The coordinated, simultaneous approach is thus likely to be most conducive to subsequent meta-analysis compared to the other two strategies. This is not to say that coordination over treatments, measures, and estimands is perfect in these studies. In the EGAP/Metaketa I on the influence of politician

performance information on voter behavior (Dunning et al., 2019a), teams did not introduce identical types of information in each country, nor did they do so in exactly the same format. Instead, each team introduced politician performance information that was specific to the type of politician running in that country's election (mayoral, legislative, etc) and introduced politician performance information that was specific to the formal duties of the politicians in competition. Each team also introduced the information in a format (video, flyer, oral presentation, scorecard) that was familiar and comprehensible to local study participants. Likewise, in Banerjee et al. (2015)'s set of experiments, study participants could choose which productive asset they wanted transferred to them and, depending on the country, were offered different options for saving money after the interventions. In some countries, study participants deposited savings in banks or other formal financial institutions; in other countries, study participants were offered savings groups or opportunities to save through the study meetings. Thus, although the interventions were the same at a high level of generalizability (e.g., all interventions in Dunning et al. (2019a) involved relevant, comprehensible, non-partisan information about politician performance, and all interventions in Banerjee et al. (2015) involved the transfer of desired productive assets and opportunities to save), they were not identical at lower levels of generality (Sartori, 1970). Nevertheless, compared to the consistency typically achieved in the other two current types of cross-context experimentation, the level of coordination in coordinated-simultaneous multi-context experiments is likely to lend itself the most toward later meta-analysis.

There are also other advantages to the coordinated, simultaneous cross-context design. Sometimes, learning quickly is a goal for researchers and practitioners. If researchers want to inform pressing policy debates, they may seek to amass knowledge about the generalizability of particular treatment effects within a short amount of time. Even if the approach involves intensive planning stages, the coordinated, simultaneous approach may be best positioned to facilitate speedy learning by yielding a number of studies around a given intervention in quasi-simultaneous fashion. EGAP's Metaketa IV, involving experiments around community policing in six countries over the course of four years, is an example of a coordinated, simultaneous effort to generate policy-relevant evidence more quickly than the decades it might take for scholars to independently congregate around the same set of questions. The studies in Paluck, Green and Green (2018), for example, span six decades. The sequential, coordinated approach, by design, yields inferences about generalizability over longer periods of time and proceeds only as quickly as a given research team is inclined to and able to acquire local knowledge and funding to move across different contexts, or only as quickly as other researchers take up the interventions and measures in previously published work and implement them in a new context. Unless the initial study happens to take place in the country where policymakers are most urgently seeking answers, the speed of learning under

the coordinated, sequential approach may be slow. Relatively speaking, the coordinated-simultaneous approach promises faster learning about particular interventions than the other two approaches.

However, as can be seeing in Table 1, there are also downsides to coordinated, simultaneous designs that researchers should take seriously and that require modification to those designs. Consider professional incentives. Career advancement for many researchers is dependent on publishing. If publication is at all tied to novelty and individual innovation,[9] then efforts to standardize interventions and outcome measures through coordinated, cross-context experimentation may cut against the professional incentives of individual researchers and research teams. True, coordinated, simultaneous designs might reduce the risks to individual researchers of not being able to publish null results (Gerber, Green and Nickerson, 2001; Franco, Malhotra and Simonovits, 2014; Dunning et al., 2019a, see also Malhotra in this volume), if the whole group succeeds in placing a publication that provides a more highly powered design.[10] But researchers may nevertheless hesitate to participate in a coordinated, simultaneous approach if they fear that they will get little professional credit for conducting only one study in a larger endeavor, or if they fear "reviewer fatigue" whereby outside journal or book reviewers do not see much contribution from papers that bear resemblance to others they have been asked to review. Without modifications, coordinated, simultaneous multi-context designs may provide insufficient selective incentives to individual researchers to contribute the public good of knowledge about generalizability. This is because authorship norms in many social sciences currently only weakly reward contributions to publications with many authors.

The other two approaches offer more selective incentives to individual researchers. A coordinated, sequential experimental program, if conducted by a single research team over time, can keep the novelty of the design and measures pegged to an individual researcher (or research lab), and thus reserve credit for those author(s). Under the uncoordinated approach, researchers can design interventions and measures however they want in order to demonstrate innovation and impress journal and book reviewers: they are not constrained by needing to repeat and standardize interventions and measurements across studies. The downside, as noted above, is that researchers' prioritization of novelty, rather than of coordinated treatments and measures, may make later meta-analysis very difficult.

A better way forward might be to design coordinated, simultaneous cross-context experiments with an eye toward incentivizing researchers to contribute to the public good of knowledge about generalizability. This might mean incorporating ways for individual researchers to claim credit for innovation even while coordinating with other researchers. For example, EGAP's Metaketa I (Dunning et al., 2019a; Dunning,

[9]Note that we are not arguing that rewards for novelty are a "good" thing but rather that they are a current feature of the discipline. See also Malhotra in this volume on incentives to published "splashy" results.

[10]Coordinated, simultaneous experiments may also allow early-career researchers to gain access to a larger pool of resources and to receive feedback on their designs from the group.

2019b) allowed individual research teams to introduce alternative treatment arms in their studies in addition to implementing the common treatment. All aspects of the designs (both the common arm and the alternative arms) were paid for by the same funder. In this way, each research team could demonstrate an ability to innovate (see e.g., Adida et al., 2017; Arias et al., 2018, in which authors describe alternative arm interventions that were independent of the coordinated common treatment arm)[11] while at the same time facilitating meta-analysis of the effects of the common treatment. A more radical change would be to have the discipline shift from rewarding novelty to rewarding scholars (at all career stages) for contributions to coordinated, sequential and coordinated, simultaneous cross-context research. This more radical change would mean having donors, journal editors, journal reviewers, search committees members and tenure letter writers look favorably upon efforts to replicate previously published designs in new contexts as well as upon efforts to participate in coordinated studies that use similar interventions and outcome measures in order to gauge generalizability across contexts. (See also Maholtra in this volume for discussions of shifting incentives in the discipline.)

These points raise the issues of costs and equity of access to conducting research valued by the discipline. Any of the three cross-context experimental designs can involve significant financial costs in the aggregate. The seven studies in the EGAP Metaketa I initiative each received $175,000-300,000, for a total cost of about $1 million US dollars across all studies. The uncoordinated approach could easily involve that much money as well, after adding up the cost of all studies around a set of research questions. The difference across the approaches thus may not be in their total costs but instead in (a) how feasible it is to conduct interim cost-effectiveness evaluations, (b) whether there are economies of scale, and (c) in whether there are dangers of exacerbating inequities within the discipline. The uncoordinated approach and coordinated-sequential approach leave space for donors (and researchers) to conduct cost-effectiveness evaluations after each individual study. In some cases this may cause further research to stop if the intervention is not cost effective. By contrast, the coordinated, simultaneous approach is more likely, by design, to expend the total sum first, leaving no space for cost-effectiveness evaluations after each study. Yet, the coordinated, simultaneous approach—though less conducive to interim cost-effectiveness evaluations—leverages economics of scale by allowing some research-related costs (e.g., the costs of checking and replicating results, the costs of hosting data repositories, the costs of refining treatments and outcome measures) to be shared across research teams. Each of the three current design types offers either the possibility for interim cost-effectiveness evaluations, economies of scale, or possibly both (see Table 1).

[11] At the time of this writing, four out of the six teams who conducted RCTs as part of the Metaketa I had published alternative arm or alternative measurement strategy papers in peer-reviewed journals.

However, *if* the discipline made coordinated, simultaneous cross-context experimentation the primary approach to assessing cross-context generalizability (because of its advantages in meta-analysis, speed and economies of scale), the approach might be prone to excluding scholars in the field with lower status or fewer resources, raising concerns of equity of access. In the coordinated, simultaneous approach (e.g., Dunning et al. (2019a), Banerjee et al. (2015), Henrich et al. (2006)), project leaders often recruit or select research teams to be part of the large, coordinated effort. Any such screening process is vulnerable to excluding scholars who are not as well networked: less-well-networked scholars might not learn about the opportunity or might not be selected because they are at a lower status institution, have a less prestigious pedigree, or the like. By contrast, in the uncoordinated or coordinated, sequential approaches, lower status, less-well-resourced scholars can find ways to conduct a new study or a replication on a lower budget. They can still join the effort, without a screening process that might be vulnerable to excluding scholars on the basis of status or resources. Costa (2017)'s meta-analysis exhibits this phenomenon well: many of the replications of Butler and Broockman (2011)'s audit study were done by early career scholars on modest budgets.

As with professional incentives, the best way forward in light of these trade-offs might be a shift in disciplinary norms (see also Malhotra in this volume). Leaders of coordinated, simultaneous cross-context experiments might want to insist on an open call for applications; and they may want to structure their selection process to weed out as many opportunities for status-related biases (e.g., by reading proposals blind to the identities of the authors). As in our discussion of professional incentives above, veto players (journal editors, journal reviewers, search committees, and tenure letter writers) could also seek to reward lower-cost efforts to replicate previous findings in new contexts, sensitive to the reality that not all scholars may have a chance to participate in coordinated, simultaneous cross-context experimentation.

In short, for the goal of assessing the generalizability across contexts of one or a few particular interventions, the coordinated, simultaneous approach seems the most promising current option, based on our list of trade-offs in Table 1. The coordinated, simultaneous approach lends itself best to meta-analysis. It presents some downsides in terms of professional incentives and concerns about equity within research communities, but these downsides could be mitigated if combined with efforts to make opportunities more equitably available and costs shared among a wide range of researchers *and* with shifts in disciplinary norms: making participation in standardization of treatments/measures and coordination more highly valued.

Nevertheless, the three current approaches are largely complementary in our view. As noted above, the coordinated-sequential approach *can* lead to standardized interventions and measures. It also offers more opportunities to less-well-networked researchers (under current practices and disciplinary norms)

and might lead to improvements in the quality of designs over time. And, because the uncoordinated approach offers incentives to researchers to innovate, it might lead to more learning about treatment effects of *many* interventions (albeit in specific contexts), whereas the coordinated, simultaneous approach lends itself best to learning about the generalizability of one or a few particular interventions across contexts. Each approach thus contributes to knowledge accumulation in different ways. Moreover, as we argue in the next section, all three approaches also fall short on some dimensions of learning about generalizability.

Shortcomings in All Current Approaches

We have imagined that researchers want to choose among the three current approaches to cross-context experimentation, and we have described some trade-offs among three current approaches that researchers might want to consider. But there are also ways in which *all* of the current approaches fall short of addressing the central goal of cross-context experimentation: to learn what works and in what contexts.

We identify four issues that all of the current approaches leave more or less unaddressed in pursuit of this goal. The first issue is **how contexts are selected**. In all three current approaches, contexts are often not selected in such a way as to *facilitate* assessments of generalizability. Instead, scholars usually first select contexts on the basis of researcher characteristics (such as language proficiency, personal connections, and the choices of academic advisors and collaborators). Conditional on the context, interventions are then designed and/or selected. This process has implications for how much anyone can learn from cross-context experimentation. There are unobservable (or at least, typically unobserved) features about the country that led to its selection, and so it is difficult to extrapolate directly to other contexts that do not share those possibly unobserved features. For example, in part due to the personal choices of a small set of researchers, many early development economics experiments were conducted specifically in Busia District in Kenya (e.g., Miguel and Kremer, 2004; Duflo, Kremer and Robinson, 2011). Whether we can generalize from those findings to other similar contexts may be confounded by the factors that went into the original choice to work in Busia. In pursuit of learning about the generalizability of treatment effects, researchers would do well to select sites more purposefully: either randomly (and thus independent of factors such as personal connections) or with the aid of theories about likely treatment moderators. (See also Hartman in this volume for a discussion of theory about likely treatment moderators.)

The second, and related, issue is **how treatments are assigned to contexts**. In current approaches, the choice of which intervention to study in a given context is often neither randomized nor paired with a site in order to explicitly test theorized treatment moderators. Instead, researchers often pair interventions and sites on the basis of the researcher's intuition about what will be *most* effective in that context. This

selection process may have implications for the extent to which anyone can learn from cross-context experimentation because at the extreme, interventions researchers believe are less effective in a given context will never be put to the test. Whether researcher expectations are right or wrong, one set of context-intervention pairs is never tested, thus making it harder to learn about what works best in each context.

The third issue is **statistical power**. The stated goals of experimentalists who conduct cross-context experiments is to learn about variation in the effectiveness of an intervention across *contexts* rather than across units or interventions (see Hartman in this volume for a discussion of generalizability across units). Yet the three current approaches described above typically involve a small number of contexts (e.g., countries) relative to the number of interventions evaluated. Any extrapolation from the experimental contexts to others is thus typically highly uncertain. Each context represents a bundle of characteristics, for example in the case of a country its regime type, region, rate of voter turnout, history of colonialism, etc. In current practice, coordinated, simultaneous experiments, which, as we describe above, offer the most promise for meta-analyses, are conducted in half a dozen or a dozen cases *at most*. This yields very little leverage to make predictions about variation in effectiveness across contexts, because with so few cases there will be insufficient variation in the country-level characteristics that we bundle together as "the context." We might be comparing two democracies to one autocracy and using those small number of cases to extrapolate to another case that is a democracy. When we consider a bundle of characteristics, we end up slicing the data so finely that there may be at most one case with that set of characteristics.

The fourth issue is the challenge of **evaluating not just whether a given intervention works in a given context but also whether it works "best" *relative* to other feasible interventions**. In order to have evidence about what is most effective in each context, we need to have within-context evidence about what works from all possible interventions under consideration by the policy community. Instead, current approaches typically confound interventions and contexts. Researchers select interventions to suit a given context and often test only a single intervention (or very few) in that context. We thus do not obtain variation that would help tailor our generalizations to new contexts across interventions. Moreover, internally-valid comparisons can be made between arms from a multiarm trial but not by comparing different arms in multiple experiments in different contexts.

How could we further adapt current practice to improve our handling of these four issues? One set of possibilities is design-driven. In order to developing our thinking in this direction, we consider the logic of a general method for experimentation that directly optimizes learning about what works best for which contexts: *multi-arm bandit* algorithms (for an introduction in political science, see Offer-Westort, Coppock and Green, 2018) and their cousin the *contextual bandits* (Langford and Zhang, 2007). These

procedures were developed by computer scientists for implementation as a single experiment to learn about which of a set of interventions is most effective for different types of individuals. We take these procedures as a point of departure for imagining an experimental design implemented across multiple contexts that might more effectively address the four issues raised above.

The *multi-armed bandit* aims to identify, through multiple periods of a randomized trial in the same context, which of a set of several interventions (say M of them) have the largest treatment effect. In the first time period, a first set of experimental units are randomized into M + 1 groups representing the M treatment and a control group with equal probability. From that first round, M treatment effects are estimated comparing each of the groups to control. In the second period, a new set of 100 experimental units are also assigned to M + 1 groups but not with equal probability. Instead, the probabilities are assigned as a function of the estimated treatment effects of each arm, called the payoffs for each arm. The higher the payoff (estimated treatment effect), the higher the probability of assignment. The second period experiment runs, and researchers estimate another set of treatment effect estimates and use those estimates to determine the assignment probabilities in the third round. For a treatment that is wholly ineffective, the probability of assigning units to that treatment will quickly go to zero, and the experimental sample is reallocated to learn about other possibly-effective treatments. The algorithm can be stopped once the probability of assignment to one treatment is sufficiently high, reflecting confidence it is the most effective.

The multi-arm bandit problem can also incorporate information about the characteristics of experimental subjects in order to address the question of what works across contexts. This approach is *the contextual bandit* (Langford and Zhang, 2007).[12] The standard multi-armed bandit research design learns what works best on average, but it does not tell us what works best *for a given experimental subject.* The contextual bandit, thus, can be used to predict treatment effects across settings that have different mixes of types of individuals in them. In the contextual bandit design, treatment probabilities are assigned not only as a function of the average effect estimate for a given intervention but also as a function of the conditional average treatment effect for that intervention in a given context. Instead of assigning a single payoff to a treatment arm, which assumes effectiveness to be constant across individuals and contexts, the payoff is modeled as a function of individual and context characteristics. Payoffs might be a function of age, gender, and occupation (individual characteristics), as well as urban/rural status and regime type (contextual characteristics). This is accomplished in different ways but often with regularized regression, which estimates regression with many predictors but constrains some coefficients to be zero to avoid

[12]Standard contextual bandit designs differ from our setting in that it focuses on a Bernoulli trial in which individuals with observed characteristics are randomized one-by-one in sequence. We thus consider it as an analogy for a context-specific multi-armed bandit design that we could notionally implement in an experimental literature.

overfitting. As a result of this procedure, when probabilities of assignment change in subsequent rounds, they are a function of how effective the intervention is *for similar contexts*. In each round, the algorithm is learning both about average effectiveness and differences in effectiveness across types of individuals and contexts.

As a thought experiment, we consider whether an adaptation of the contextual multi-armed bandit could address the four issues we raised above. The selection of cases, assignment of a menu of interventions to each case, and assignment of interventions to individuals within each case would each be systematized in order to facilitate learning about what works in what context. A hypothetical principal investigator, or team of principal investigators, would randomly sample contexts (such as countries) from the universe of contexts she wishes to extrapolate to. This would improve extrapolations from the study experiments to others. From the set of M interventions, each case would be randomly assigned a subset of those interventions to test (1 to M of them) as well as a control group. This would avoid any confounding bias of linking context and intervention. Within each study of a context, individuals would be randomly assigned to one of the interventions selected for the context or to the control group, with individual assignment probabilities a function of estimated conditional average treatment effects given individual and context characteristics based on past experimentation. This approach would improve precision, because more than one intervention would be tested in each site, and one would more efficiently learn about which intervention was most effective in which context by tying assignment probabilities to past evidence on conditional average treatment effects. To be clear, we are not suggesting centralizing control of interventions, outcomes, and contexts from researchers, but rather considering as a thought experiment what a single planner would do as a way to rethink how knowledge accumulates — and how we could change professional incentives to make it accumulate more quickly.

The idealized design we describe would still fall short of facilitating learning about treatment effect heterogeneity across what Hotz, Imbens and Mortimer (2005) call "macro variables"—factors that are invariant within cases but vary across them (such as country regime type). The design would still be underpowered to assess treatment effect heterogeneity across these factors, given the relatively small number of sites with variation in these macro features. However, by optimizing learning about heterogeneous treatment effects within each context, researchers could use the approach to produce average predictions for other contexts. For example, democracies contain more wealthy people and more wealthy districts than autocracies. If we learn about treatment effect heterogeneity according to the income of individuals and according to the average income of districts individuals live in, then we might make predictions at the country level on the basis of those effects, weighting by the number of people in each context that are wealthy (poor) and living in wealthy (poor) districts.

In this idealized design, generalization could follow the model described in Hartman in this volume, rather than focusing on differences in average effects across contexts. Adjustments to the sample data would be made according to theory or evidence about the set of features related to differences in treatment effects. As the Hartman chapter highlights, it is crucial to collect information on the sampling set (the set of features related to inclusion in the experimental sample) and the heterogeneity set (the set of features related to differences in treatment effects), which together are used to adjust sample data to match the target context. Moreover, it is important to collect data on the features of the target population, to enable extrapolation and the assessment of the ability to extrapolate. Other design considerations, such as selecting contexts that represent the full set of types of individuals that are present in the target population, are also relevant to this idealized design. For example, even if contexts were randomly sampled, if an unlucky draw led to no contexts with religious minority groups in them, the fact that coordinated experiments were conducted in all of the sample contexts would not yield data that allow us to generalize to any context with religious minority groups. Stratified sampling could be used to mitigate these types of risks.

However, our suggestion to implement an adaptation of the contextual multi-arm bandit may be infeasible for various reasons. For one, research teams with knowledge and skills needed to implement experiments may not be available in all the relevant contexts with a given time period. For another, research teams may resist being subject to the random assignment of interventions to study sites or to choosing study sites in ways related to theories about treatment effect heterogeneity rather than based on their personal connections to particular contexts.

So, are there alternative ways forward that adopt *some* of the advantages of the bandit design into existing practice? We offer a few ideas. Researchers, research communities, and funders have a small set of levers that can shape the research designs and interventions used across different contexts. Research communities, through professional associations, can shape what studies are conducted to incentivize improving generalizability, keeping in mind the concerns raised about professional incentives and equity within the research community raised earlier in this chapter. Funders could do the same by shaping funding proposals or directly funding studies using designs listed below.

Below are several possible types of designs and linked studies that research communities and funders could incentivize in order to incorporate elements of the contextual multi-armed bandit design into cross-context experimentation:

1. Multi-arm designs in a single context with the aim of learning about the effectiveness of more than one treatment in that context. For example, in the Metaketa I study, a second intervention

was implemented in each country in addition to the common coordinated intervention. The use of at least two interventions in each context provided evidence of what works best of the two interventions in that context, as well as internally-valid comparisons of how the effect sizes differ. The implementation of multiple interventions in each context, using theory-based or random assignment of interventions to the selected contexts, could be implemented as part of coordinated or uncoordinated designs.

2. Replications of existing findings simultaneously in both a new context and again in a context that has already received the intervention.[13] In contrast to the coordinated-sequential model where generalizability is sought by trying the intervention in new contexts one after another, replicating studies in both new *and* repeat contexts would more closely approximate elements of an adapted contextual multi-arm bandit. For instance, one might repeat audit interventions again in the United States, re-weighting interventions for what worked on whom in the previous study, *and* conduct replications in other countries. Repeating the experiments in the United States would also contribute to our understanding of whether effects are generalizable over time.

3. Designs that are powered to detect *individual-level* treatment effect heterogeneity. Designs could be powered to detect differences in effects according to individual characteristics such as age, gender, or civic engagement score; or institutional features of towns or districts in which individuals live, such as whether leaders are elected or the quality of primary schools. Theory may play an important role here, in identifying potential sources of heterogeneity.

4. Designs that, when selecting interventions to study, incorporate priors from previous experiments about conditional effectiveness and match those to the characteristics of the new context. For example, when selecting from voter mobilization strategies for a new study on increasing turnout in Nigeria, such designs would look not only at average treatment effects of voter mobilization interventions from previous experiments but would also look at the characteristics of voters in Nigeria and estimate from previous experiments the predicted conditional average treatment effect based on the characteristics of voters Nigeria. This might involve observing that there is a large youth bulge in Nigeria and then using the fact that certain interventions were especially effective among young voters in previous experiments to select interventions for the Nigeria study. (See Hartman this volume for a discussion of projection or reweighting methods for generalizing from existing studies.)

[13]The definition of replication is contested. Some label replication in a new context a form of "conceptual replication" (see Crandall and Sherman, 2016).

In addition to these data-driven approaches, there has been a resurgence in thinking about the role of theory in experimentation. The design steps we have suggested in this chapter should lead to improved inferences about the relative effectiveness of some interventions. But we will not learn "what works best in what context" in a larger sense in finite time, given the vast number of interventions, outcomes, and contexts. To illustrate the problem from one relatively large and mature literature, Vivalt (Forthcoming) examines over 15,000 coefficient estimates from 635 papers reporting on 20 types of development economics interventions. Remarkably, the author found only 307 coefficients on outcome-intervention pairs that are shared by at least one other estimate in the dataset. From the rest of the coefficients, we learn little about how generalizable the findings are because we have only one estimate of the effect of the outcome-intervention pair. Without researcher coordination on a small set of outcomes and interventions to focus on, it would take incalculable time and resources to create a dataset of comparable estimates within and across contexts to learn, in a large sense, what works best in what context. And, in order to extrapolate from any existing corpus of experiments to a new context or new variation of an intervention, we must be guided in some way by theory about how that new context relates to those in our corpus. See Humphreys and Wilke (2019) for four roles for theory in experimentation with generalizability in mind, including guiding inferences by indicating the variables needed to extrapolate findings and guiding design decisions such as site selection.

In sum, the goal of assessing generalizability across contexts is a real challenge, and one difficult to overcome even in a single, large-scale, multi-site experiment with a multi-armed bandit design. We would encourage researchers to combine improved data-driven approaches with the development of theory in order to move us forward. In designing multi-context experimentation going forward, we would in particular encourage researchers and donors to think more about how to address the four issues we have identified above (the selection of contexts to study, the assignment of interventions to contexts, power, and learning about the comparative effectiveness of interventions in a single context) in conceiving and incentivizing new varieties of cross-context designs.

Conclusion

Experimentalists often care both about understanding the effects of a particular intervention in a single context and about the generalizability of those findings to other contexts. Considerable attention has been paid to the question of how to use data-driven approaches to extrapolate from a single experiment to new units and interventions (see Hartman 2019 in this volume for a detailed review). Recently, there has been increased attention to a set of complementary approaches to learning about generalizability across

contexts, such that we might be able to make predictions about treatment effects in new contexts from a *set* of experiments. In some recent examples (Dunning, 2019b; Banerjee et al., 2015), research teams have joined together to mount coordinated experiments quasi-simultaneously in six or more different countries, in order to quickly draw inferences about how effects of a single intervention vary across contexts.

In this chapter, we reviewed these recent developments and their practical trade-offs with other approaches to accumulating knowledge about treatment effects across contexts. We drew attention to the advantages of coordinated, simultaneous multi-context experiments, particularly in terms of their ability to facilitate meta-analysis (through standardized interventions and measures and through required data transparency), their relative speediness, and their economies of scale. But we also highlighted ways these designs might be modified to align with professional incentives and to ensure inclusivity within the discipline. We furthermore highlighted four areas in which all current political science approaches to assessing generalizability across contexts arguably fall short: in selecting study sites, in assigning interventions to study sites, in ensuring high enough statistical precision, and in evaluating which intervention works best. We then engaged in a thought experiment about an ideal multi-context design and proposed several concrete steps that scholarly communities and donors might pursue going forward, including conducting more multi-arm trials in single contexts and leveraging findings about heterogeneous treatment effects across *individuals* in order to learn about likely treatment heterogeneity across contexts. Knowledge about generalizability of treatment effects across contexts is accumulating, but there is still more we could learn from our current efforts.

References

Adida, Claire, Jessica Gottlieb, Eric Kramon, Gwyneth McClendon et al. 2017. "Reducing or Reinforcing In-Group Preferences? An Experiment on Information and Ethnic Voting." *Quarterly Journal of Political Science* 12(4):437–77.

Arias, Eric, Horacio Larreguy, John Marshall and Pablo Querubin. 2018. Priors Rule: When do Malfeasance Revelations Help or Hurt Incumbent Parties? Technical report National Bureau of Economic Research.

Banerjee, Abhijit, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Parienté, Jeremy Shapiro, Bram Thuysbaert and Christopher Udry. 2015. "A multifaceted program causes lasting progress for the very poor: Evidence from six countries." *Science* 348(6236):1260799.

Blair, Graeme, Rebecca Littman and Elizabeth Levy Paluck. 2019. "Motivating the adoption of new community-minded behaviors: An empirical test in Nigeria." *Science advances* 5(3):eaau5175.

Butler, Daniel M and David E Broockman. 2011. "Do politicians racially discriminate against constituents? A field experiment on state legislators." *American Journal of Political Science* 55(3):463–477.

Costa, Mia. 2017. "How responsive are political elites? A meta-analysis of experiments on public officials." *Journal of Experimental Political Science* 4(3):241–254.

Crandall, Christian S. and Jeffrey W. Sherman. 2016. "On the scientific superiority of conceptual replications for scientific progress." *Journal of Experimental Social Psychology* 66:93 – 99.

Duflo, Esther and Michael Kremer. 2005. Use of Randomization in the Evaluation of Development Effectiveness. In *Evaluating Development Effectiveness*, ed. O. Feinstein G. Ingram Pitman, G. New Brunswick, NJ: Transaction Publishers pp. 205–232.

Duflo, Esther, Michael Kremer and Jonathan Robinson. 2011. "Nudging farmers to use fertilizer: Theory and experimental evidence from Kenya." *American Economic Review* 101(6):2350–90.

Dunning, Thad et al. 2019b. "Voter information campaigns and political accountability: Cumulative findings from a preregistered meta-analysis of coordinated trials." *Science advances* 5(7):eaaw2612.

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 University Press.

Egami, Naoki and Erin Hartman. 2018. Covariate Selection for Generalizing Experimental Results. Technical report Working Paper.

Franco, Annie, Neil Malhotra and Gabor Simonovits. 2014. "Publication bias in the social sciences: Unlocking the file drawer." *Science* 345(6203):1502–1505.

Gerber, Alan S, Donald P Green and David Nickerson. 2001. "Testing for publication bias in political science." *Political Analysis* 9(4):385–392.

Hartman, Erin, Richard Grieve, Roland Ramsahai and Jasjeet S Sekhon. 2015. "From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178(3):757–778.

Henrich, Joseph, Richard McElreath, Abigail Barr, Jean Ensminger, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, Michael Gurven, Edwins Gwako, Natalie Henrich et al. 2006. "Costly punishment across human societies." *Science* 312(5781):1767–1770.

Hotz, V. Joseph, Guido W. Imbens and Julie H. Mortimer. 2005. "Predicting the efficacy of future training programs using past experiences at other locations." *Journal of Econometrics* 125(1-2):241–270.

Humphreys, Macartan and Anna Wilke. 2019. "Field Experiments, Theory, and External Validity." Working paper.

Langford, John and Tong Zhang. 2007. The epoch-greedy algorithm for contextual multi-armed bandits. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*. pp. 817–824.

Miguel, Edward and Michael Kremer. 2004. "Worms: identifying impacts on education and health in the presence of treatment externalities." *Econometrica* 72(1):159–217.

North, Douglass C. 1991. "Institutions." *Journal of economic perspectives* 5(1):97–112.

Offer-Westort, Molly, Alexander Coppock and Donald P. Green. 2018. "Adaptive Experimental Design: Prospects and Applications in Political Science." Working paper.

Paluck, Elizabeth Levy. 2010. "Is it better not to talk? Group polarization, extended contact, and perspective taking in eastern Democratic Republic of Congo." *Personality and Social Psychology Bulletin* 36(9):1170–1185.

Paluck, Elizabeth Levy and Donald P Green. 2009. "Deference, dissent, and dispute resolution: An experimental intervention using mass media to change norms and behavior in Rwanda." *American political Science review* 103(4):622–644.

Paluck, Elizabeth Levy, Graeme Blair and Daniel Vexler. 2010. "Entertaining, informing, and discussing: Behavioral effects of a democracy-building radio intervention in Southern Sudan." *Unpublished manuscript* .

Paluck, Elizabeth Levy, Seth A Green and Donald P Green. 2018. "The contact hypothesis re-evaluated." *Behavioural Public Policy* pp. 1–30.

Pettigrew, Thomas F. and Linda R. Tropp. 2006. "A meta-analytic test of intergroup contact theory." *Journal of personality and social psychology* 90(5):751.

Sartori, Giovanni. 1970. "Concept misformation in comparative politics." *American political science review* 64(4):1033–1053.

Tankard, Margaret E and Elizabeth Levy Paluck. 2016. "Norm perception as a vehicle for social change." *Social Issues and Policy Review* 10(1):181–211.

Trujillo, Matthew D and Elizabeth Levy Paluck. 2012. "The devil knows best: Experimental effects of a televised soap opera on Latino attitudes toward government and support for the 2010 US Census." *Analyses of Social Issues and Public Policy* 12(1):113–132.

Vivalt, Eva. Forthcoming. "How Much Can We Generalize from Impact Evaluations?" *Journal of the European Economics Association* .