No cover image available

Oxford Handbook of Engaged Methodological Pluralism in Political Science

(In Progress)

Janet M. Box-Steffensmeier (ed.) et al.

https://doi.org/10.1093/oxfordhb/9780192868282.001.0001

Published: 2023 **Online ISBN:** 9780191964220 **Print ISBN:** 9780192868282

CHAPTER

Tensions in Knowledge Accumulation Using Coordinated Intervention Experiments to Improve Public Policy

Jake Bowers, Natasha Greenberg, Morgan Holmes, Daniel N. Posner

https://doi.org/10.1093/oxfordhb/9780192868282.013.22

Published: 20 June 2024

Abstract

This chapter has two main goals. It introduces and explains the value of coordinated randomized experiments for informing public policymaking with a special focus on international development, and then articulates tensions inherent in these coordinated experiments. It explains why coordination can help build the evidence base for policy innovation by helping funders and policymakers test existing theories of change in ways that facilitate the cumulation of knowledge across studies. It then articulates tensions or trade-offs that reasonable decision-makers might face given choices between coordinated and uncoordinated experiments, and provides some rough guides to help decision-makers navigate those tensions. We write from the perspective of encouraging coordinated experiments and offer practical alternatives that might be easier to implement than extant models for policy-oriented decision-makers.

Keywords: knowledge accumulation, evidence-based public policy, meta-analysis, randomized experiments, public policy, evidence-based programming, foreign assistance

Subject: Political Methodology, Politics

Series: Oxford Handbooks

Collection: Oxford Handbooks Online

Overview

Opportunities for learning about the impact of public policies have improved significantly over the past several decades as program evaluations have shifted from post-hoc synopses of "how things went" to rigorous randomized controlled trials (RCTs) in which outcomes in units exposed to a policy or program are compared to a set of counterfactual units that are not. Although the results from such randomized evaluations offer a solid basis for learning about the impact of a given policy in a particular place at a certain

moment in time, they are vulnerable to questions about external validity: would the same conclusion have been reached if the policy had been implemented in a different time or place? This is a particularly important question for policymakers and funders, who may wish to apply lessons from evaluations undertaken in another setting to the context in which they are undertaking their own programming. For example, a decision-maker might ask: does a rigorous randomized study of a journalist training program in Bangladesh tell me anything about the likely impact of an analogous program here in Guatemala? Does an RCT of cash transfers to poor households in Seattle provide guidance for a program I am designing for rural Georgia? Do the findings of a randomized evaluation of a program conducted to encourage people to get their annual flu shots in 2015 (pre-COVID) provide any guidance for the likely impact of a similar program I am launching today?

One response to this challenge is to attempt to pool knowledge from evaluations of similar policies or programs in multiple settings in the hopes of identifying general lessons. This can be done loosely through informal literature reviews or more formally via statistical meta-analyses. However, as we elaborate below, the usefulness of such attempts at synthesis depends critically on the extent to which the policies studied, and the research designs employed are sufficiently similar to make cross-study comparisons meaningful.

To maximize such comparability, scholars and international donors in recent years have pioneered programs of coordinated research involving multiple, simultaneously executed, randomized experiments with harmonized interventions and research designs. By investigating the same policies in the same ways in multiple contexts, such initiatives simultaneously address the problems of external validity and comparability—and, by doing so, offer significant potential payoffs for broad learning and improvement of policy interventions across contexts. Indeed, our view is that such coordinated RCTs are the crucial next step in the evolution of evidence-based policymaking.

This chapter discusses such coordinated models, highlighting both the contributions they offer for evidence-informed public policymaking and some of the tensions that arise in designing and implementing them. For example, decision-makers will need to decide when the knowledge-base is sufficiently well-developed to justify launching a costly new coordination effort or whether enough has been learned to stop an existing effort. Decision-makers aiming to best serve a specific population in a specific context will need to decide whether to tightly coordinate interventions (which might make them less locally relevant, and hence less effective in any given context) or to coordinate less strongly (which may make it more challenging to compare findings across contexts and generate actionable general evidence). Decision-makers overseeing coordinated investigations of a particular intervention may be tempted to stop them as soon as the intervention shows disappointing results in one of the trials. Should they? When using the results of previous studies, decision-makers interested in a single place may be tempted to ignore results from contexts that differ from their own location, thus discouraging coordination. Again, should they?

We approach these tensions from the point of view of a funder, implementer of programs, or policymaker deciding how to manage a program of coordinated RCTs or weighing whether or not to participate in (or sponsor) such a program in the first place. Accordingly, our objective is not simply to articulate the trade-offs that such decision-makers will need to confront in embracing this approach but also to provide guidance about how to think about the choices they will need to make.

We believe that rigorous evidence from past research should provide a basis for the creation of new policies and programs, and that novel policy interventions should be evaluated using transparent, high integrity, and rigorous processes. That is, we write this chapter assuming that both evidence-as-insight and evidence-as-evaluation should be key parts of the policy innovation process,[1] and that both should be incorporated into policy and funding decisions.[2]

Preliminaries and Context

We begin by reviewing the special benefits that RCTs offer for providing an evidence-base for policy innovation. We also discuss some of the challenges that face funders, implementers and decision-makers who want to use evidence from RCTs to inform their decisions on programming.

Randomized Controlled Trials Improve Evidence for Policy Decisions

Government funded RCTs have a long history of informing public policy, starting with a few large studies focused on health and welfare, and diversifying since in size and topic.[3] Each RCT done in a transparent way offers easy-to-interpret evidence about the causal effect of an intervention: when the vaccination rates of people randomly assigned exposure to a new vaccine message are higher than the vaccination rates of people randomly assigned exposure to an old message, we know that the new message caused an increase in vaccinations. If we had not randomized—for example, if we had allowed people to *choose* which message to see (perhaps we, the health agency, posted the new messages at public libraries)—we would have had to spend time explaining how pre-existing differences between people might or might not explain the observed increase in vaccination rates: might people who visit public libraries in our state be more likely to keep up their vaccinations regardless of messages? This question is moot in an RCT: our control of messaging and our randomization of who receives it means that a person who loves the public library is equally likely to appear in either the treatment or the control groups: differences in vaccination rates after exposure to the randomized message cannot be explained by these or other differences in type of person. Thus, we can confidently conclude that exposure to the new message increased vaccination rates—at least among the people in our state at that moment in time.[4]

The Challenges of Interpreting the Results of Uncoordinated RCTs

The qualifying clause at the end of that last sentence is crucial: our hypothetical study of vaccine messaging provides strong support for the claim that the new message works in the setting in which it was tested. But what about in other settings? Can we rule out that the positive result arose only because of circumstances specific to the time and place where the study was conducted? This is a crucial question that donors and other policy innovators often face when they seek to learn from prior research to inform their own programming.

To learn whether our findings were driven by factors specific to our research site, we could run the study in multiple times and places and see if the findings hold up. Or we could review the published literature to see whether other researchers have undertaken studies about the impact of similar messaging in other settings and examine whether their findings match our own. If they do, then it would give us confidence that our findings about the efficacy of the new messaging may be broadly useful beyond our specific context.[5]

However, the published literature may not provide a useful guide. Academic careers depend on publishing articles that pass peer review, and academic journal editors prefer to publish articles that show strong and statistically significant results. Articles presenting weak or not statistically significant findings (indicating that what appeared to be a good idea did not end up working out) tend not to be published (Franco, Malhotra, and Simonovits 2014; Gerber and Malhotra 2008). This means that the academic literature provides a biased sample of what has been tested, and thus an unreliable yardstick against which to assess the findings of our own study. The published record might contain several other studies that, like our own, find that the new messaging increases vaccination uptake. But this may hide the fact that an even larger number of studies—unpublished, and thus unknown to us—estimate little or no impact.

This so-called file drawer problem (Rosenthal 1979) is compounded by the professional disincentives for replication, which reduces the number of studies that are undertaken to test a given policy or hypothesis. Big scholarly rewards accrue to the first researcher to test a promising hypothesis or to evaluate a commonly implemented policy. Fewer academic benefits come to the second (or third or fourth) researcher who investigates the impact of the same program in a different setting. Academic researchers might publish tests of the impact of *modifications* of the original policy or program, but not straight-up replications. The academic published literature is therefore thin not just on null results but also on studies that are highly comparable.

Yet even if we were to identify a complete (or at least representative and comparable) set of studies that assessed the impact of our vaccine messaging strategy in different settings, we may still face difficulties in comparing our findings against those of the other studies due to differences across the studies in their design and implementation. Variation in how vaccination rates were measured (for example, via self-reports vs. through administrative records), how the messaging was disseminated (via posters vs. a radio campaign; using public health officials vs. enlisting local celebrities; over a two week period vs. over several months), how the sample was selected (young people vs. nursing home residents; people in warmer vs. colder climates; in urban vs. rural settings), not to mention subtle differences in how the data was analyzed and in the message itself, may all generate differences in the reported findings.

This is not a problem of disincentives for researchers to study the same policy or program: even if all of the researchers were attempting to investigate the identical messaging program, subtle differences in how they designed and executed their investigations, along with minor tweaks in the messaging treatment to bolster its impact in the specific communities in which they were working, can all lead to different research findings even if the underlying impact of the policy would be the same in the absence of these differences. Researchers interested in discerning general impacts by collecting and comparing the results of these efforts would then have difficulty in figuring out whether these differences should be interpreted as inconsistency in the underlying impact of the messaging (i.e., the lack of a general effect) or simply variation in the research designs and measurement strategies employed by the different research teams. Researchers seeking to undertake such syntheses have developed sophisticated tools to deal with some of these challenges (Blair, Christensen, and Rudkin 2021; Lipsey and Wilson 2001; Wood 2008). But the inconsistencies that naturally arise when researchers independently evaluate the impact of a policy in different contexts—or even in the same context!—stands as a major impediment to the accumulation of general knowledge.

Coordinated Evaluations Promise to Improve Policy Decisions

The inconsistencies just discussed stem from the fact that researchers, working independently, make different choices about the details of their research designs and implementation strategies. The problem for knowledge accumulation stems not from the fact that these decisions are good or bad, right or wrong, but simply from the fact that they are different across the studies whose results we seek to synthesize. If the researchers could somehow have gotten together prior to launching their studies to coordinate the details of how they would structure their interventions, draw their samples, measure their effects, analyze their data, and implement other details of their research, then their findings would be much more easily synthesized, and the general knowledge gained about whether or not the policy works would be on much stronger footing: differences in results between studies would no longer arise from differences in the details, only from differences in the contexts.

This insight—that coordinating the details of independent evaluations of the same programs and policies can be a huge boon to learning—has inspired a new approach to organizing research aimed at informing policy. Coordination of this type is difficult for individual academic researchers because of the need to publish strong, statistically significant, and novel results quickly in peer reviewed outlets and to establish a reputation as an independent scholar having an impact on the field; but it is well-suited for governmental agencies, which frequently deploy the same policies in multiple settings, have an interest in learning about the impacts of their programming, and often employ researchers outside of academia, whose personal fortunes are less tied to considerations about publishability.[6] It can also be a workable strategy for NGOs that are committed to knowledge accumulation to determine the effectiveness of their programming, and that can play a coordinating and incentivizing role for independent researchers.

We discuss two such efforts below. The first is the pioneering "Metaketa" project spearheaded by the Evidence in Governance and Politics (EGAP) organization. The second is a pragmatic adaptation of this approach that we call a "rolling Metaketa." Whereas the EGAP initiative has successfully been implemented and provides a concrete example of how such an approach functions in practice, the rolling Metaketa initiative is prospective. We present it to lay out an approach that we believe holds great promise—especially for government organizations—and whose adoption we advocate.

The Metaketa Model for Coordinated Experiments

The Metaketa Initiative works by commissioning multiple field experiments that test a common policy, program, or hypothesis.[7] The field experiments include a common, coordinated treatment but are carried out in different settings. All of the research teams commit to testing the same intervention, employing the same outcome measurement, and doing the studies at more or less the same moment in time. The scholarly disincentives for coordination are addressed by providing funding for the studies, by encouraging each team to test additional treatments that go beyond the common treatment arm (thus providing opportunities for independent authorship of papers, outside the joint Metaketa project, testing novel hypotheses or mechanisms), and by offering co-authorship on a paper reporting the combined results—a paper that could not exist in the absence of the coordination. Once the individual studies are completed, their results are combined to produce an overall estimate of the effect of the common intervention.

The inaugural Metaketa, described in Dunning et al. (2019), investigated whether providing citizens with information about politicians affects their voting behavior. It coordinated seven separate randomized evaluations undertaken by teams of researchers in six different countries (Benin, Brazil, Burkina Faso, India, Mexico, and Uganda). Although each study provided citizens with slightly different kinds of information about politicians' performance, all of the projects measured exposure to the information in a consistent way and on a common scale. Thanks to their shared design and coordinated measurement strategies, the results of the seven studies could be analyzed together in a formal meta-analysis, thus making possible a more general conclusion about the efficacy of the intervention.

EGAP's experience with the first Metaketa was so positive that the organization subsequently launched four others, on the themes of taxation, the governance of natural resources, community policing, and how women can be mobilized to participate in consultative processes aimed at improving public services provision.[8]

A Practical Adaptation: The Rolling Metaketa Model

In a rolling Metaketa, research teams coordinate over time as well as across locations; rather than require that the harmonized studies be launched more or less simultaneously, as in the original EGAP model, the rolling model simply requires that the interventions and evaluations be undertaken so as to maximize their comparability. This approach is ideally suited to government organizations that implement similar programming on a handful of common topics year after year in different places.

For example, one office in the U.S. Federal government, the Office of Evaluation Sciences (OES) is already doing multiple studies on the same topic, although it is not coordinating those studies as self-consciously as one would in a rolling Metaketa. OES initiates studies of programming undertaken by various government agencies following a standardized project process to facilitate research integrity and, it turns out, enables comparability.[9] For example, between 2015 and 2019, the OES undertook randomized evaluations of eight different behaviorally informed direct communications to promote vaccination uptake (Kappes et al. 2023). The fact that multiple studies of the same topic were done by the same team, following the same project process allowed Kappes et al. (2023) to accumulate evidence from those different studies even though each study was fielded by different individuals rotating onto and off of the team, supporting different agencies, during different years, in different locations.

Such efforts can (and, we believe, should) be greatly expanded and formalized. We point to only two of many other opportunities for how this might be done within the U.S. federal government. United States Agency for International Development (USAID) initiates similar programming in missions around the world in almost every major sector. Within the US, the Office of Planning, Research and Evaluation of the U.S. Department of Health and Human Services also studies the same topics across multiple contexts. Such programming is designed and motivated by the policy concerns of each agency (and, in the case of USAID, each mission-level decision-maker) at a particular moment in time. But by recognizing that these concerns are shared broadly, and by agreeing to coordinate efforts to learn about the impact of what often turn out to be very similar interventions, knowledge can build faster and in a more directed fashion than via approaches that collect disparate, uncoordinated studies for summarization and meta-analysis.

To get a sense of the opportunities for leveraging such an approach, consider the fact that more than twenty interventions aimed at promoting political participation among women and youth have been launched by USAID missions in recent years, and many of these have been accompanied by rigorous evaluations of their impact. By simply coordinating the design and implementation of these evaluations across missions, USAID could turn what they are already doing into a rolling Metaketa. Unlike an EGAP Metaketa, in which the various projects are initiated at the same time, USAID would roll out the coordinated evaluations over time as different missions, operating on their own timetables, initiated their programming in the common area. The key to making it work is to ensure that each mission adheres to a common design and data collection protocol and that careful records are maintained so that the findings of each evaluation can feasibly be synthesized.[10] This model would likewise be well suited to U.S. Department of State programming or that of other agencies, as well as the work of other governmental and large private donors with overarching topic-oriented learning agendas and many interventions on similar topics.

The benefit to these donors and decision-makers would be the generation of evidence about what works (and does not) that is not just rigorous but general—or, if not general, then offering particularly useful insight into the conditions under which particular interventions are more and less likely to have an impact. The cost would be simply imposing a standardized template on the design and implementation of programming that is already happening, along with an improvement in record-keeping by a team overseeing the coordination and some coordination across missions, embassies, or other places where interventions are being designed and implemented.

The rolling Metaketa model requires that research teams at each place understand themselves to be a part of a coordinated process of learning, and that they commit to building interventions, measuring outcomes, and designing studies that relate directly to both past studies and future studies. Although challenges of coordination are central in any Metaketa (Dunning et al. 2019), they are especially acute when the projects are launched in a rolling fashion. How can a researcher in one time and place see themself as a part of a coordinated effort contributing to broader learning—especially when there is no direct contact with the other researchers whose rigorous evaluations are also contributing to the coordinated project (some of whom have not even begun their planning processes)? The OES example shows us that it helps when the production of research is organized by an institution that exists to serve decision-makers and when the organization has a transparent project process. The U.S. government's Evidence-Based Policy Making Act of 2018 and the efforts to implement that law, such as production of multi-year learning agendas for agencies and annual evaluation plans, can be a driving force for achieving this objective.[11]

To the extent that research comes to be conducted by teams of professionals within government, it is likely to be governed by project processes that will enable studies to build on one another in ways that might be easier than if they had been done by people based only in universities responding to the incentives of academic careers. Among U.S. federal agencies, USAID, the State Department, and the Department of Health and Human Services (HHS) are already structured to foster this, with technical bureaus that centralize knowledge and approaches and give technical assistance on particular topics, aiming to generate and disseminate best practices. These bureaus could take the lead in setting the parameters for the common project designs and implementation protocols that are crucial to the comparability of the evaluations that emerge over time. Technical bureaus could also be responsible for maintaining records (data, documentation, and so on) of the independent evaluations that make up the rolling Metaketa, and for undertaking the meta-analyses once a sufficient number of studies have been concluded. Such coordination and oversight by a separate institution is particularly important for government agencies like the U.S. Department of State and USAID, whose staffing structures have some degree of built-in staff rotation, which can create challenges for initiatives (like a rolling Metaketa) that unfold over time.

Tensions Inherent in Designing and Implementing Coordinated Studies

Coordinated intervention experiments such as those pioneered by EGAP, and that could feasibly be launched in rolling form by many government agencies, offer the promise of significant benefits to learning. However, funders and policymakers interested in adopting this approach will confront tensions and trade-offs that naturally arise in designing and implementing such coordinated studies. We describe some of these tensions below. We also engage with some of the practical problems involving such coordination in the next section.

Tension between Starting a New Coordinated Effort vs. Continuing to Invest in an Existing Effort

The decision to initiate a program of coordinated research represents a large commitment of money and staff effort. Does the past research teach us enough about an existing theory of change such that we should launch a new coordinated effort focusing on testing a new one? Or should we continue to build knowledge about the existing theory of change (via pilot and/or laboratory studies) before launching an expensive research effort involving coordinated field trials? Assuming we decide to initiate such an effort (for example, via a rolling Metaketa), should there be a stopping rule? How many null results or even positive results do we need before we stop fielding additional studies and return our focus to developing a new theory of change?

Here is an example. Say a prominent theory of change suggests that empowered journalists are the key to preventing democratic backsliding. Imagine that a coordinated effort to build evidence for or against this theory has begun, with the first five studies collectively reporting a small positive effect. Should the team working to improve democracy in a single country join this study to add precision to the existing overall estimate (just in case they discover a negative or null effect)? Or should the team say: "We know enough about journalists and democracy to adapt the approach that has worked in five places over the past decade of coordinated studies. We should invest research resources in learning about something new: maybe about the effects of artificial intelligence [AI] generated mis-information on civil society organizations, now that the question about journalists is settled enough for us to scale up in our context."?

Tactics to help decision-makers evaluate this trade-off.

We suggest a couple of strategies for decision-makers facing such choices. If a community of practice could be convened for periodic conversations, this group could evaluate how much is known about a given theory of change and associated policy interventions and their effects, as well as about the costs of getting it wrong by stopping too early. This assessment could occur every few years or, in the case of a rolling Metakata or a series of studies all occurring within a given organization like USAID or OES, every fifth study. Overarching questions of this group—which probably should include representatives from multiple contexts as well as academics—would be whether evidence has provided robust enough evidence on the theory of change under study to make formal recommendations on its use; what the overall learning agenda for a given topic includes and a determination of whether to prioritize certain questions as those most important to study; when an intervention has been sufficiently developed to warrant a coordinated meta-analysis, and what types of interventions are and are not appropriate to be tested through RCTs and potential meta-analyses. The idea would be to create a consensus on learning to guide coordinated experimentation for the next period of time.[12]

Tension between Sticking with a Treatment Used Before (to Maximize Comparability) vs. Changing the Treatment to Allow Learning or to Be More Relevant to a Context

The previous tension revolved around questions of when to initiate a program of coordinated research and when to stop it once it has been initiated. But stopping is not the only option: a decision-maker/coordination body could decide to continue the research process but change key elements of the intervention. The trade-off then becomes one between maintaining fidelity to the prior protocols and measurement strategies to maximize comparability across the whole set of studies and tweaking things to take advantage of learning that has occurred since the launch of the coordinated research effort, or to better suit the local context.

For example, in the inaugural Metaketa project, the objective was to learn how citizens responded to information about politicians' performance (Dunning et al. 2019). While it was reasonably straightforward to agree on whether information provided in different contexts constituted "good news" or "bad news" about performance, the specific kind of information that was provided (for example, about misallocated spending in Mexico vs. the alignment between voters' and candidates' policy preferences in Uganda vs. the quality of municipal services provided by the previous incumbent party in Burkina Faso) was necessarily different and context-dependent. Rigid insistence that all interventions be identical (for example, all involving information about budget irregularities) would have led to interventions that were not locally appropriate, or even irrelevant, to the concerns of citizens in a given setting. The Metaketa organizers had to weigh the trade-off between tweaking the interventions to suit each context (thus maximizing local relevance and suitability) and winding up with weak treatments in some cases (thus undermining the contribution of those studies to the coordinated research effort).

While such challenges of harmonization will inevitably arise in any coordinated research effort, it is especially challenging in the rolling Metaketa model, where the details of the core intervention is likely to have been designed to suit the facts on the ground in an initial set of contexts but may be less well suited to the facts on the ground in settings that join the project in its later stages (that, in some cases, the original intervention designers may not have even anticipated would join the coordinated effort).

Changing the intervention—whether because of learning that has occurred due to the results of the prior studies or because the intervention doesn't seem to suit the context—makes it more difficult to combine the results of this one study with the past studies, and so diminishes the power of coordination to speed learning about the theory of change that provides general guidance for the interventions.

Tactics to help with the trade-off.

We think that the decision-maker (and coordination body) should make changes in the interests of the welfare of the people in the site but should also have some way to summarize how much is known about the early results. A null result can mean many things. For example, if the previous four studies had small samples, then a large fifth study could very well change our overall estimate of the effect of the intervention.[13] This means that the coordinating body mentioned above could work on communicating not just whether each of the previous results was statistically significant but also the precision of the combined estimate so far (after the first four studies). Further, we think that decision-makers in specific sites should be encouraged to be creative in their intervention creation but be encouraged to *add an arm* to the experiment to compare the harmonized arm to the intervention that differs from the previous arm. The experimental design devices that we describe below (factorial designs, unequal and changing probabilities of treatment, and placebos) all provide ways for a decision-maker focusing on the best intervention for a given context to both serve that context and also to contribute to the broader learning agenda of the field and/or agency.

Tension between Learning from Other Contexts vs. a Focus on One Place and Time

A naive application of the rolling Metaketa model would attempt to take advantage of any and all opportunities to add new studies to the broader research program. However, a savvy funder or policymaker knows that causal effects operate in a context, and that the payoffs to running an investigation in a given setting may vary from place to place or from one moment in time to another. Cartwright and Hardie (2012) demonstrate this compellingly, showing how positive effects of small classrooms on test scores in Tennessee did not translate well into the context of California.[14] Two of the authors of this chapter have been involved in replications which similarly emphasize the importance of context to the causal effect. When Raffler, Posner, and Parkerson (2023) replicated the Björkman and Svensson (2009) study of community-based monitoring of health care workers in Uganda, they did not find the same positive health outcomes that inspired their replication: baseline health care had improved between the two studies, leading to a kind of ceiling effect for much of the second study sample. A study on the impact of short message service (SMS) reminders on vaccination uptake undertaken by Dai et al. (2021) found that receiving a text message about the importance of getting vaccinated was associated with a six percentage point increase in COVID vaccination among older UCLA Health System members. This was a welcome result given the substantively large size of the estimated effect and how inexpensive such reminders are to implement. However, these findings did not translate well to younger people in Rhode Island who had not yet received a COVID vaccine after a month of COVID vaccine availability (Rabb et al. 2022). It would appear that reminding people to take advantage of a policy in which they already want to participate (the participants in the Dai et al. (2021) study had been anxiously awaiting the release of the first COVID vaccines) was quite different from motivating those already ambivalent about COVID vaccines who had already had ample opportunity to become vaccinated, had they chosen to do so.

These examples suggest that it might not be in the best interest of the public in a given context for a decision-maker to participate in a rolling Metaketa. A previously positive effect might easily vanish in a new context, and a previously null effect makes it hard to justify repeating what appears to be a previous failure.

Tactics to help with the trade-off.

One way around these problems, especially in the rolling Metaketa model, would be to create a complex theory of change that encompasses multiple different preconditions and factors and provides guidelines about the conditions under which the intervention is most likely to be successful. This could lead to a model that has a menu of interventions with a decision tree to help to identify which circumstances most closely match with the newest proposed application site; a decision tree which would point to the recommended intervention.

In the absence of such a decision tree, a solution on a study-by-study basis may be to introduce theory-based arms to test whether our prediction that intervention X may not work as well in context Y is true. The design team could try to explicitly link context to the theory of change by formally representing the kinds of contextual variations we might expect. For example, given prior beliefs that SMS messages might fall flat for unmotivated people but catalyze action among those previously motivated, one could imagine simulating the effects of a study where a hypothetical population is a mix of those motivated and unmotivated to see how many unmotivated people in the population would sap the intervention of its hoped for overall effect. And then one could try to learn about the motivated to unmotivated ratio in the given population using some quick survey or other measurement. Or one could dramatize the effect that a single null result might have on the distribution of possible outcomes that the community of practice might have imagined would be plausible given past research. The same committee charged with identifying key questions to answer and assessing when a rolling Metaketa has adequately answered a question could also assist with the design of these arms to test how the theory works.

Tension between Learning More about a Given Site vs. Learning More about How Sites Differ

Some decision-makers overseeing or funding a coordinated research effort might feel a tension between going broader or deeper (i.e., maximizing heterogeneity across units to increase learning about generalizability vs doubling down on statistical power in fewer places or selecting multiple sites within the same "unit" to learn more about the generalizability of the findings within that setting). If a decision-maker can select sites purposefully, should she deploy her resources to include more sites, with more heterogeneity across them, or more units within sites?

Tactics to help with the trade-off.

Our current vision of a rolling Metaketa design, building on the original simultaneous Metaketa design, involves going deep instead of broad. The idea is that each study should provide enough information to detect realistic, policy-relevant, effects in a given context and also that each study be able to directly contribute to decision-making in that context upon completion: we envision coordination occurring over time and over such individually useful studies. The primary reason to do more than one study is to address the problems of replication, publication bias, and measurement explained by Dunning et al. (2019).[15]

Practical Challenges to Implementing Coordinated Research Strategies

Tension between Long-Term Investment in Research and Immediate Investment in Programming

Policymakers, funders, and federal accountability bodies often focus on somewhat short time horizons and a desire to see results quickly. There can be tensions in deciding whether to take the time and spend the additional funds to accompany programming with research. While it is well established that research is essential in identifying safe and efficacious medical interventions, this has not been the norm or practice in some of the social sciences and international development fields. In the absence of clear research standards and protocols, one could argue that these fields have not displayed strong evidence that their interventions are consistently working, and that existing funding is being used in the most effective manner. Unlike smallpox, in which a solution was researched, tested, and then implemented, leading to the eradication of the disease, the issues these programs attempt to address still plague society. While they exist in very complex sociopolitical systems, making them harder to isolate as well as address, it would behoove the community that funds and implements this work to agree on a more consistent standard and methodology for testing the effectiveness of interventions.

Tactics to help with the trade-off.

It can be a challenge to get funding for research to accompany programming for even a single intervention; getting a commitment for research for a series of interventions is an even higher bar. One important step in addressing this is confronting the misperception that if a program includes research, all of the funds support pure research, rather than directly and immediately benefitting beneficiaries. In our vision, the research is not being conducted in a lab and then later applied to real situations. Rather, we see practical, action-oriented research that is being pursued hand-in-hand with an intervention, in an ethical manner that both directly serves beneficiaries and helps to establish whether or not the intervention is having the intended effect. The cost for the research components that accompany the intervention are often a drop in the bucket in the grand scheme of the amount of funds going to programming. If the research shows that an intervention is not effective, it could save funders and taxpayers huge amounts of money, prevent beneficiaries from being subjected to ineffective programming, and spur the generation of creative new ideas.

Tension (or Challenge) of Establishing High Fidelity When Replicating Models

In the existing Metaketas, although researchers were organized around the same model in the same timeframe, the programs often were not identically applied because the different program implementers were working in different contexts with unique challenges and circumstances. This lack of exactly identical implementation was expected to some extent: after all, even if every site printed flyers with information about the incumbent mayor, those flyers were translated into multiple languages and were disseminated in different geographic, urban, and rural contexts. In other projects, such as the Metaketa project on policing, the academic teams had little control over the decisions of the different police forces. This heterogeneity and lack of control can make it difficult to identify one central model to replicate for the future. In addition, if some interventions were more effective than others, it complicates the analysis of what led to these differences: was it the intervention or something about the context?

Tactics to help with the trade-off.

In field experiments, the experimental intervention rarely is identically applied to the experimental subjects: for example, different field staff knock on different doors and have slightly different conversations when an intervention is provided door-to-door and different sized groups of people with different group dynamics gather to receive group-based interventions. Randomization averages out these many differences between arms within a single study. Yet, the comparisons across places (say, after every five studies in the rolling Metaketa model), will not have randomization across sites, and the details of implementation of even a well-harmonized common intervention will differ systematically across the sites. So this tension is practical and highlights the need for a coordinating body to help sites minimize differences in intervention and/or build in auxiliary data collection to help answer questions about the cross-site variation: for example, sites could be encouraged to add experimental arms that address questions about why treatment effects differ across sites alongside the harmonized arm. The coordinating body should also help individual sites attend to the theory of change and learning agenda that animates the work: slight differences in intervention may still advance the overall goal of learning whether, say, democratic backsliding is best prevented by a focus on journalists or youth movements, even if the fine details of the interventions across sites differ.

Tension (or Challenge) between the Incentives of Academics and the Incentives of Policy Decision-Makers

Academic careers depend on the reputation of a given scholar within a larger academic community. Thus, a young scholar must focus research on theories of interest to this community and must publish their work in peer reviewed outlets. Academic publications can take years to appear, and academics often do not want to share their results or data in the meantime as such sharing could impede their ability to publish and advance in their career. This is in direct tension to the needs of funders who want to make the results immediately public, explain how the funds were used, and what implications it has for future funding. This tension is not unique to the problem of coordinated experiments, but it does make it harder to execute such coordination.

Tactics to help with the trade-off.

Some possible solutions to this tension are to hire tenured academics who conduct these studies as one aspect of their work but may have a slightly lower amount of pressure to publish, to hire staff in centers affiliated with universities whose careers do not depend on publishing (where funders could have an important role in supporting these centers and thus signaling the value of this work to universities and the academic community at large).[16] Some universities in the United States were founded with applied public service as a core component of their missions. So, we imagine that incentives could align for many academics in many places if their institutions have a public service mission.

Another tool is ensuring the government agencies or funders of the research have data access plans and policies that allow for publication before sharing data publicly, as USAID has.[17] Research dissemination plans, laid out at the outset of a study, are also a good opportunity to ensure the funders, implementers, and academics have a shared understanding of when, how, and with whom data and findings will be shared. Oftentimes the internal sharing of findings and data is sufficient to meet the needs of funders for urgent answers and those funders are willing to enact public data sharing policies that allow for academic publication.

Devices to Make Design Reflect Policy Goals

In this section we propose some other reflections that might help organizations begin to implement either simultaneous or rolling Metaketa style coordination across the evaluations they are already undertaking.

Experimental Design Can Be Flexible and Reflect Policy Goals

The trade-offs between designing single studies to evaluate interventions in specific places and times vs. coordination are not as stark as they may seem. For example, researchers routinely field experiments testing more than one intervention in such a way that estimates of the impact of both interventions have roughly the same precision as they would if only one intervention were studied. The technical devices used to do this include the factorial design (most well known as the 2 x 2 design), adaptive experimental designs (Kasy and Sautmann 2021), the use of placebos, and multi-arm studies in general.

The factorial design simply involves randomly assigning two treatments independently of each other. For example, the following Table 1 describes an idealized 2 x 2 factorial design in which 200 people are assigned to two treatments, one about vaccines and another about tax payments. In this case, we can treat the design as two different experiments, each with 200 people from the perspective of assessing the effect of tax payment info and the effect of vaccine information separately.[18]

Table 1 A Factorial Design with Two Independent Treatments

	Vaccine Info.	**Status Quo**	
Tax Payment Info.	50	50	100
Status Quo	50	50	100
	100	100	

The factorial tactic thus enables decision-makers to field a harmonized arm (say, the arm focusing on the impact of receiving information about vaccines) alongside an arm designed specifically for the context or an arm that is hoped to improve upon the harmonized arm (in this example, the arm involving the information about tax payments). The difficulty here is to ensure that the two arms do not conflict: if learning about tax payments dramatically changes how people react to vaccination information then; we might want to avoid the factorial design. However, if tax payment information and vaccination information have little to do with each other in a given context, then the factorial design can offer great benefits in regards combining harmonization alongside flexibility.

The basic placebo tactic differs from the simple factorial design and involves two active interventions and one control arm: say, one arm involves door-to-door visits to encourage vaccination, another involves door-to-door visits to encourage payment of taxes, and a third involves no visits at all. If vaccination outcomes for the tax arm are the same as we would see in the control group arm, and if tax payment outcomes in the vaccination arm are also the same as we would see in the control group, then we would have two studies in one: a study of a door-to-door tax payment intervention and a door-to-door vaccination intervention. As a side benefit of this design, we can also learn about the effect of the active "dose" of the intervention—what happens when someone opens the door and interacts with the enumerator/field staff.[19]

Finally, a simple multi-arm study can be enough to enable a consistent intervention to be implemented over many sites or country-contexts. In existing Metaketas, multi-armed studies have been used to maintain a consistent coordinated intervention across countries while using additional arms to test additional hypotheses or customizations of the main, common treatment that funders or governments wish to evaluate in that particular setting. In the end, there is a consistent arm that is comparable across countries and each individual country's stakeholders can have their questions addressed as well. We also note that a multi-arm study does not require that all arms contain the same numbers of subjects in all sites. The literature on adaptive experiments reminds us that valid experiments can have probabilities of assignment that are unequal and are changing according to some rule. So, if a multi-arm rolling Metaketa is launched, and one of the arms appears to perform worse than other arms, future sites can maintain coordination by including fewer cases in that arm.[20]

Communication and Record Keeping Are Key

We suggest that a coordination body be created to organize the efforts to field coordinated studies across time and locations. This body should record the results of previous studies and serve as the repository for the materials for the next study: for example, providing templates for pre-analysis plans, sampling protocols, and survey questionnaires, as well as advice about common problems encountered in the previous studies. The body can provide guidance about creative and flexible approaches to combining adjusted and contextually designed interventions in the same study alongside the harmonized interventions. This body can also foster communication across study teams working in different locations by convening periodic meetings to discuss progress in regard to an overarching learning agenda and make the kinds of decisions about stopping and starting coordination that we identified as an important tension in this process.

Next Steps in Policy-Relevant Coordinated Experimentation?

Loose coordination of studies over time is already occurring as policymakers seek guidance from research. Some of this coordination occurs because multiple research projects are being done by the same organization (examples include the more than a hundred randomized field experiments fielded by the OES across the federal government, the many studies focusing on youth empowerment being developed by USAID, the research done on human trafficking under the State Department's Program to End Modern Slavery, and eventually, other work by U.S. federal agencies organized under the Evidence Act). Some of this coordination occurs as scholar-practitioner collaborations replicate the most promising previous work as a conservative approach toward implementing already-tested interventions in new contexts (e.g., Raffler, Posner, and Parkerson (2023); Rabb et al. (2022)). Because of the transparency standards maintained by many of these recent studies, scholars have been able to execute large-scale meta-analyses of the studies done across organizations to synthesize knowledge from the many different studies (for example, DellaVigna, and Linos (2022) look across 126 RCTs done by two such organizations, the OES and the Behavioral Insights Team). Tightly organized simultaneous research projects have been completed or are in progress under the auspices of EGAP: these projects have involved more than twenty research teams across roughly twenty countries focusing on five topic areas and theories of change.

The first Metaketa effort was created to overcome problems in knowledge accumulation common in academia—lack of replication, differences in measurement and data collection, and publication bias. When translated to the world of public policy we see new challenges arising including the problem that not all decision-makers are ready to field the same intervention at the same time. The pragmatic adaptation of the Metaketa model, that we call the "rolling" model, allows for each decision-maker to participate in a coordinated effort at their own pace. Yet, just as the incentives of academia lead to publication bias and lack of replicability, the incentives of funders and policymakers can lead away from research and coordination. We hope that this paper spurs coordination by helping teams confront those incentives and pressures productively, both by naming those tensions and also by suggesting starting places for resolving them. For example, we recommend coordination roles be created and filled in agencies and/or by bodies affiliated and funded by agencies, and that those coordinators convene meetings to decide whether we have learned enough about a given theory or need to learn more; or that those coordinators make life easier for individual sites by providing example analysis plans, simulation code, and other materials. We imagine the coordination bodies might have different institutional homes depending on the specifics of the theory of change driving the coordination: we could imagine them existing mostly within government agencies, or at universities, national academies, or non-profit organizations.

Although we have focused on tensions and challenges, we want to end with a side benefit of coordination. Experience with the original Metaketa model has revealed that, when results differ between places, teams found that they had to collect more information to describe the differences across the sites than they might have done had they focused only on a single site, assuming that the tough work of synthesis and cumulation would be undertaken by other teams at other times. Recent work suggests that a study is more useful to future decision-makers if it collects more background descriptive information about the units and context in which it was fielded (Chassang and Kapon 2022). Coordinated research projects like the ones we describe and advocate are more likely to collect this kind of background information than independent evaluations of single interventions in part because such coordination encourages individual research teams to see themselves as a part of a broader and cooperative effort to build generally useful evidence about theories of collective interest. We propose that this unintended outcome—the establishment of a community of practice dedicated to theory and learning—can expedite our collective knowledge acquisition about the world. Moreover, we hope such a community of coordinated researchers and practitioners can influence government institutions and catalyze concrete interventions that improve the lives of people around the world.

Acknowledgments

Greenberg is an employee of the U.S. Department of State and Holmes is an employee of USAID; the views expressed herein belong solely to the authors and do not necessarily reflect those of the U.S. Government.

References

Balu, Rekha. 2020. "10 Things Your Null Result Might Mean." *EGAP Methods Guides*. https://egap.org/resource/10-things-your-null-result-might-mean/ (November 13, 2023).

Björkman, Martina, and Jakob Svensson. 2009. "Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda." *The Quarterly Journal of Economics* 124 (2), 735–69. https://academic.oup.com/qje/article-abstract/124/2/735/1905094.
Google Scholar WorldCat

Blair, Graeme, Darin Chrisetnsen, and Aaron Rudkin. 2021. "Do Commodity Price Shocks Cause Armed Conflict? Evidence from a Meta-Analysis." *American Political Science Review* 115 (2), 709–16.
Google Scholar WorldCat

Bowers, Jake, and Paul F. Testa. 2019. "Better Government, Better Science: The Promise of and Challenges Facing the Evidence-Informed Policy Movement." *Annual Review of Political Science* 22 (1), 521–42. https://doi.org/10.1146/annurev-polisci-050517-124041.
Google Scholar WorldCat

Broockman, David E., Joshua L. Kalla, and Jasjeet S. Sekhon. 2017. "The Design of Field Experiments With Survey Outcomes: A Framework for Selecting More Efficient, Robust, and Ethical Designs." *Political Analysis* 25 (4), 435–64. https://www.cambridge.org/core/journals/political-analysis/article/design-of-field-experiments-with-survey-outcomes-a-framework-for-selecting-more-efficient-robust-and-ethical-designs/BBD56764268C914806D23AB5D7403636 (November 3, 2023).
Google Scholar WorldCat

Cartwright, Nancy and Jeremy Hardie. 2012. *Evidence-Based Policy: A Practical Guide to Doing it Better*. Oxford University Press.
Google Scholar Google Preview WorldCat COPAC

Chassang, Sylvian, and Samuel Kapon. 2022. "Designing Randomized Controlled Trials with External Validity in Mind." Unpublished paper.

Dai, Hengchen, Silvia Saccardo, Maria A. Han, Lily Roh, Naveen Raja, Sitaram Vangala, Hardikkumar Modi, et al. 2021. "Behavioural Nudges Increase COVID-19 Vaccinations." *Nature* 597 (7876), 404–9. http://dx.doi.org/10.1038/s41586-021-03843-2.
Google Scholar WorldCat

DellaVigna, Stefano, and Elizabeth Linos. 2022. *RCTs to Scale: Comprehensive Evidence from Two Nudge Units*. *Econometrica* 90 (1), 81–116.
Google Scholar WorldCat

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, Craig McIntosh, and Gareth Nellis. 2019. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*. Cambridge University Press.
Google Scholar Google Preview WorldCat COPAC

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social *Science*s: Unlocking the File Drawer." *Science* 345 (6203), 1502–05.

Gerber, Alan S., and Neil Malhotra. 2008. "Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals." *Quarterly Journal of Political Science* 3 (3), 313–26.
Google Scholar WorldCat

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: WW Norton.
Google Scholar Google Preview WorldCat COPAC

Green, Donald. 2018. “10 Things to Know About Conducting a Meta-Analysis.” *EGAP Methods Guides*. https://egap.org/resource/10-things-to-know-about-conducting-a-meta-analysis/.

Gueron, Judith M., and Howard Rolston. 2013. *Fighting for Reliable Evidence*. New York: Russell Sage Foundation.
Google Scholar Google Preview WorldCat COPAC

Kappes, Heather Barry, Mattie Toma, Rekha Balu, Russ Burnett, Nuole Chen, Rebecca Johnson, Jessica Leight, et al. 2023. “Using Communication to Boost Vaccination: Lessons for COVID-19 from Evaluations of Eight Large-Scale Programs to Promote Routine Vaccinations.” *Behavioral Science & Policy* 9 (1), 11–24. https://doi.org/10.1177/23794607231192690.
Google Scholar WorldCat

Kasy, Maximilian, and Anja Sautmann. 2021. “Adaptive Treatment Assignment in Experiments for Policy Choice.” *Econometrica* 89 (1), 113–32. https://www.econometricsociety.org/doi/10.3982/ECTA17527.
Google Scholar WorldCat

Lipsey, Mark W., and David B. Wilson. 2001. “Practical Meta-Analysis.” *Applied Social Research Methods Series* 49, 247.
Google Scholar WorldCat

Nickerson, David W. 2005. “Scalable Protocols Offer Efficient Design for Field Experiments.” *Political Analysis* 13 (3), 233–52. https://www.cambridge.org/core/journals/political-analysis/article/scalable-protocols-offer-efficient-design-for-field-experiments/A9274FF8730F293654B35CEFDD8959A6. (November 3, 2023).
Google Scholar WorldCat

Offer-Westort, Molly, Alexander Coppock, and Donald P. Green. 2021. “Adaptive Experimental Design: Prospects and Applications in Political Science.” *American Journal of Political Science* 65 (4), 826–44.
Google Scholar WorldCat

Office of Evaluation Sciences. 2019. *Unexpected and Null Results Can Help Build Federal Evaluation Plans and Learning Agendas*. Office of Evaluation Sciences. https://oes.gsa.gov/assets/files/unexpected-results-2-pager.pdf.
Google Scholar Google Preview WorldCat COPAC

Posner, Daniel N. 2019. “Proposal for a ‘Rolling Metaketa’ at USAID.” Unpublished memo.

Rabb, Nathaniel, Megan Swindal, David Glick, Jake Bowers, Anna Tomasulo, Zayid Oyelami, Kevin H. Wilson, and David Yokum. 2022. “Evidence from a Statewide Vaccination RCT Shows the Limits of Nudges.” *Nature* 604 (7904), E1–7. http://dx.doi.org/10.1038/s41586-022-04526-2.
Google Scholar WorldCat

Raffler, Pia, Daniel N. Posner, and Doug Parkerson. 2023. “Can Citizen Pressure Be Induced to Improve Public Service Provision?” Unpublished paper.

Rosenthal, Robert. 1979. “The File Drawer Problem and Tolerance for Null Results.” *Psychological Bulletin* 86 (3), 638–41. https://psycnet.apa.org/fulltext/1979-27602-001.pdf.
Google Scholar WorldCat

Wood, John “Andy.” 2008. “Methodology for Dealing with Duplicate Study Effects in a Meta-Analysis.” *Organizational Research Methods* 11 (1), 79–95.
Google Scholar WorldCat

Notes

1 In this chapter we use "policy innovation" as a short hand to refer to the process of coming up with a new idea for how to encourage change (for example, increase support for democracy or increase vaccination rates) or provide benefits (for example, ensure that those eligible for educational benefits get them) or other governance-related action (for example, prevent fraud in unemployment benefits requests). Policy innovation tends to occur via complex collaborations within and outside of governments involving philanthropists, non-governmental organizations (NGO)s, civil servants, academics, elected officials, and so on. We also say "policymaker" or "decision-maker" to include both funders of programs but also the key decision-makers supervising the implementation of a given program.

2 We understand evidence-based policymaking to involve evidence in at least two forms: (1) The results of past finding, the basis of evidence for considering a new policy innovation and (2) evidence about whether or not a given policy intervention was effective. We follow Bowers and Testa (2019) in calling these two forms of evidence "evidence-as-insight" and "evidence-as-evaluation".

3 Gueron and Rolston (2013) present a history of early RCTs in U.S. public policy. DellaVigna and Linos (2022) present a meta-analysis of 165 experiments completed between 2015 and 2019 by the Office of Evaluation Sciences supporting the U.S. federal government and the Behavioral Insights Team North America supporting U.S. local governments.

4 If we did not find differences across people exposed to the old and new messages, the signal would be less clear. But we would still be confident that the lack of an effect was not caused by differences in the characteristics of the people exposed to each message. For more on the use of unexpected and null results see (Balu 2020; Office of Evaluation Sciences 2019).

5 See, for example, the What Works Clearinghouse which collects single studies of educational interventions done in different schools at different moments in time and also supports aggregating the results of those studies to produce single overall estimates via statistical meta-analysis (Green 2018).

6 For example, the Office of Evaluation Sciences (OES) in the U.S. Federal Government has fielded more than a hundred randomized field experiments, each of which has been designed to address the policy goals of a particular agency. Academics join the OES as fellows for one or two years, during which time they focus primarily on working groups to publish reports to improve policymaking. No OES report carries the names of any authors. And, although the OES project process is rigorous and involves review by academic and non-academic researchers across disciplines, every project is published regardless of statistical significance or novelty. Many federal agencies are also now building their own bases of evidence as a consequence of the broader appreciation of data and evidence and the Evidence-Based Policy Making Act of 2018 (see http://evaluation.gov for learning agendas and evaluation plans by many different federal agencies).

7 Further details are provided at https://egap.org/our-work/the-metaketa-initiative/.

8 The struggles of successfully completing so many studies in coordination have also generated useful reflections about this model of coordination and how to improve it (see for example this summary https://egap.org/resource/beyond-the-metaketa-initiative-reflections-from-meetings-with-egap-members/).

9 See https://oes.gsa.gov/projectprocess/ for details.

10 See Posner (2019) for a fuller discussion of the rolling Metaketa model as it might be applied to the democracy and governance sector at USAID.

11 For example, https://www.evaluation.gov/.

12 It would be tempting to formalize this process by eliciting prior beliefs from all of the participants and then calculating posterior distributions of effects implied by the priors and the results collected so far. Efforts like this could certainly help the conversations that we envision. That said, we don't see a metric for "precision of the posterior" or "difference between posterior and prior" that ought to drive action on its own in the absence of the broader discussion that we imagine.

13 We are assuming four null results or small effect sizes, not four harmful results. We hope that large harmful effects would not be replicated four times (and in fact, hope that large, unexpected harms are prevented during the monitoring of the field work, and that such harms could trigger re-evaluation of the theory of change and intervention if they are detected). For more on interpreting null but not harmful results, see Balu (2020).

14 Cartwright and Hardie (2012) describe the results of the Situation, Task, Approach, and Results (STAR) RCT in Tennessee which showed that smaller classroom sizes increased academic performance over larger classroom sizes for children in grades K–3. They further describe how evaluations of the causal effect of smaller classrooms in California did not show such an effect. Why did small classrooms work well in Tennessee but not as well in California? Cartwright and Hardie speculate that the effect of classroom size depended on a supply of experienced teachers and quality classrooms: when California mandated small classrooms, they ended up with a shortage of classrooms and also of experienced teachers, and so some of the small-group instruction in California was occurring in hallways by first year teachers, and thus was of lower quality than the larger classroom instruction occurring in quality classrooms by experienced teachers. Thus, where the theory of change that "small classrooms improves outcomes" was well established by the Tennessee study, the California study made vivid the importance of contextual conditions, to help elaborate what might have been an overly simple theory of change linking classroom size to academic outcomes.

15 Samii and Wilke (this Handbook) point out that harmonization of interventions helps decision-makers learn about differences in treatment effect across contexts if those effects differ across contexts.

16 For examples of this kind of center, see the AidData center at the College of William and Mary funded by a consortium of funders including both the U.S. Dept of State and USAID, and The Center on Human Trafficking Research & Outreach at the University of Georgia.

17 See, for example, USAID Public Access Plan, 15 which specifies dissemination plans. This kind of transparency can also help academics and practitioners navigate this tension.

18 We could also assess the effects of receiving both sources of information but would have much less statistical power for such an analysis.

19 For more on placebo-controlled designs, see Broockman, Kalla, and Sekhon (2017); Gerber and Green (2012); and Nickerson (2005).

20 See Kasy and Sautmann (2021) and Offer-Westort et al. (2021) for more on adaptive designs to guide policymaking. See Rabb et al. (2022) for an application of an adaptive design in a trial of SMS messages meant to inform decision-making by a public health agency.