An Approach to Small-Scale Mixed-Methods Experimentation

Transparency for Development, Phase 2

Stephen Kosack,[1,2] Jessica Creighton,[2] and Courtney Tolmie[3]

Introduction

Faced with a promising but complex intervention, how can further refinement be evaluated? The typical approach is experimentation. Rigorously evaluated experimentation, for several centuries the province mostly of medicine and related research, is today a reality in a variety of fields of social science and practice. Its primary form, the randomized controlled trial (RCT), stems from its medical roots. By design, RCTs are a highly specialized instrument of inquiry: they seek reliability by focusing on a simple, singular causal relationship. Their relevance typically depends on the relevance of this causal relationship and whether it can be accurately represented and measured in one treatment or in a handful of modifications (or "arms"). But the growth of experimentation has brought randomized controlled trials into evaluations of complex interventions in policy areas like health care, education, water, or sanitation, which often occur at the group or society level, at a large scale, and the implementation of which can take myriad forms. For precisely evaluating the benefits of complex programs, RCTs remain the gold standard, frequently used when, for example, a health or education program is under consideration for scaling, is already being done at large scale but is of uncertain benefit, or is almost perfected save for a very specific design question. But often large-scale randomized controlled trials of complex interventions are inappropriate. For an intervention whose benefit is already widely accepted, they may be too expensive; for one whose benefit is uncertain, they may be too large-scale. When, for whatever reason, evaluation of a complex intervention is important but small-scale experimentation is more appropriate than a full randomized controlled trial, how might further refinement of that intervention be most rigorously and reliably evaluated?

This paper describes one approach to such an evaluation. The approach is designed for exploring, through small-scale experimentation, focused design changes to an intervention of a common and widely evaluated kind that has shown early promise in two large-scale experimental trials. The approach is experimental, and thus retains several advantages of the RCT, including a perfectly identified cause as well as a control group to provide a counterfactual. But it is much smaller scale

[1] University of Washington, Seattle, WA
[2] Harvard Kennedy School, Cambridge, MA
[3] Results for Development Institute, Washington, DC

than a traditional RCT—involving trials in just fifteen villages, five each in three countries, as opposed to the dozens or hundreds of treatments involved in the typical large-scale RCT. Thus rather than seeking a precise estimate of an average impact across a large number of treatments, it focuses on understanding as fully as possible the variation around that impact, particularly variation in the intervention's intended mechanism. It does so by augmenting the typical experimental approach used in RCTs with a number of empirical methods designed to understand, reliably and from multiple perspectives, the intervention, the environment in which it was implemented, and the causal pathways it took on its way toward any effect. The root methodology is John Stuart Mill's methods of agreement and difference (Mill 1843), which take advantage of easily observed regularities in otherwise dissimilar situations. Into that basic comparative logic, the approach here tries to incorporate a variety of perspectives on rigorous and reliable social inquiry. Observed regularities are determined through structured observation of variation across a small number of carefully selected places, and, in a subset of those places, of variation in the implementation of the intervention, any causal process that results, and effects both expected and not. The methods by which these observations are collected are designed to verify or explore key expectations—observable implications of the conclusions of current theory and evidence about the intervention and its interaction with the place in which it is implemented. They are designed to be unbiased, systematic, and replicable across diverse settings, and integrated so as to use the best features of each to compensate for the disadvantages of the others.[4] In addition, the overall approach is designed in part to be open to subjectivity and induction, in order to be helpful to exploring a second kind of hypothesis: those inductively generated by participants, about factors or dynamics that they might have found influential to its impact but had not been considered in earlier theory and evidence.[5] By placing the observable implications of current theory and practice under maximal deductive and inductive scrutiny, from as many perspectives as possible, while also exploring explanations only apparent to intervention participants themselves, the approach here seeks to reliably identify plausible hypotheses and eliminate implausible hypotheses, and thereby advance current understanding of the implications of a complex, contextually dependent intervention beyond what is possible or practical with RCTs.

The remainder of the paper is as follows. Section 2 introduces the intervention and the research questions about it that this mixed-methods approach is designed to evaluate. Section 3 describes the basic goals and logic of the approach and its antecedents in scholarship and practice of international development and comparative methods more broadly. Sections 4-8 detail five elements of the approach that are intended to allow it to identify plausible hypotheses and eliminate implausible hypotheses about the intervention's impacts: careful selection of contexts; a control group; and broad and deep understanding from multiple perspectives of differences in contexts; the intervention (the event); and the causal processes from the event to the outcomes, if any. Section 9 concludes.

---

[4] In addition to Mill (1843), the approach draws from the conclusions and recommendations of scholarship in comparative methodology, including King, Keohane, and Verba (1994), Lieberman (2005), Seawright (2016) among others, as well as in international development, including scholarship at the World Bank (e.g. Woolcock 2013; Ananthpur, Malik, and Rao 2014) and the UK Department for International Development (e.g. Stern et al. 2012).

[5] In addition to being an important check against important sources of bias in the evaluation (e.g. confirmation bias, myopia), including such space for the views of participants themselves—about the process and the environment—is one of the currently accepted first principles of international development, as reflected in the 2005 Paris Declaration on Aid Effectiveness and the Accra Agenda for Action (http://www.oecd.org/dac/effectiveness/34428351.pdf).

1. Questions about a Promising Intervention

The approach developed here was designed to be specific to a particular intervention; this section describes the intervention, some existing evidence about its effects, and several unanswered questions about it of real-world relevance.

The intervention has the goal of engaging with members of rural communities in developing countries to improve their health care. The general family of methods of encouraging such engagement is called "transparency and accountability interventions"; they are an increasingly common approach that has been widely evaluated with large-scale randomized controlled trials. Transparency and accountability interventions typically seek improved governance within specific policy areas: resource extraction, for example, or the delivery of services like health care, education, water, sanitation, or infrastructure. They are particularly deployed for problems in government functions that involve face-to-face contact between citizens and the street-level bureaucracy. Such functions are often difficult to solve with resources and technical expertise; the central thesis of transparency and accountability interventions is that citizens themselves may be in a favorable position to improve many of those aspects of underperformance. But evidence about this thesis is mixed: a number of large trials of discrete transparency and accountability programs have found significant and substantial benefit; others, no benefit. The Transparency for Development project is a response to this mixed record: it is designed to explore whether such a transparency and accountability intervention can work, and, if so, where and why. From 2013-16, in the first phase of the project, a six-month intervention was co-designed with and implemented by civil society organizations in Indonesia and Tanzania in two large-scale randomized controlled trials, each involving 100 treatment and 100 control communities. These interventions provided selected community members with carefully curated information about problems with their maternal and newborn health care and a facilitated forum for discussing what, if anything, to do about those problems; the trial and its evaluation sought to determine if that discrete program empowered citizens to improve their maternal and newborn health care, and if so, where and how.

Reliable inferences about the impact of this intervention on health care and health outcomes await an endline survey (scheduled to be completed in late 2017 in Indonesia and mid-2018 in Tanzania). Yet relative to expectations, early signs are promising in several respects; in particular, several qualitative components of the evaluation, precursors to some of the techniques described below, strongly support three early conclusions:

1. Community members who volunteered to participate in the intervention often subsequently acted in ways that public health scholarship and experience suggests can improve health and health care outcomes (such as improving the facility or the community's relationship with the provider, or organizing transportation or raising funds to improve access to hard-to-reach facilities)[6];

[6] The sufficiency of this path for improving health and health care remains to be seen. Yet the identifiable problems with the performance of a public health care clinic may include many for which the actions of average citizens might plausibly make a difference. Filth, for example, is readily apparent, as is a lack of running water, privacy, toilets, or placenta pits.

2. those actions are varied, rather than converging around one approach[7]; and

3. those actions most often focused on the community and local government, such as the village leadership or officials in the local health facility; actions that engage with actors further up the government hierarchy were the minority.

Because the goal of transparency and accountability interventions is improved governance, the third preliminary conclusion raised an important question: why did so few community members choose to engage with officials further up the government hierarchy to resolve difficulties in their maternal and newborn health care?

One possibility is that official channels don't work: engaging with government officials—called the "long route" of accountability, in contrast to the "short route" of engaging directly with frontline service providers—is an ineffective way to resolve local issues with local service delivery; recognizing this, community members are opting for community-level approaches that they know to be more promising.

But a second widely-discussed possibility is that long route engagement is generally also locally effective, but citizens face barriers to productive engagement with government officials. Interventions like the Transparency for Development project's interventions in Indonesia and Tanzania are designed and implemented to be locally focused and community-designed and -led. Thus their design or implementation might tend to nudge participants toward more locally focused engagement than they would prefer if given options.[8] The practical implication of this hypothesis is

---

Some problems may indeed be more apparent to those who have experienced them than those with medical expertise: an unhelpful provider, for example, or unpredictable variation in the hours the clinic is reliably open and staffed, the prices it charges, or difficulty obtaining transportation to the clinic. These kinds of problems may also be within the capacity of average community members to improve, assuming that they value public health care and do not have an alternative to which to turn to get it. They can clean or repair the clinic, dig a well, talk to the nurse about their attitude and ask them how they can work together better, complain to the district health officer or their political representative, organize a transportation pool, put up a privacy wall, post hours and charges. Those kinds of solutions might make a difference both to the clinic itself and to the people who work there, inasmuch as they are able thereafter to do work that is more clearly noticed and valued by the community and for which community members are willing to contribute their own time and effort. Also within the reach of average community members' influence might be gaps in the knowledge of mothers about modern tenets of healthy birth that are generally accepted in the West (e.g. Centola 2011). Average citizens might even find solutions to more structural, supply-side problems, like the facility lacking electricity or being too far away to reach in a pinch; in many countries, for example, there are frequently forums in which community members gather to request development projects for their communities, and electrical supply of the clinic, or even an entirely new clinic, are among the acceptable requests.

[7] A common contention in existing work on transparency and accountability was that one approach was typically more successful (for example, collaborative or oppositional, or focused on service providers, local government, or regional or national government).

[8] A third possibility is that participants will engage productively with the long route on their own, even without CSO assistance, but not right away: instead citizens may seek to engage with local providers or officials first, and only go further up the hierarchy as they encounter obstacles to improvement at the lower levels. In this case, the pattern we are observing in Indonesia and Tanzania is temporary, and we will notice greater long-route engagement after the intervention has had more time to play out.

that civil society organizations should try to make the long route a more realistic possibility by relaxing barriers that community members face to productively engaging the long route.

The second phase of the Transparency for Development project is exploring the second possibility: can civil society organizations implement discrete programs to help citizens productively engage the long route, and in ways that have a discernable effect on health care?

The Interventions

To explore the promise of civil society organizations trying to enable and encourage citizens to productively engage the long route, three new interventions were designed[9] for evaluation in three new small-scale experiments in three new countries. In most ways these interventions are similar to the interventions that were showing signs of promise in Indonesia and Tanzania, but they have been adjusted, through a process of co-design with civil society organizations in three additional countries—Ghana, Malawi, and Sierra Leone—to try to increase one aspect of their impact: their ability to sustainably engender more positive engagement between citizens and government actors focused on improving their maternal and newborn health care.[10]

In particular, the three interventions:

1. follow the basic structure of the interventions designed for the Indonesia and Tanzania trials: they seek to improve health and health outcomes, via community empowerment, by providing information and a forum for discussing it and what to do in response to it;

2. have designs that all hew to the original design principles of the Tanzania and Indonesia interventions: they are co-designed with local partners; are health-focused not service-delivery-focused; locally relevant; community-led; non-prescriptive; and largely free of outside resources[11]; and

3. involve additional elements, developed during co-design discussions, that staff of the civil society organizations in Ghana, Malawi, and Sierra Leone predicted, based on their experience and understanding, would encourage and enable long-route engagement by communities with a pre-selected actor in the district level government (or equivalent, the next level up in the "long route," who had expressed support for the intervention and an interest in learning what comes of it). These include information given to participants about

[9] The phase 1 interventions in Indonesia and Tanzania were designed to be very similar but were co-developed, through a process of iterative design and piloting (Pritchett, Samji, and Hammer 2017) between practitioners of transparency and accountability at the Results for Development Institute in Washington, DC and in two civil society organizations, PATTIRO in Indonesia and CHAI in Tanzania, with the structured participation as well of researchers at the Harvard Kennedy School and the University of Washington. The Intervention Design Report as well as the intervention manuals and all materials that resulted from this two-year co-design process are available on t4d.ash.harvard.edu.
[10] What Lant Pritchett and colleagues have called "crawling the design space" (Pritchett, Samji, and Hammer 2017).
[11] Though the additional elements incorporated into each to relax constraints to long route action entailed substantially more outside material, relational, and technical resources than the designs used in the first phase of the project.

the health system hierarchy and how it functions, as well as additional meetings between participants and the pre-selected government actor.[12]

Three aspects of these three interventions are particularly relevant to the evaluation. First, they are all part of a larger program of research into the impact of transparency and accountability interventions on health and health care, which, as described above, also includes two much larger-scale randomized controlled trials of a similar intervention. Thus similarities between the five—the two large trials in Indonesia and Tanzania and the three much smaller trials in Ghana, Malawi, and Sierra Leone—will be helpful for understanding them as well as the broader family of interventions of which each is an example.

Second, because of the co-design processes used to develop these interventions, each of them, while deliberately similar in myriad ways, is also necessarily a somewhat different example of that broader family of interventions. In particular, the three (and those in Indonesia and Tanzania) differ in ways that reflect to some degree the unique places in which each was designed and implemented: its geography, its economy, its politics, its organizations, and the experiences, history, and traditions of the people who live there. To the degree that these or other factors have any influence on the intervention's mechanisms or impacts, dissimilarities in them limit the conclusions that can be drawn from the impact of any specific implementation of an intervention from this general family, including these.

Third, all are discrete by design. They (and those in Indonesia and Tanzania) are systematized and routinized for widespread training and facilitation, and they are implemented by specific organizations in specific places, at a specific point in time, for a specific problem in health, and at a specific societal level: the small rural community. These specifics and variations within them all potentially limit the generalizability of any of the conclusions of this evaluation, notwithstanding its attempts to account for them.

Questions

The evaluation developed here is designed to explore six questions about these three interventions, their impacts, and the interactions of each with the contexts in which each was implemented.

The first three questions replicate research questions from the Tanzania and Indonesia trials about how and why (if at all) the interventions lead participants to engage in actions that improve health and health care, whether it unfolds differently or has different kinds of effects in different places, and whether participants perceive their experience of it to be empowering and improving of their relationship with their government:

[12] As mentioned above, the co-design process for these three new interventions also adhered to the same design principles as in the first phase of the project; thus the design alterations focused on relaxing constraints to long-route action but were not, for example, more prescriptive, less community-driven, less focused on creating a design that was scalable and adaptable across contexts, or more nationally as opposed to locally relevant. The one exception is the final principle: all three designs relaxed constraints to long route action providing more outside material, relational, and technical resources than the designs used in the first phase of the project.

1. What mechanisms are triggered by the interventions that public health scholarship and experience suggests can improve health and health care outcomes?

2. What is the role of context in shaping or determining these mechanisms?

3. What are the implications of the interventions for citizens' perceptions of empowerment and efficacy, both within communities and between communities and the state?

Inquiries into these three in Ghana, Malawi, and Sierra Leone, as well as in the Indonesia and Tanzania trials—which use similar methods to those described below and will therefore be comparable to the results from Ghana, Malawi, and Sierra Leone—will provide an overall picture of the causal process(es) triggered by the intervention and its implications across varied communities and places.[13]

The second set of questions stems from the additional goal of relaxing the barriers to productive "long-route" engagement by participants.

4. Does adding to the intervention a formal connection with government actors who are—or who appear to be—willing collaborators lead participants to engage with the intervention differently or lead them to undertake different actions toward improving health care (in particular, more long-route approaches)?

5. To the extent that participants act to engage the long-route in response to the intervention, does the involvement of these government actors lead to an institutional response geared toward improving health care?

6. Do any of the differences in the designed process by which the three interventions engage government actors show promise for enabling and encouraging participants to engage long-route approaches and/or for encouraging an institutional response to participants' actions?[14]

3. The Basic Approach

[13] The Evaluation Design Report for the first phase RCTs is available at t4d.ash.harvard.edu.

[14] More specifically, alongside these specific questions, this inquiry is generally structured to continue the process of realizing the Primary Objectives of the T4D research program. In particular, it will provide additional assessment of the project's theoretical contentions and thus contribute to the development of a robust and empirically verified theory of T/A's impact (the Project's Primary Objective 1). It will also involve a process of refining our adaptable T/A intervention developed in Phase 1 into a form that might be useful to other CSO partners (Primary Objective 2). Although the major examination of impact is a part of Phase 1, Phase 2 will contribute to the overall picture of how T/A interventions influence health care quality and outcomes (Primary Objective 3), by exploring the degree to which these interventions trigger processes that are known to influence those outcomes. And finally, the research products of Phase 2—will provide material for dissemination and outreach to T/A and sectoral practitioners, scholars, and other stakeholders (Primary Objective 5). In consultation with the T4D steering committee, we determined that the remaining primary objective of the original proposal (Primary Objective 4, to determine generalizability and scalability of T/A interventions) was a lower priority for Phase 2 than the exploration of the new focal questions around encouraging and enabling long-route community action.

> "*Theophrastus* said, that human Knowledge, guided by the Sences, might judg of the Causes of things to a certain degree; but that being arriv'd to first and extreme Causes, it must stop short and retire, by reason of its own infirmity, or the difficulty of things."
>
> — Montaigne, "Apology for Raimond de Sebonde"[15]

Faced with a complex but promising intervention, how can current methods of inquiry and knowledge creation be integrated to most rigorously and reliably evaluate further refinements in that intervention? The approach developed here is one of further, small-scale experimentation in additional contexts and with design changes that are distinct but focused on the same improvement. The evaluation challenge is to generate evidence useful for knowledge and practice about the efficacy of these innovations in particular and of this kind of intervention in general. In this way, relative to the typical focus of large-scale randomized controlled trials on internal as opposed to external validity, the goal here is a refocused and somewhat expanded scope of inquiry: refocused on the variation around the causal pathways by which the intervention has an impact, so as to better understand their nature, implications, and—this being an evaluation of a type of intervention common in the world today—whether they come with hitherto unknown side effects; and expanded to be more generally valid by including further contexts and potentially further causal pathways. It seeks that understanding through gathering and examining a set of observations that altogether, with as little bias as possible, reflect the true variation around a complex intervention, implemented in a small number of communities, in ways that are similar but different in one complex respect: their effort to relax barriers by those communities to productive engagement with government officials.

Antecedents

The approach seeks to integrate methods and thinking about them from several well-developed and widely used comparative research techniques from across the social sciences. It follows from a recent focus by academic methodologists on integrating the insights of diverse comparative techniques into "mixed" or "multi" method research. One result of the experimental turn in many of the social sciences was a wide range of interventions whose causal impacts are well-identified but whose causal processes are not well-understood or that vary with the particular setting in which the experiment is conducted; mixed methods research has long promised greater insight into the complex implications of a complex intervention's causal pathways, not only its average effects.

The conclusions of this scholarship are diverse. Although many scholars favor single-method evaluations[16] and raise important questions about the advantages of mixed methods,[17] a number of

[15] Montaigne went on to reflect on the slow but steady progress of human knowledge, arts, and sciences in understanding and accurately measuring certain things, as well as a tendency to overestimate the things within human capacity to understand, a tendency he experienced and which he believed tempers progress in knowledge.
[16] In development, for example, Banerjee and Duflo (2011) or Karlan and Appel (2011).
[17] For example, Beck (2006, 2010), Ahmed and Sil (2009), Kuehn and Rohlfing (2009). See Seawright (2016) for a discussion.

scholars and organizations argue that methodological eclecticism has substantial advantages in at least some research situations. In particular, the sort of evaluation in question here—of a complex intervention designed to improve a complex outcome through a complex adaptation[18]—has been specifically cited by scholars and practitioners favoring mixed methods research, who argue that they can offer more generalizable, useful, and ideally more practically applicable understanding that stands up to scrutiny from multiple perspectives (Seawright 2016; Woolcock 2013; Lieberman 2005; King, Keohane, and Verba 1994).[19]

Transparency and accountability interventions are inherently complex (e.g. Björkman and Svensson 2009; Lieberman, Posner, and Tsai 2012; Fox 2015),[20] and the early indications from the large trials in Indonesia and Tanzania do nothing to suggest that their impacts are simpler than was previously thought.[21] In both countries, participants in different communities engaged with the intervention in highly varied ways. Some were indifferent; some dropped out; many made plans and acted on them in ways that have clear causal linkages to improved health and health care but are otherwise dissimilar—problem-solving, complaining, seeking reform, or just asking others for help. The variation clearly reflects a complex intervention inducing a complex reaction.

The Transparency for Development project is designed to offer useful evidence about the efficacy of transparency and accountability interventions for improving health, including whether such interventions improve health and health care, and, if so, where and why. Above we noted an additional pattern in the Indonesia and Tanzania trials of particular practical import to the community of practice around transparency and accountability: participants engaged relatively little with government actors in trying to improve their health care. Two alternative explanations for this

---

[18] For the interventions in this second phase of the evaluation, the complex adaptation was to encourage productive long-route engagement by participants.

[19] Within the field of international development, the approach is consistent with those advocating for an eclectic methodological approach, including those who emphasize the importance of evaluation methodologies tailored to the intervention (Levy) and of small-scale experimentation of complex interventions informed by learning and iteration. Pritchett et al. (2017) argue for an experimental approach alongside and often instead of large-scale randomized controlled trials. the Doing Development Differently manifesto (http://doingdevelopmentdifferently.com) focuses on helping organizations learn by figuring out for themselves what is working and what isn't (see also Stern et al. 2012); the Goldilocks Project (Gugerty and Kaplan 2017) seeks to offer guidance for non-governmental and civil society organizations seeking to offer evidence of their efficacy meeting measurable donor goals. The broader move toward experimentalism in the social sciences (see, e.g. the argument in Shapiro 2016 that experimentalism is vital for understanding how to improve the human condition) has also led many to firmly believe that mixed methods are never appropriate. Recently the preferred method has been RCTs; in political science, see, for example, the work of Alan Gerber and Don Green (e.g. Gerber, Green, and Larimer 2008; Gerber and Green 2012) or Chris Blattman (Blattman, Hartman, and Blair 2014; Blattman and Annan forthcoming) or in economics, Abhijit Banerjee and Esther Duflo (e.g. Banerjee and Duflo 2011) or Dean Karlan (e.g. D. S. Karlan 2005; Ashraf, Karlan, and Yin 2006). But as noted, there are also many who have argued for methodological eclecticism and have provided extensive guidance on the best way to integrate mixed methods (Seawright 2016; Lieberman 2005; King, Keohane, and Verba 1994).

[20] Transparency and accountability interventions are complex: they share common features, but those common features—the presentation of specific information and a facilitated forum for discussing what, if anything, to do about that information—leave room for myriad design choices and have myriad influences on the place they are implemented.

[21] The Transparency for Development transparency and accountability intervention was tasked with improving a very specific set of measurable outcomes in health and health care for mothers and children. More generally, transparency and accountability interventions seek community empowerment and improved state-society relations.

sort of pattern are prominent in debates among scholars and practitioners in the transparency and accountability field: long-route engagement not working vs. the existence of barriers to productive long-route engagement that civil society organizations can help relax. The design implications of civil society organizations seeking to relax barriers to long-route engagement raised a number of additional questions with readily apparent observable implications. The Transparency for Development project's second phase seeks to further expand the evidence base around the same questions as in the Indonesia and Tanzania trials while also focusing particularly on the efficacy of civil society organization efforts to relax barriers citizens face to productive long-route engagement. One option is to evaluate such efforts with further large-scale randomized controlled trials. But this option is impractical: too large and expensive, considering that whether the intervention works at all, on average, is still in question.

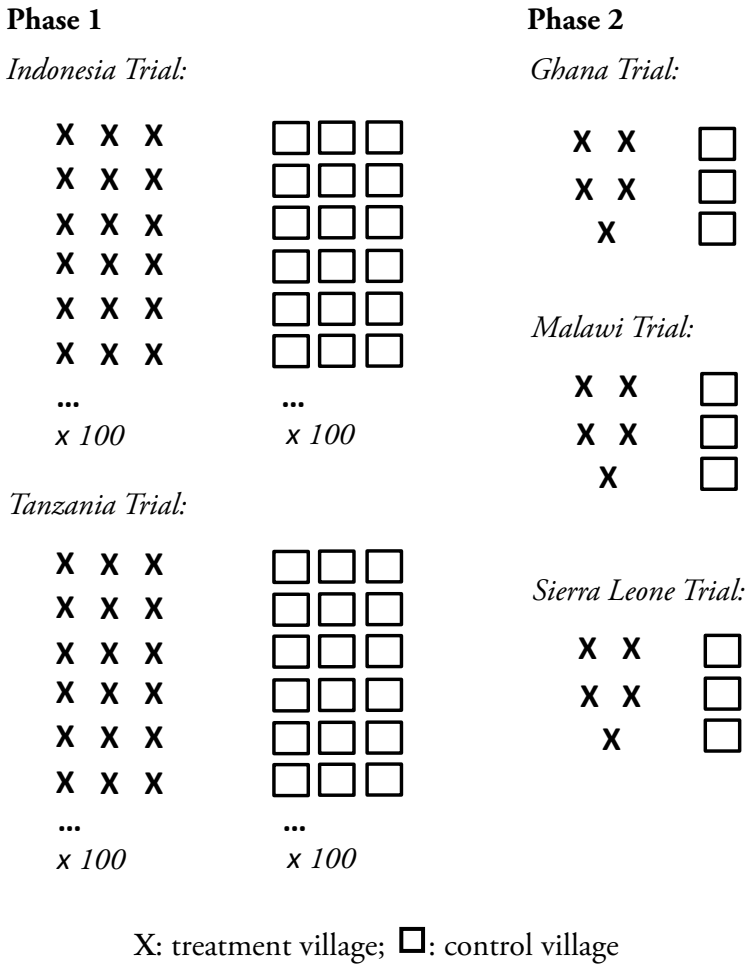
Instead, the approach to answering these questions in the project's second phase draws from the scholarly and practical traditions behind the arguments mentioned above: it is an attempt to be tailored to the intervention, experimental but at a small scale, and it is designed to build an understanding of the intervention's myriad potential causal implications, through careful integration of multiple methods providing different but complementary perspectives on those implications.

Part of the appropriateness of this approach to the evaluation rests on its similarities to and linkages with interventions that are common practice and, in particular, are very similar to those implemented in the earlier, large-scale experimental trials in Indonesia and Tanzania that provided the signs of its early promise noted in section 1 above.[22] The second phase of the project involves three new interventions that, relative to those earlier interventions, involve specific innovations are all geared toward a specific end: improving the intervention's ability to sustainably engender more positive engagement between citizens and government actors, and thereby lead to measurably greater community empowerment, better health care, and better health outcomes. In all other respects, the three are highly similar to those already being evaluated at large scale in Tanzania and Indonesia. In addition, the evaluation approach adopts several features of those earlier experimental evaluations; in particular, it seeks to understand the impact of a discrete cause—the three new interventions—on a discrete and measurable set of implications of that cause, measures by varying that cause between treatment and control groups and observing the implications in both.

But the approach here has important differences to the earlier trials in Indonesia and Tanzania, with more modest goals reflecting its far smaller scale. The 200 randomly selected treatment and control communities in Indonesia and Tanzania—a "fully powered" randomized controlled trial—will allow a high degree of confidence that any average difference in measured community empowerment, health care, and/or health outcomes between treatment and control communities is the causal result of the interventions in those two countries. The fifteen additional treatment communities receiving the three new interventions are not enough to permit that kind of definitive proof. Thus the goal of the approach is plausibility, the sort of plausibility that might eliminate important possibilities—such as that the intervention has no impact at all, or that particular factors that influence its impact—perhaps motivating larger-scale implementation and evaluations if things are promising. Figure 1 shows the relative scale of the Phase 1 and Phase 2 trials.

[22] As recommended most recently by Seawright (2016).

Figure 1 - Treatment and Control Villages in the Phase 1 and Phase 2 Trials



The conclusions resulting from the approach will therefore depend to a significant degree on the conclusions of those larger trials. In particular, the surveys at baseline and endline of the randomized controlled trials will be the final word in whether this intervention works, because the patterns in those surveys will offer a highly reliable estimate of whether the intervention improved maternal and newborn health and health care, on average, in the treatment communities relative to the control communities. In comparison, the particular approach here seeks to answer, on a smaller scale, supplementary questions resulting from those larger trials.

Yet even at smaller scale, the central challenge in evaluating each of these three new interventions remains the same: to generate evidence about them and their immediate implications that is useful for knowledge and practice in the community of practice of transparency and accountability with only fifteen examples of how communities in three countries experienced each one. In other words,

how to learn as much as possible from a set of pilots? More specifically, the goal is an evaluation that reveals as much as possible about whether the set of interventions plausibly cause (or trigger progress toward) a set of reliably measurable outcomes like a greater proportion of long-route approaches or greater institutional responsiveness to those overtures, cause those outcomes only in certain kinds of places, or are irrelevant to those outcomes, as well as whether key differences in the intervention's design are likely to be influencing how it causes those outcomes. In other words, did three transparency and accountability interventions in Ghana, Malawi, and Sierra Leone plausibly improve government responsiveness and community empowerment geared toward improving maternal and newborn health, did they plausibly improve them only in certain places or only when they had certain design features, or did they appear to be irrelevant to those outcomes? And did they appear to have any other effects?

We base the logic for exploring these questions on two simple contentions. First, whenever a similar event occurs in places that have little else in common except the event, and a similar observable outcome follows in all those places, the event is plausibly the cause of the outcome, or at least among its causes.[23] Second, whenever two or more events or places are similar except for one characteristic, and a similar outcome follows only when that characteristic is present, that characteristic is plausibly necessary to the outcome. These basic insights, articulated in Western scientific lexicon by John Stuart Mill in 1843 as, respectively, the *method of agreement* and the *method of difference*, provide a basis for a set of questions that can be asked of any circumstance in which an event might be a cause of an outcome, but in which key differences in the event may influence the outcome and key characteristics of the environment may influence both the event and the outcome.

The first contention provides guidance for establishing unusual regularities in otherwise regular situations—for example, whether an event is plausibly the cause, or at least among the causes, of an outcome. When an event occurs in even a small number of places, an outcome follows in those places, and the event can be shown to correspond to the outcome, the event is plausibly among the causes of that outcome. For example, let's say that I set out to assess the causal relationship between driving over the speed limit and getting pulled over by a cop, and I run a test in which I try to drive 20 miles over the speed limit in five cities. If I succeed in driving 20 miles over the speed limit in all five cities, and in each one I end up getting pulled over and given a speeding ticket, I can conclude that driving 20 miles over the speed limit is plausibly among the causes of getting a pulled over and receiving a speeding ticket:

Place	*Event*	*Outcome*
City 1	Speeding	Pulled over; received speeding ticket

[23] The event is plausibly the cause of the outcome, but it is not guaranteed to be: it is always possible that some other factor, which happens to be present wherever the event occurred, is the cause, and the focus on the event leads to a spurious correlation between it and the outcome. This "omitted variable" bias, which is a primary motivator for RCTs, is often a concern with small-*N* research. It can be mitigated to some extent by observing the event and any change in the outcomes from multiple vantages, and by process-tracing the causal path from the event to the outcome and verifying its observable implications. We take both approaches here; they are outlined in more detail below.

City 2	Speeding	Pulled over; received speeding ticket
City 3	Speeding	Pulled over; received speeding ticket
City 4	Speeding	Pulled over; received speeding ticket
City 5	Speeding	Pulled over; received speeding ticket

But characteristics of those places can also shape both the event and the outcome in ways that change the relationship between the two. For example, I may set out to conduct the uniform test above by trying to drive 20 miles over the speed limit in all five cities, but run into unexpected difficulties that differ across the cities: in one I run into heavy traffic that prevents me from speeding over a long distance, and I am not pulled over until the traffic lets up and I am able to speed; in another I run into traffic that briefly prevents me from speeding, but am stopped as soon as I am able to speed; in two more I get pulled over after speeding for over an hour; and in the last city I am pulled over before I even have the chance to speed:

Place	*Characteristic*	*Event*	*Outcome*
City 1	Sustained traffic	No speeding, then brief speeding	Pulled over after brief speeding; received speeding ticket
City 2	Brief traffic	No speeding, then brief speeding	Pulled over after brief speeding; received speeding ticket
City 3	No traffic	Sustained speeding	Pulled over after sustained speeding; received speeding ticket
City 4	No traffic	Sustained speeding	Pulled over after sustained speeding; received speeding ticket
City 5	No traffic	No speeding	Pulled over despite not speeding

If I had been able to run my test in 100 cities, these sorts of differences may not have mattered—I might have seen that, on average, when I sped, I got pulled over and received a speeding ticket. But what can I conclude from these five examples about whether speeding will get me pulled over? Does speeding cause a cop to pull me over, does it cause a cop to pull me over only over long distances or only in certain places (for example, those where police have incentives to issue speeding tickets), or do I get pulled over regardless of whether I speed?

The second of Mill's contentions can be of great help in making sense of this more complex set of relationships between place, event, and outcomes, by providing guidance for establishing whether variation in the outcome is plausibly related to variation in the event, the environment, or both. Whatever is characteristic of some events or places but not others is a potential cause of any difference in either the event or the outcomes observed between that place and others. But where we observe variation between events or places that is not reliably associated with differences in outcomes, we might eliminate those as plausible explanations for those differences. The cities in which I attempt to speed likely vary on countless dimensions, but in hindsight, one observable characteristic, traffic, clearly influenced the event: when traffic was heavy, I could not speed, and I did not receive a speeding ticket.

Yet the way I conducted my test makes it hard to eliminate explanations. Although I can conclude from the observed variation that traffic impedes speeding and that speeding is plausibly necessary to receiving a speeding ticket, I cannot conclude that any of the characteristics I observed—distance I drove, traffic, and speeding itself—did not play a role in whether I was stopped. I was pulled over after driving short distances and long distances; where there was traffic and where there was not; and when I sped and when I did not. Nor can I account for other potentially important characteristics of the cities where I conducted my test that I did not observe, such as the incentives of police officers in the different cities for pulling drivers over.

The approach to the three small-scale trials in Ghana, Malawi, and Sierra Leone is designed to allow it to identify plausible hypotheses and eliminate implausible hypotheses more effectively than my speeding test. As noted, these trials are nested within a broader project that also includes two large-scale randomized controlled trials. In addition, it includes five improvements to the sort of test I conducted about speeding and traffic stops: 1) purposeful selection of contexts, 2) a control group, and broad and deep understanding from multiple deductive and inductive perspectives of differences in 3) context, 4) the event, and 5) the causal process from the event to the outcomes, if any. These five are each detailed in subsequent sections of this paper.

The five improvements draw from the enormous body of methodological inquiry in communities of scholarship and practice that has focused for decades on how to draw reliable inferences about complex relationships between places, events, and outcomes in a small sample.[24] Such discussions have a long tradition in research about international development, the practice of which involves a large number of complex interventions that typically play out differently in different places. The mixed-method debate in international development has particularly emphasized both experimentation and participation—in intervention design and implementation, as well, increasingly, in evaluation, reflecting the adoption of participation and partnership as core values in international development more generally.[25] Among others, DFID (e.g. Stern et al. 2012) and the World Bank (e.g. Woolcock 2013; Ananthpur, Malik, and Rao 2014) have recently offered new scholarship and tools for mixed methods and participant-oriented approaches as part of their research operations, as well as asking for those approaches to be used in both internal and external evaluations of their work. Even organizations that have long championed strict adherence to a specific form of large-scale randomized controlled trial, such as J-PAL or IPA, are incorporating more qualitative techniques into their work and exploring ways that other organizations can more easily incorporate them as well.[26]

[24] Notable contributions in the extensive literature on the advantages and disadvantages of research combining qualitative and quantitative methods of inquiry and inference, and how best to integrate them to understand causal processes, include King, Keohane, and Verba (1994), Lieberman (2005), Beck (2006, 2010), Brady (2008), Ahmed and Sil (2009), Kuehn and Rohlfing (2009), Mahoney (2008, 2010), and Seawright (2016). Such questions are also being debate in private and nonprofit policy research organizations such as the Urban Institute, Mathematica, or National Opinion Research Center at the University of Chicago.

[25] See for example the 2005 Paris Declaration on Aid Effectiveness and the 2008 Accra Agenda for Action (http://www.oecd.org/dac/effectiveness/34428351.pdf).

[26] For example, the Goldilocks Project (http://www.poverty-action.org/goldilocks).

The approach herein builds from these discussions in the following ways. In keeping with the existing literature's general focus, it is designed to observe the processes resulting from a perfectly identified cause—a treatment offered in a number of carefully selected communities, which are then compared to a control group to provide a counterfactual—in such as way as to gain the most thorough understanding practicable and possible of them and their implications. It seeks to do so by carefully integrating rigorous and reliable observations focused around the observable implications of current understanding of those causal processes (King, Keohane, and Verba 1994; Lieberman 2005; Brady 2008; Mahoney 2008; Seawright 2016). These observations and their integration is an attempt to heed Seawright's recent (2016) call for a back-to-basics approach that clearly anticipates the core advantages and disadvantages of each approach, and integrating them to allow each to reflect what it is best equipped to reflect—ideally with enough regularity to survive statistical scrutiny. In particular, it combines several small-scale trials designed to explore small changes to a promising intervention with several large-scale randomized controlled trials designed to explore that intervention's causal mechanisms and impacts, which is what Seawright recommends for the situation here, in which an RCT provides the precise estimate of an average effect and a qualitative component is designed to explore context and causal mechanisms.[27]

Seawright's focus on careful integration leads him to criticize another common approach to mixed-method inference in qualitative studies that is used extensively in the approach developed here: triangulation, or verifying an observation (or observable implication of a hypothesized process) with the findings of different methods. As detailed below, the approach uses triangulation to integrate the findings of different methods, such as structured surveys and observations employing anchoring vignettes with unstructured interviews, focus groups, observation, and participation in the intervention, around a number of questions that are difficult to reliably observe, such as whether the interventions nudge community decision-making toward long-route approaches or whether such engagement, when it happens, is productive. Seawright rightly notes the difficulty that apples-and-oranges techniques have adjudicating among explanations found by one technique but not another: triangulation is clearly inferior to a division of labor among methods when the advantages and disadvantages of each are clear and orthogonal. For example, the small-scale pilots in Ghana, Malawi, and Sierra Leone are highly inferior for assessing overall impact relative to the large Indonesia and Tanzania trials. But the approach also includes questions whose observable aspects are generally difficult to reliably observe. When integrating methods whose relative value to a reliable perspective on an observation is unknowable, or at least unknown, no one approach is sufficient to accept or reject a given hypothesis. But triangulation can help to eliminate implausible assumptions or hypotheses and suggest those that are plausible by focusing on agreement between multiple perspectives. As with an event that is observed to cause the same outcome in disparate places, an observation about whether an event, an outcome, or a characteristic of a place is consistent

[27] Specifically, in this situation Seawright (2016) recommends an integration wherein the RCT is used to infer causation and the qualitative methods are used "to design, test, refine, or bolster the analysis producing that inference," in particular by verifying that key assumptions in plausible causal chains were met. The approach developed here to verify these key assumptions is similar to a kind of experimentation Seawright describes as setting the "agenda" for comparative-historical analysis, in which "an experiment is replicated, using similar groups of subjects, in multiple societies, social classes, institutional contexts, and so forth" and patterns of similarities and differences become the variation to be explored with within-case evidence and process tracing "to help unravel the causal and historical structure behind the contemporary pattern of causal effects."

with a hypothesis about that event, outcome, or characteristic is more reliable when disparate perspectives from different methods point in the same direction, either for or against the hypothesis. Their agreement suggests something more general. Indeed this basic intuition about agreement and disagreement underlies Mill's development of the comparative method itself (Mill 18xx). Triangulation can thereby help to identify plausible and implausible hypotheses, distinguishing both from hypotheses of uncertain plausibility: those on whose validity two methods, again of indeterminate relative value, point in different directions, so that it is in doubt but cannot be eliminated.

In seeking to reliably observe predicted implications of an intervention for an event, an outcome, or a characteristic of a place, the approach here is designed to do more than contribute to causal inference; it seeks descriptive inference as well. Several decades ago, King, Keohane, and Verba (1994) argued for a unified conceptualization of scientific empirical research, whether qualitative or quantitative, as having four characteristics: a goal of inference (causal or descriptive); transparent and clearly defined research procedures; uncertain conclusions; and strict adherence to methods. They developed two major criteria for judging the resulting inferences, whether causal or descriptive: inferences are valid if they are unbiased—correct on average—and efficient—deviating little from that average.

The approach developed here seeks to follow these guidelines. It seeks observations about the interventions, the places they were implemented, and their implications, particularly for hypothesized processes and outcomes, that are unbiased and efficient. It is also an attempt to follow most of the rules that King, Keohane, and Verba lay out for good social science: to ask a question of demonstrated real-world importance; to approach it with theory that is consistent both internally and with prior evidence, falsifiable, testable (in the sense of generating multiple observable implications about data that were not used to create the hypotheses), and concrete; and to measure it with data (meaning information regularly and reliably recorded and reported) that is maximal in the depth and breadth of the observable implications it accurately describes, and valid, by consisting largely of reliable, replicable measurements defined ex ante.

Such rules are broadly consistent with arguments specifically in the field of international development about how to increase quality and rigor in mixed method research, with an important exception. An example of their consistency is Spencer et al. (2003), an analysis influential to work commissioned by the United Kingdom's Department for International Development on "broadening the range of designs and methods for impact evaluation." Spencer and colleagues emphasize a shared set of goals—truth, applicability, consistency, and neutrality—for all research, including "scientific" and "naturalistic," that are largely in line with the rules of good social science laid down in King, Keohane, and Verba.[28] DFID (Stern et al. 2012) uses the Spencer et al. (2003)

[28] DFID's report (Stern et al. 2012), echoing Spencer et al (2003), argues that these criteria simple go by different names among "scientific" or "naturalistic" researchers: truth is *internal validity* in scientific researcher and credibility in naturalistic research; applicability is *external validity* in scientific research and transferability to naturalistic research; consistency is *reliability* in scientific research and dependability in naturalistic; and neutrality is *objectivity* in scientific research and confirmability in naturalistic.

framework to develop a central quality-assurance framework for evaluating five design approaches in research: experimental, statistical, theory-based, case-based, and participatory.

The approach developed here is a combination of all five. As noted, it is experimental and statistical, as it involves several experiments of several interventions and the collection of quantitative data; theory-based, in drawing from existing theory and evidence to identify gaps in current knowledge, influence the structure of the intervention, and motivate the purposive selection of cases; and case-based in most of its methods of selection and understanding.

Finally, the approach here is participatory, in relying heavily on the views of participants themselves about the process and the environment and leaving room for hypotheses inductively generated from the participants. In addition to being an important check against important sources of bias in the evaluation (e.g. confirmation bias, myopia), including such space for the views of participants themselves—about the process and the environment—is, as noted above, one of the currently accepted first principles of international development reflected in the Paris Declaration on Aid Effectiveness.[29] Indeed, one of the few examples of the integration of an RCT that incorporated purely inductive techniques is an evaluation of an international development intervention (Ananthpur, Malik, and Rao 2014). That intervention, like the interventions this approach is designed to evaluate, was intended to induce citizen participation to improve the quality of local-level government, in that case through a citizenship training and facilitation program in rural India. A randomized controlled trial of 100 treatment and 100 control villages was combined with an in-depth ethnography in which researchers lived in or near five treatment and five control villages, spent four years exploring the village and its history, and the dynamics triggered around the intervention.[30] Something similar was included with the larger evaluations in Indonesia and Tanzania as an independent data source: in each country, researchers lived in four treatment and two control communities for between 9 months and one year around the intervention, seeking to understand the participants' perspective on it to the maximal degree an outsider can. For reasons of resource constraints, the approach developed here combines the two: a single researcher will collect

[29] See also Spencer et al. (2003).

[30] Ananthpur, Malik, and Rao (2014) describe their approach to the ethnography: "From 2007-2010, each GP was assigned a field investigator, typically someone with an MA degree in a social science or in Social Work, who was from the region and therefore very familiar with the milieu and dialect and easily able to blend into the community and establish rapport. The investigator either resided in the GP or in a location that was a short, easily accessible distance away. In the first round of reports each investigator mapped the village's social and political structure, outlining the various caste and religious groups residing in the GP, relationships within and between them, structures of social networks and power, major events in the GP's history including its experience with development projects, etc.

"Subsequently, once a month, the investigators sent in a 5-10 page report on important changes that had taken place. S/he was instructed to record important local events, interview important actors in those events, investigate new village constructions and the financing behind them, track electoral activities and expenditures, examine changes in levels of local activism, and investigate other issues that were relevant to the political and economic life of the GP. In treatment GPs they were, in particular, asked to closely track the work of the KSIRD RPs, and to follow up on how their work percolated into the village, and the sequence of changes that were initiated by the work of the RPs. From 2010-2011, the team was reduced to three investigators who visited all the GPs on a rotating schedule sending in reports every three months. Consequently we have a total of about 400 reports divided equally between treatment and control GPs. These village reports, supplemented by regular field visits by the principal investigators, constitute our qualitative data that we distill and draw on for this qualitative section of the paper" (p. 9-10).

both the objective and subjective data. It tries to compensate for this compromise by including open-ended questions in every data collection tool, and asking that researcher to be responsible not only for exploring predetermined questions but also for inductively developing an understanding of the village and how the intervention influenced it and was influenced by it.

These participatory elements are consistent with the typical arguments about appropriate use of mixed methods in international development. But they are inconsistent with King, Keohane, and Verba's (1994) rules, because they deliberately leave substantial room for unobserved and unmeasurable concepts to "get in the way."[31] The methodological reason the approach makes use of participatory elements was noted in the introduction. It is designed to explore two kinds of hypotheses: those widely circulating in current theory and practice, whose observable implications it explores by employing rigorously deductive data collection, as well as emergent hypotheses, inductively generated by participants, about factors or dynamics that they thought were influential to its impact but that had been unconsidered in earlier theory and evidence.

This deductive and inductive data collection is also designed to allow for reliable comparisons to the villages in Indonesia and Tanzania that experienced the first phase intervention. In addition to open-ended induction, the particular data collection methods used include surveys (some with anchoring vignettes), focus groups, interviews, structured observation of meetings, networks, as well as "social action plans" developed by participants as part of the intervention. In both form and content these techniques echo those used to collect information on the impact of the first phase interventions, in a nested sample called the "T4D Onion," which is displayed in Figure 2.

[31] Some practical mixed-method efforts adopt this perspective as well. For example, the Goldilocks Project recommends that NGOs undertaking mixed-method evaluations narrow in on only the information that they commit in advance to acting on, so as to prioritize efficiency and practicality.

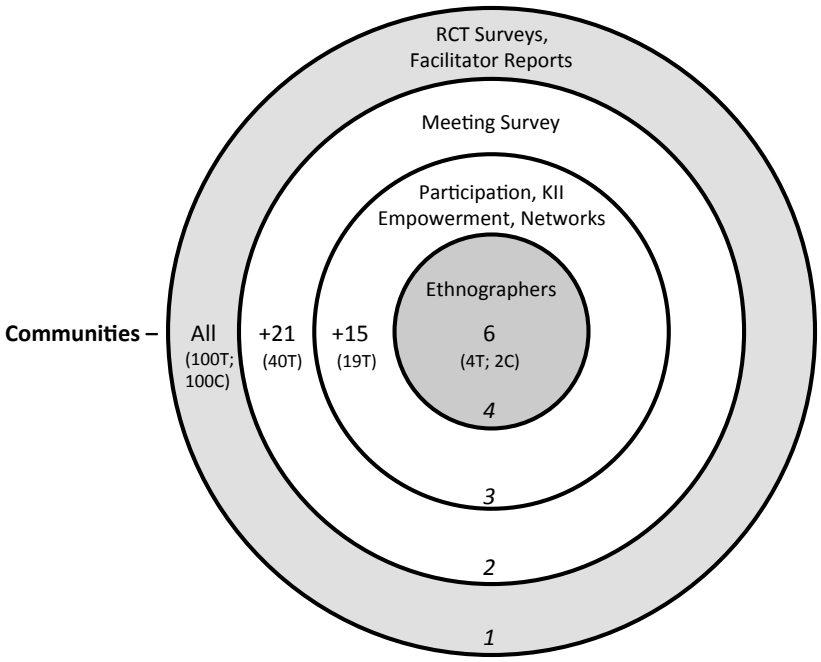


Figure 2 - The T4D "Onion."[32]

In sum, the goal of the evaluation approach in the second phase of the Transparency for Development project is not to equal the rigor and internal validity of the larger Phase 1 trials. Rather, it is to all these three smaller trials to produce neutral, consistent, unbiased conclusions that approximate truth, are broadly applicable, both on their own and in combination with the findings of the larger Phase 1 trials. The approach seeks this goal through a nested design that allows comparisons of the three small trials with the two much larger randomized controlled trials in Indonesia and Tanzania, and with five improvements to the sort of haphazardly chosen and implemented test of speeding and traffic stops described earlier in this section: 1) purposeful selection of contexts, 2) a control group, and broad and deep understanding from multiple perspectives of differences in 3) context, 4) the event, and 5) the causal process from the event to the outcomes, if any. Recall that in the five-city experimental speeding test it was not clear whether distance, traffic, city, speeding itself, or unobserved characteristics like police incentives was playing a role in whether I was pulled over and given a ticket. Applied to that speeding test, the approach developed here would add five features: more careful selection of the cities, broad and deep measurement of differences in those cities, controlled variation in the event, observation of police reaction to my speeding, and a control group.

The remainder of this document describes something similar for the three small trials in Ghana, Malawi, and Sierra Leone. Each of the five improvements contributes either to observing elements of the research questions described in the previous section, or to understanding other elements unforeseen but potentially important.

[32] For details on the elements of the onion, see the Phase 1 Evaluation Design Report available at t4d.ash.harvard.edu.

The table below notes the research questions to which each aspect is designed to contribute.

			Understanding of variation in		
Questions	*Careful selection of villages*	*Control group*	*Context*	*Intervention (the event)*	*Causal Process*
1. What mechanisms are triggered by the intervention that public health scholarship and experience suggests can improve health and health care outcomes?		■	■		■
2. What is the role of context in shaping or determining these mechanisms?	■		■	■	■
3. What are the implications of the interventions for citizens' perceptions of empowerment and efficacy, both within communities and between communities and the state?			■		■
4. Does adding to the intervention a formal connection with government actors who are—or who appear to be—willing collaborators lead participants to engage with the intervention differently or lead them to undertake different actions toward improving health care (in particular, more long-route approaches)?	■		■	■	■
5. To the extent that participants act to engage the long-route in response to the intervention, does the involvement of these government actors lead to an institutional response geared toward improving health care?		■	■		■
6. Do any of the design differences in the manner in which the three interventions engage government actors show promise for enabling and encouraging participants to engage long-route approaches and/or for encouraging an institutional response to participants' actions?			■	■	■

Altogether these aspects are designed to 1) allow its conclusions to be less biased and more efficient, 2) to permit it to more clearly identify a cause and a counterfactual, 3) to offer a greater likelihood that its methods will uncover useful patterns in otherwise dissimilar situations, and 4) to contextualize those patterns in light of understanding existing theory and practice, placing the observable implications of current theory and practice under maximal deductive and inductive scrutiny by observing as many of those implications as possible, from as many perspectives as possible, while also searching for explanations only apparent to participants themselves.

4. Careful Selection of Contexts

The first of these additional elements is more careful selection of contexts, in this case villages within regions in three countries in which the interventions are implemented. Recall the two insights that provide the underlying logic of our approach. First, whenever a similar event occurs in places that have little else in common except the event, and a similar outcome follows in all those places, the event is plausibly the cause of the outcome, or at least among its causes. Second, whenever two or more events or places are similar except for one characteristic, and a similar outcome follows only when that characteristic is present, that characteristic is plausibly necessary to the outcome.

Ideally, we would implement the intervention in a group of places that allow us to understand the role of as many of the relevant dimensions as possible: i.e. several groups of places that are 1) similar to each other on all the relevant dimensions but 2) as a whole vary from each other on these dimensions. The large number of randomly selected communities in the Indonesia and Tanzania trials was large enough to provide variation on many potentially relevant dimensions.

In order to achieve something similar in these three smaller trials, we will still rely on random selection to some extent, in order to minimize bias in the selected communities. But the far smaller number of communities will require us to select purposefully on some dimensions, as well as to check the variation across randomly selected communities and adjust it, to achieve a sample that is balanced across countries and groups of communities and includes the desired variation on relevant dimensions.

The first selection question is the choice of countries. Ghana, Malawi, and Sierra Leone, though all African, vary widely in geography, history, culture, politics, and development—including the nature and development of the their health systems.[33] Thus the three offer a wide diversity of settings well-suited to an examine common patterns in the interventions' impacts with an approach based on Mill's Methods. Yet the three countries were also selected purposefully to be similar in one crucial respect: each has a local civil society organization with substantial local knowledge and experience—community-level knowledge of the local context and experience working on similar interventions—

[33] By World Bank classifications, Ghana is a lower-middle-income country; Malawi and Sierra Leone are both low-income countries. Ghana is a democracy, rated as fully "free" by Freedom House; Malawi and Sierra Leone are both partial democracies, rated "partly free," with substantially less political competition.

as well as a willingness to actively engage in the aforementioned co-design process, from which the innovations to the Indonesia and Tanzania interventions derived.

The second question is the choice of region in each country in which to implement the intervention. Here again the choice was made purposefully in one important respect: each region was selected to have at least one public official who was willing, in principle, to engage positively with the intervention. In developing the interventions that are under scrutiny in Indonesia and Tanzania, we hypothesized that that a number of contextual factors would be important to how a transparency and accountability program might induce an improvement in a public service like health care (Kosack and Fung 2014). One of these hypothesized factors is the willingness of officials to be responsive to citizen demands was key to productive long-route engagement. This willingness varies—not all governments are designed and managed to be maximally responsive to average citizens—and for citizens seeking improvements to their health care, official willingness may influence whether engaging with the long route is the path of least resistance to positive change. Thus given the research questions of this evaluation—about the potential for civil society organizations to sustainably engender more positive engagement between citizens and government actors—the willingness of government officials was determined to be a crucial contextual factor. Consequently, in each country one administrative district was selected whose district government included at least one willing public official—a "government champion."[34]

But willingness is only one of a large number of community differences that may be relevant to how this transparency and accountability intervention plays out and whether it makes a difference to outcomes of interest; other factors include:

1. The existing quality of the health care system
2. The degree to which the particular health subsector is perceived by the community to be a problem, both absolutely and relative to other issues.
3. Whether there are multiple health clinics or other kinds of health care accessible to the community (exit options)
4. The willingness of front-line service providers such as nurses or midwives to engage with community efforts to improve health services
5. Whether there is an elected leadership and degree of political competition
6. Level of trust
7. Civil society
8. Socioeconomic characteristics (ethnicity, religion, wealth, education)

For maximal variation on these factors, within the chosen region we select villages randomly, not purposefully, and then check for their representativeness of overall variation and balance between groups of villages. The general idea is similar to the randomization process in an RCT, though the

[34] In Kosack and Fung (2014) this is "world 4" of a five "world" theoretical framework of the political economy contexts in which transparency was likely to follow a different path of least resistance to reform. Such environments are of particular practical import to the project, to the extent that places where governments see the value of citizen efforts to improve health care are places where a transparency and accountability program is most likely to have the chance to be implemented at scale, and thus are the places where evidence about the effectiveness of an intervention from the transparency and accountability family would be most salient to official decision-making, to the extent that government responsiveness depends on structure as well as intention (e.g. Evans 1995)

implementation includes a bit more room for imbalance correction. As the particular selection process also involves assigning villages into treatment and control groups, it is described in detail in the following section.

5. A Control Group

The second of the additional elements in the approach—a control group—helps with inferences about how the intervention played out and whether it made a difference to outcomes of interest. Qualitative and mixed method research often suffers from difficulties in causal identification, as they often lack a true counter-factual. By selecting of a group of treatment villages to receive the intervention, as well as others that will not and that can serve as a control group, the approach here retains the experimental advantage of a perfectly identified cause. To reiterate, the resulting inferences will not have the high degree of internal validity enabled by larger RCTs; rather, their goal is to provide important facets of the answers to the research questions, by focusing on whether observable implications of hypotheses about them were in fact observed. In particular, comparison with a control group will help with inferences about any implications of the intervention for health and health care as well as any institutional response to it.

a. Implications for health care

The first research question is about mechanisms by which the intervention could plausibly affect health and health care in general, wherever it was implemented. For this question, as with the Phase 1 interventions, the relevant comparison is a set of places where the intervention did not occur but which are otherwise similar to those places where it did occur.

But the Phase 2 control groups will differ from Phase 1's in two fundamental ways. First, the Phase 2 communities are not numerous enough to provide a reliable signal of trends in health and health care outcomes in the absence of the intervention. The purpose of the Phase 2 control groups is therefore to provide only a comparison set, to put whatever changes are observed in the treatment communities into context. Interviews and focus groups in the control communities will provide a basis for comparing them to the treatment communities prior to the latter receiving the intervention, and interviews and observations as part of a facility survey implemented at the same two time-points as in the treatment communities will allow a comparison with any observable evolution in the health care systems over the period of the intervention. In particular, the facility observations will allow us to assess certain aspects of facility quality, such as cleanliness, availability of water and electricity, and availability of drugs, supplies, beds and other medical resources.

The second difference is the selection of the control communities. In Phase 1, control groups were determined by randomly selecting a set of communities and using stratified random sampling to assign half of them to receive the treatment, and then verifying that as a group the treatment communities were similar to the control communities on observable characteristics thought to be relevant. Because of the far smaller number of communities in the Phase 2 tests, random selection will not necessarily result in a set of control communities that are similar to the treatment

communities on relevant characteristics. Instead, to ensure that these control communities provide a relevant comparison, we will select them purposefully, to vary on one dimension potentially relevant to the community's ability to engage in effective social actions: existing healthcare system capacity. The focus on health system quality stems from the goal of measuring change over time in implications for health care, as well as the hypothesis that it will be another contextual factor that is important to how the intervention plays out and its ultimate effects. In particular, we expect the most community-driven improvement in the facility to occur where that facility is sufficiently high-capacity to demonstrate its value to the community, but not so high-capacity that community members cannot easily imagine making it better. For both these reasons, we will seek rough parity between treatment and control communities on existing health system quality.

In addition to health system quality, balance between treatment and control communities will subsequently be checked on other dimensions noted in the previous section that are observable and potentially relevant to the community's capacity to engage in social actions, such as community wealth, education, exit options, and proximity to larger towns or cities.

The selection process will proceed as follows:[35]

1. Within the selected district,[36] we will identify all the public health facilities providing primary care services and identify the villages officials served by the facility ("catchment-area" villages).

2. We will randomly select five treatment facilities and three control facilities, as well as two additional facilities to serve as backups in the event that the overall sample is severely unbalanced. Where proxy data for existing health system quality is available in advance,[37] we will attempt first to stratify facilities, by creating small groups each of two or more facilities whose quality is most similar to each other, and then randomly selecting treatment or control facilities from within these groups.

3. From the catchment-area lists, we will exclude villages based on certain criteria, including 1) pre-existing similar programs and 2) population (very large and very small).

4. We will randomly select one village from the catchment area of each of the selected facilities. Those villages selected from the catchment area of a treatment facility will receive the intervention; those villages selected from the catchment area of a control facility will not.

5. The variation between the five treatment facilities in health system quality will be assumed to represent the variation in the district more generally; this will be verified in the control

[35] This process is subject to revision based on our access to pre-existing district and country data, as well as actual circumstances in each location.

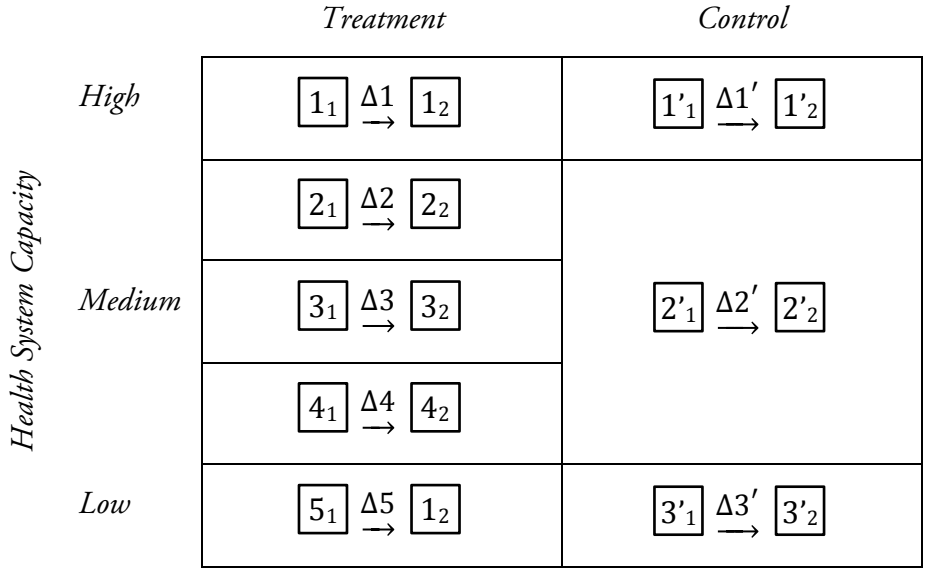
[36] The district in which our partner CSO has identified its "government champion"—the government official who has expressed an interest in the intervention and a willingness to try it out.

[37] Such as the physical condition and availability of supplies of the facility, or its distance from major population centers or referral hospitals.

communities and, if it appears based on early research that the facilities are biased in one direction, we will adjust the sample accordingly.

6. Finally, we will then check for balance in the control and treatment groups across observable aspects of health care quality and the additional contextual features noted above. To the extent that the initially selected control communities are substantially different from the treatment communities, we will conduct additional research in the backup control facilities and replace mismatched control communities with better-matched backups to ensure a control set that is as balanced as possible.[38]

The table below shows the comparisons that this design will allow, with several example hypotheses for what we would expect to observe in these comparisons. (Boxes are villages; villages 1, 2, etc. are treatment villages and 1′, 2′ etc. are control villages; subscripts are times 1, before and after the intervention and 2, after the intervention. The deltas—$\Delta 1$, $\Delta 2$, etc.—are changes in the outcomes of interest in village 1, 2, etc. between time 1 and 2, before and after the intervention.)



Comparisons among changes in each treatment and control village will allow us to test a number of widely circulating hypotheses about the way the intervention influences health care.[39] For example,

[38] Full balance on all relevant dimensions is likely to be impossible, given the same sample size. Instead the goal is a control set that is as balanced with the treatment set as possible.

[39] King, Keohane, and Verba (1994) note that such comparisons are most unbiased and efficient when there is homogeneity among observable units; in our case, for example, that an observed change in health care would be observed and measured similarly across villages. The data collection tools developed for this evaluation seek this "unit homogeneity" by employing standardized questions and responses across places, several of them calibrated with anchoring vignettes. (The exceptions are the open-ended questions that run throughout the tools and that are designed to allow for inductively generated hypotheses.)

if the intervention improves health care, we should observe agreement among the changes in the treatment villages relative to the control villages:

$$\Delta 1 > \Delta 1', (\Delta 2, \Delta 3, \Delta 4) > \Delta 2', \Delta 5 > \Delta 3',$$

where the deltas describe measure improvements in observed characteristics of health care. Second, if those effects differ measurably according to the existing capacity of the public health system, such that the most community-driven improvement occurs where the health system is sufficiently high-capacity to offer decent service and thus demonstrate its value to the community, but not so high-capacity that community members cannot easily imagine making it better, we should observe that the improvements in villages 2, 3, and 4 are greater than those in villages 1 and 5:

$$\Delta 1 < (\Delta 2, \Delta 3, \Delta 4) > \Delta 5.$$

b. Institutional response

The control communities will also provide an important facet of the answer to question 5: the institutional response. Change in government responsiveness will be assessed through a series of key informant interviews, in which a sample of government actors and citizens are asked a set of structured questions about government responsiveness and presented with a set of vignettes to calibrate their responses.

It is possible, of course, for institutional responses to be particular: targeted at the particular communities that requested them.[40] But the nature of a truly institutional response is that it influences all communities within the government's jurisdiction, including the control communities. Thus spillovers are expected to some degree, and indeed are partly the point of the intervention. The control communities will help to understand the nature of these spillovers. In particular, they will help us to determine three possible types of institutional response:

1. a response that affected all communities within the government's jurisdiction;

2. a response that represented a marginal improvement, affecting only the treatment communities and leaving the control communities unchanged; or

3. a response that came at the expense of the government's attention to other communities.

[40] Indeed more locally particular government action (at the expense, for example, of more concentration on public goods) is thought to be characteristic of less institutionalized and encompassing governance, and thus may be the dominant mode of response in the less democratic countries where we are implementing Phase 2 (e.g. Bueno de Mesquita et al. 2003; Acemoglu and Robinson 2006).

Control communities provide an opportunity to check for these kinds of spillovers, by suggesting whether communities that did not receive the intervention saw their health-relevant governance improve, worsen, or stay the same:

	Change in [Perceived] Responsiveness	
Type of Institutional Response	*Treatment Communities*	*Control Communities*
Encompassing	Increase	Increase
Particular	Increase	Unchanged
Reallocation	Increase	Decrease
None	Unchanged	Unchanged

The table below summarizes the evidence used to assess these changes:

Change in	*Evidence*
Health system quality (question 1)	Facility survey (before and after intervention, in both treatment and control communities)
Government responsiveness (question 5)	Key informant interviews with government actors, citizens, and outside observers (before and after intervention, in both treatment and control communities)

6. Observation of and data on the causal pathway(s):

The small sample size in Phase 2 makes it difficult to rigorously assess these questions of change—in health and government responsiveness—from a single perspective like a survey or a set of interviews. There is always the potential for bias in data from a single survey, such as omitted variables that influence the dynamics of the intervention but were not considered (or cannot be considered) in the survey. To mitigate this possibility, we seek to assess change from several perspectives and, with rare exceptions, will limit our conclusions to those on which those multiple perspectives agree. Another of these perspectives is the causal process, or dynamics, by which the intervention led to any observed change. Observations of the causal process and evidence of it will focus on whether observable implications of expected mechanisms were in fact observed in treatment communities, and whether they were different in different contexts (research question 2). These observations and evidence will further buttress our understanding of questions 1 and 5—mechanisms leading to changes in the health system and in government responsiveness—and will allow us to assess question

3—participants' perceived empowerment—and 4—encouraging and enabling long-route approaches (see section 2 on Research Questions above).

The expected causal process triggered by the intervention derives from the logic model of the intervention [include figure when this is finalized]. The process is complex, with multiple pathways to impact each with multiple steps, and somewhat dependent on a number of assumptions about the intervention's implementation and reception by community participants. The goal of the evaluation is to gather enough reliable information on the observable elements of different assumptions and hypotheses about pathways within this causal process to eliminate those that are implausible (for which the evidence disagrees with the hypothesis), plausible (with which the evidence agrees), or uncertain (for which evidence from different perspectives point in different directions). Evidence gathering will therefore focus in particular on observable elements of assumptions and pathways within this complex causal process that make it more or less likely to succeed in improving maternal and newborn health care, via community empowerment, by providing information and a forum for discussing it and what to do in response to it. It focuses on the following questions:

1. Did facilitators collect local data and stories for use in the intervention meetings?
2. Who are the participants (age, gender, education, formal or informal community leaders, etc.)
3. Do participants attend the meetings?
4. Do participants productively participate in meetings?
5. Do participants 1) understand the information presented, 2) use it to identify barriers to improvement and 3) devise viable social action plans for alleviating those barriers?
6. What actions are directed toward government (long-route actions)?
7. Do participants carry out those social action plans?
8. What are the pathways of each action? In particular:
    1. the ultimate target and any intermediate allies or opponents (including whether it represents a long or short route approach, self-help, or exit);
    2. for each,
        1. the approach toward each actor (collaborative or oppositional),
        2. the request of that actor, and
        3. the response of that actor (action that engages the problem or engages with others to fix the problem; lip-service or rhetoric; or nothing); and
    3. the ultimate response (including whether it achieves the participants' objectives and/or is plausibly linked to improvements in health care)
9. Do participants adapt their plans in response to successes and challenges?

In addition, because the focus of Phase 2 is on long-route approaches, we will also focus on a set of questions around any government response triggered by the intervention:

10. In any meeting that includes government participants, what (if anything) do those participants say that they will do?
11. What do those government participants actually do?
12. Does government responsiveness change, as perceived by both citizens and public officials?

13. Does provider responsiveness change in response to any government action (as opposed to short-route engagement by citizens), as perceived by citizens, providers, and public officials?

Finally, we will examine implications of the process for participants' perceptions of empowerment:

14. Do participants' perceptions of their own empowerment change after going through the process?

To increase reliability of the evidence on each question, each of these questions will be assessed from different perspectives with multiple methods. In all, the evidence on the causal process will derive from seven methods. The first two were introduced above:

1. A facility survey to assess certain aspects of facility quality, such as cleanliness, availability of water and electricity, and availability of drugs, supplies, beds and other medical resources. (Facility Survey; FS)
2. Specialized key informant interviews to assess the degree of government responsiveness, in which a sample of government actors, citizens, and other observers are asked a set of structured questions about government responsiveness and presented with a set of vignettes to calibrate their responses (Government key informant interviews; Gov-KIIs)

Additional forms will include:

3. The social action plans that the treatment communities develop (Social Action Plans; SAPs)
4. Key informant interviews of the facilitators, the participants, and a sample of those the participants engage as part of their actions (KIIs)
5. Structured observation of the intervention meetings (a stripped down version of the "standard coding scheme" used to observe selected meetings in the Phase 1 evaluation; SCS)
6. A short module to assess participants' perceptions of their ability to improve their communities, presented with a set of vignettes to calibrate responses (ES)
7. Facilitator reports (FRs) and scorecard data collective by the partner organization

The table below summarizes the questions to which each form will contribute evidence:

	Form of evidence						
Question	*FS*	*Gov-KIIs*	*SAPs*	*KIIs*	*SCS*	*ES*	*FRs*
1. Did facilitators collect local data and stories for use in the intervention meetings?					■		■
2. Who are the participants (age, gender, education, whether they are formal or informal community leaders, etc.)				■	■		■
3. Do participants attend the meetings?					■		■
4. Do participants productively participate in					■		■

Form of evidence

Question	*FS*	*Gov-KIIs*	*SAPs*	*KIIs*	*SCS*	*ES*	*FRs*
meetings?							
5. Do participants 1) understand the information presented, 2) use it to identify barriers to improvement, and 3) devise viable social action plans for alleviating those barriers?					■		■
6. What actions are directed toward government (long-route actions)?			■	■	■		
7. Do participants carry out those social action plans?		■	■	■	■		■
8. What are the pathways of each action? In particular: 1. the ultimate target and any intermediate allies or opponents (including whether it represents a long or short route approach, self-help, or exit); 2. for each, 1. the ask 2. the approach toward each actor (collaborative or oppositional), and 3. the response of that actor (action that engages the problem or engages with others to fix the problem; lip-service or rhetoric; or nothing); and 3. the ultimate response (including whether it achieves the participants' objectives and/or is plausibly linked to improvements in health care)		■	■	■	■		
9. Do participants adapt their plans in response to successes and challenges?		■	■	■	■		■
10. In any meeting that includes government participants, what (if anything) do those participants say that they will do?				■	■		
11. What do government participants actually do?		■		■			
12. Does government responsiveness change, as perceived by both citizens and public officials?	■	■		■		■	

	Form of evidence						
Question	*FS*	*Gov-KIIs*	*SAPs*	*KIIs*	*SCS*	*ES*	*FRs*
13. Does provider responsiveness change in response to any government action (as opposed to short-route engagement by citizens), as perceived by citizens, providers, and public officials?	■	■		■			
14. Do participants' perceptions of their empowerment change after going through the process?				■		■	

As the table makes clear, these questions will be largely answerable from more than one perspective, thus minimizing the potential bias in any one vantage.

7. Measurement of Differences in Context

An understanding of variation in context underpins the approach to all of the research questions, not only the role context may nor may not play in the process triggered by the intervention (research question 2). Recall the two contentions that underpin the logic of this evaluation. First, whenever a similar event occurs in places that have little else in common except the event, and a similar outcome follows in all those places, the event is plausibly the cause of the outcome, or at least among its causes. Second, whenever two or more events or places are similar except for one characteristic, and a similar outcome follows only when that characteristic is present, that characteristic is plausibly necessary to the outcome.

The simplest scenario is one where, like the speeding example in section 2, the same event is associated with the same outcome across different contexts. Wherever this is the case across varied contexts, there is reason to think that those contextual differences were relatively unimportant, casting doubt on any hypotheses in which they play an important role. For example, if we observe similar movement in key outcomes—citizen empowerment, government responsiveness, and/or progress toward health system improvements—following from a similar process triggered by interventions in all five communities within a country, we can conclude that any contextual differences between those communities were unimportant to the process the intervention triggered, because a similar process led to similar outcomes no matter these contextual differences. For example:

$$(\Delta 1_1 \cong \Delta 2_1 \cong \Delta 3_1 \cong \Delta 4_1 \cong \Delta 5_1) \cong (\Delta 1_2 \cong \Delta 2_2 \cong \Delta 3_2 \cong \Delta 4_2 \cong \Delta 5_2) \cong (\Delta 1_3 \cong \Delta 2_3 \cong \Delta 3_3 \cong \Delta 4_3 \cong \Delta 5_3)$$

where subscripts 1, 2, and 3 index for the three countries—Ghana, Malawi, and Sierra Leone.

Things are more complex when the relationship between the process and the outcomes differs between places. As noted in discussion of the logic of this approach, the second of Mill's contentions can help these cases, allowing us to learn from the particular similarities and differences that we observe in complex and contingent relationships.

For example, one possibility is that in all three countries, the relationship between the process and the outcome differs between communities in ways that reflect one key difference in the characteristic of these communities. Above we laid out one such potential characteristic: the quality of the existing health system prior to the intervention. We hypothesized that the most community-driven improvement might occur where the health system is sufficiently high-capacity to offer decent service and thus demonstrate its value to the community, but not so high-capacity that community members cannot easily imagine making it better. If we see this pattern—the largest improvements in the communities where the health system was of medium quality—repeated across all three countries, we can conclude that it is plausibly a generally important contextual characteristic, i.e.:

$$(\Delta 1_c - \Delta 1'_c) < \left((\Delta 2_c, \Delta 3_c, \Delta 4_c) - \Delta 2'_c\right) > (\Delta 5_c - \Delta 3'_c),$$

where subscript c indexes for country—Ghana, Malawi, and Sierra Leone.

Another possibility is that the relationship between the process and the outcomes is similar across all the communities within each Phase 2 country, but differs dramatically across each of the three countries, i.e.:

$$(\Delta 1_1 \cong \Delta 2_1 \cong \Delta 3_1 \cong \Delta 4_1 \cong \Delta 5_1) \neq (\Delta 1_2 \cong \Delta 2_2 \cong \Delta 3_2 \cong \Delta 4_2 \cong \Delta 5_2) \neq (\Delta 1_3 \cong \Delta 2_3 \cong \Delta 3_3 \cong \Delta 4_3 \cong \Delta 5_3)$$

In this case, we may conclude that any contextual characteristic that is similar between the countries is not plausibly an important characteristic to the differing patterns of process and change in outcomes across the three countries. But that is about the limit of what we can say: any characteristic whose influence we did not specifically seek to observe in the causal process tracing and that differs and between the countries is plausibly important. Mill's contentions rely on patterns in the similarities between event, outcome, and place; but where those patterns overlap, it is unable to tease out which similarities are important.

This limitation is particularly apparent in a final scenario: where there are no similarities either between countries or between communities within a country in the relationship of process and outcomes. In this case, the method developed here will not help us to draw conclusions, or even eliminate plausible hypotheses, about the role of context in the relationship between process and outcomes. Instead, in this scenario, we will need to rely on the process tracing to suggest plausible sources of variation, which then have to be developed by us or others into new hypotheses for more focused exploration in other contexts.

To measure relevant contextual characteristics of communities, we will rely on key informant interviews and focus group discussions. In addition to measuring certain aspects of health and

health system quality, and thus putting any changes we observe into context (see section 5 above), these interviews and focus group discussions will also assess other easily observable and potentially relevant contextual characteristics, including those described in section 4:

1. The degree to which the particular health subsector is perceived by the community to be a problem, both absolutely and relative to other issues.
2. Whether there are multiple health clinics or other kinds of health care accessible to the community (exit options)
3. The willingness of front-line service providers such as nurses or midwives to engage with community efforts to improve health services
4. Whether there is an elected leadership and degree of political competition
5. Level of trust
6. Civil society
7. Socioeconomic characteristics (ethnicity, religion, wealth, education)

This primary research will be supplemented by desk research on national characteristics that might be relevant to any differences between the Phase 2 countries, such as political institutions and the degree of political competition, the degree of decentralization, wealth, population density, and colonial history.

8. Controlled Variation in the Event

The final feature of the evaluation is exploration of variation in the event itself: the intervention. The intervention will vary both between Phase 1 and Phase 2 and among Phase 2. First, all Phase 2 interventions differ from the Phase 1 in one respect: they are designed to encourage and enable more long-route approaches by participants (research question 4). Second, the Phase 2 interventions will differ from each other in subtle, but potentially important, ways, some of which may show particular promise (research question 6). For example, one may include an extra set of meetings between participants and government actors; another may include extra interaction by staff from the civil society organization with government officials early on to prepare them for the intervention. The differences in how the interventions enable long-route approaches are limited, generally to one notable change or set of interrelated changes. But because the design of this evaluation bases inferences on patterns in the relationship between the process and the outcomes, variation in the intervention is a complication that, depending on the patterns observed, may reduce the confidence of any conclusions about the intervention's impact, particularly conclusions about the role of context in shaping this relationship (research question 2).

Still, to the extent that the intervention is implemented consistently *within* each Phase 2 country, this variation may also permit exploration of the ways in which design differences are associated with different relationships between the process and outcomes.

First, variation within countries can support inferences about design differences that do *not* matter. The simplest scenario is that the relationship between process and outcomes is the same across places where the intervention is different; in this case, it is plausible to infer that the design differences do

not matter to the relationship of process and outcomes. This is an important possibility, as one of the key questions of this evaluation is whether the additional efforts to encourage and enable government engagement change the way the intervention is received, particularly whether it leads to more long-route approaches (research question 4). In particular, we can compare the 15 Phase 2 communities to a set of matched pairs from the Phase 1 countries. If the prevalence of long-route approaches is no higher in Phase 2 than in similar Phase 1 communities, we can conclude that the Phase 2 design elements did not encourage and enable more long-route approaches.

The opposite, however, is not necessarily the case: if we find more long-route approaches among all Phase 2 countries compared to a matched set of Phase 1 communities, this does not mean necessarily that the design difference is driving that increase. Any remaining contextual differences not accounted for with the matching of Phase 2 with Phase 1 communities—including, necessarily, all national-level contextual differences—are also candidate explanations. The same is true of any promising design differences among Phase 2 countries. In one Phase 2 country that has a slightly different Phase 2 design, we may observe that differences in the outcome relative to the control communities are much higher. But those differences may not be the result of the design differences, but rather differences in that country context.

Our goal in such circumstances will not be conclusive evidence for a particular design difference or contextual factor, but rather to eliminate implausible factors and design differences. We will do so by 1) carefully considering variation in the context that does not coincide with variation in the intervention, and 2) deriving *ex ante* observable implications of the major design differences among the Phase 2 countries, which we will build specifically into the process tracing (section 6).

The first, carefully considering contextual against design variation, begins again from Mill's insights, specifically that any contextual factor that does not vary with differences in the relationship of the process and the outcome is an implausible cause of those differences. To illustrate, let us say that in Country 1 of Phase 2, the intervention is designed to involve government officials much earlier in the process (or differs in an important way from the other two), and we observe far greater changes in outcomes relative to the control communities in Country 1 than in Country 2 or 3. Either this design difference or some contextual feature of Country 1 could be responsible for the intervention's seemingly greater impact. Let us furthermore say that our Phase 2 countries differ primarily along three observable dimensions: wealth, population density, and degree of political competition, as follows:

	Country 1	*Country 2*	*Country 3*
Wealth	Medium	Medium	Medium
Population density	Medium	Medium	Low
Political competition	High	Low	Low

In this scenario, we could conclude that wealth and population density are far less plausible causes of the greater changes in outcomes in Country 1 than political competition, because Country 2 and 3 share features of wealth and population density but are different on political competition. Thus we are left with political competition and design differences as the most plausible explanations.

The second element allows further progress, by verifying observable implications of the importance of the design difference in our process-tracing (section 6). In this scenario, we could look for two observable implications of the main design difference, namely earlier involvement of government officials: 1) that we observe far higher responsiveness (as measured by the government key informant interviews—see section 5 above), because this earlier involvement helps to build trust and get the process off on the right foot; and 2) that it makes the institutional response more particular as opposed to encompassing (section 5), because government officials have been primed to be responsive to the communities making the requests and not (necessarily) others. If we observe evidence of these two in the process-tracing, we could conclude that design differences are an additional plausible explanation (alongside political competition) for the greater changes in outcomes that we observe in Country 1.[41]

Using this basic approach, we will draw conclusions about the plausibility and implausibility of various contextual and design differences in determining any differences we observe in the relationship of process and outcomes. To reiterate, the emphasis here is on *plausibility.* There is always a chance that even a contextual factor that is not correlated with any difference in the relationship of process and outcomes is still influencing that relationship indirectly, just as there is always a chance that a design difference that appears important is only spuriously related to the relationship between process and outcome. Such spurious correlations and interactive or mediating factors can be examined if the sample of examples is large enough. Given our small sample size, the goal of the evaluation is to make as much progress as possible on the research questions, as well as to set the stage for more targeted exploration of the most promising design factors and contextual differences our small sample reveals.

9. Conclusion

Faced with a promising but complex intervention, how can further refinement be evaluated? The approach developed here is designed to explore, through small-scale experimentation, focused design changes to a community-based health care intervention, of a common and widely evaluated type, that has shown early promise in two large-scale randomized controlled trials in Indonesia and Tanzania, each involving 100 treatment communities. For precisely evaluating the benefits of complex programs, such large-scale experimental trials remain the gold standard, and the conclusions of those larger trials will be the final word in whether this intervention works, on average, to improve maternal and newborn health and health care. Yet early indications from those trials have also raised

[41] Note that in this case we would conclude that the design difference was plausibly related to the greater "impact" of the intervention in Phase 2, but not necessarily that it was a promising innovation for encouraging an institutional response to participants' actions (research question 6) because the response was particular rather than encompassing (see section 5).

a question of particular practical import—whether civil society organizations implementing interventions like this one can engender more positive engagement between citizens and government officials—for which further large-scale trials are inappropriate: insufficient for understanding the causal processes that may result from these design changes and how and why they might vary among contexts; and too expensive, given that the efficacy of the intervention in general is still in question. The alternative approach developed here is designed to be far less costly and, relative to the typical large-scale RCT, to refocus and somewhat expand the scope of inquiry: refocusing it on the variation around the causal pathways resulting from the intervention so as to better understand their nature, implications, and whether they come with hitherto unknown side effects; and expanded to be more generally valid by including further contexts and potentially further causal pathways.

In brief, this alternative approach is experimental, and thus retains several advantages of the randomized controlled trial that is the typical approach to such an evaluation, including a perfectly identified cause as well as a control group to provide a counterfactual. But rather than seeking a precise estimate of an average impact and a small number of sources of difference in that impact, it focuses on understanding the variation around that impact as fully as possible, particularly variation in the intervention's intended mechanisms. It does so by augmenting the typical experimental method (surveys of participants at baseline and endline) with a number of empirical methods designed to understand, reliably and from multiple perspectives, the intervention, the environment in which it was implemented, and the causal pathways it took on its way toward any effect.

The root methodology is John Stuart Mill's methods of agreement and difference, which take advantage of easily observed regularities in otherwise dissimilar situations. Into that basic comparative logic, the approach here incorporates a variety of perspectives on rigorous and reliable social inquiry. Observed regularities are determined through structured observation of variation across a small number of carefully selected places, and, in a subset of those places, of variation in the implementation of the intervention, any causal process that results, and effects both expected and not. The methods by which these observations are collected are designed to verify or explore key expectations—observable implications of the conclusions of current theory and evidence about the intervention and its interaction with the place in which it is implemented. They are designed to be unbiased, systematic, and replicable across diverse settings, and are integrated so as to use the best features of each to compensate for the disadvantages of the others. In addition, the overall approach is designed in part to be open to subjectivity and induction, in order to be helpful to exploring a second kind of hypothesis: those inductively generated by participants, about factors or dynamics that they might have found influential to its impact but had not been considered in earlier theory and evidence. The goal is a set of observations that altogether reflect, with as little bias as possible, the true variation around a complex intervention, implemented in a small number of communities, in ways that are similar but different in one complex respect: their effort to relax barriers by those communities to productive engagement with government officials.

The approach is necessarily somewhat specific to the intervention being evaluated. Yet with minor modifications, it may be relevant for many evaluations of complex interventions designed to improve complex outcomes through complex, contextually dependent adaptations—particularly those for which large-scale randomized controlled trials have produced vague or inconclusive results. One

result of the experimental turn in many of the social sciences is a wide range of interventions whose causal impacts are well-identified but whose causal processes are not well-understood or that vary with the particular setting in which the experiment is conducted. In this respect, the overwhelming reliance of evaluations in recent years on the randomized controlled trial has severely limited understanding of the implications of complex interventions, including many that are being regularly used in the world today. This is particularly true in international development, the practice of which involves a large number of complex interventions that, like the type of intervention under consideration here, typically play out differently in different places.

For such interventions, additional experimentation, assessed with rigorous comparative social science research techniques, may yield rigorous and reliable inferences, even at small scale. These methods do not equal the internal validity of a randomized controlled trial, nor are they intended to. But the argument here is that comparative methodologies, drawing on the insights of the enormous and generations-old body of methodological inquiry in communities of scholarship and practice, can substantially improve the ability of even a small test to produce neutral, consistent, unbiased conclusions that approximate truth and are broadly applicable. By placing the observable implications of current theory and practice under maximal deductive and inductive scrutiny, from as many perspectives as possible, while also exploring explanations only apparent to intervention participants themselves, such techniques can reliably identify plausible hypotheses and eliminate implausible hypotheses, and can thereby advance current understanding of the implications of complex interventions beyond what is possible or practical with RCTs.

References

Acemoglu, Daron, and James Robinson. 2006. *Economic Origins of Dictatorship and Democracy*. New York: Cambridge University Press.

Ahmed, Amel, and Rudra Sil. 2009. "Is Multi-Method Resarch Really 'Better'?" *Qualitative and Multi-Method Research* 7 (2): 2–6.

Ananthpur, Kripa, Kabir Malik, and Vijayendra Rao. 2014. *The Anatomy of Failure: An Ethnography of a Randomized Trial to Deepen Democracy in Rural India*. Policy Research Working Papers. The World Bank.

Ashraf, Nava, Dean Karlan, and Wesley Yin. 2006. "Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines." *The Quarterly Journal of Economics* 121 (2): 635–672.

Banerjee, Abhijit V, and Esther Duflo. 2011. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. New York: PublicAffairs.

Beck, Nathaniel. 2006. "Is Causal-Process Observation an Oxymoron?" *Political Analysis* 14: 347–52.

———. 2010. "Causal Process 'Observation': Oxymoron or (Fine) Old Wine." *Political Analysis* 18 (4): 499–505.

Björkman, M., and J. 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.

Blattman, Christopher, and Jeannie Annan. forthcoming. "Can Employment Reduce Lawlessness and Rebellion? A Field Experiment with High-Risk Men in a Fragile State." *American Political Science Review*.

Blattman, Christopher, Alexandra Hartman, and Robert Blair. 2014. "How to Promote Order and Property Rights under Weak Rule of Law? An Experiment in Changing Dispute Resolution Behavior through Community Education." *American Political Science Review* 108 (1): 100–120.

Brady, Henry E. 2008. "Causation and Explanation in Social Science." In *The Oxford Handbook of Political Methodology*, edited by Janet Box-Steffensmeier, Henry E. Brady, and David Collier. New York: Oxford University Press. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199286546.001.0001/oxfordhb-9780199286546-e-10.

Bueno de Mesquita, Bruce, Alastair Smith, Randolph Siverson, and James Morrow. 2003. *The Logic of Political Survival*. Cambridge, MA: MIT Press.

Centola, Damon. 2011. "An Experimental Study of Homophily in the Adoption of Health Behavior." *Science* 334: 1269–72.

Evans, Peter B. 1995. *Embedded Autonomy: States and Industrial Transformation*. Princeton University Press.

Fox, Jonathan A. 2015. "Social Accountability: What Does the Evidence Really Say?" *World Development* 72 (C): 346–361.

Gerber, Alan, and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W. W. Norton.

Gerber, Alan, Donald P. Green, and Christopher Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large- Scale Field Experiment." *American Political Science Review* 102 (1): 33–48.

Karlan, Dean, and Jacob Appel. 2011. *More Than Good Intentions: Improving the Ways the World's Poor Borrow, Save, Farm, Learn, and Stay Healthy*. New York: Penguin.

Karlan, Dean S. 2005. "Using Experimental Economics to Measure Social Capital and Predict Financial Decisions." *The American Economic Review* 95 (5): 1688–1699.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, N.J.: Princeton University Press.

Kosack, Stephen, and Archon Fung. 2014. "Does Transparency Improve Governance?" *Annual Review of Political Science*.

Kuehn, David, and Ingo Rohlfing. 2009. "Does It, Really? Measurement Error and Ommitted Variables in Multi-Method Research." *Qualitative and Multi-Method Research* 7 (2): 18–21.

Lieberman, Evan S. 2005. "Nested Analysis as a Mixed-Method Strategy for Comparative Research." *American Political Science Review* 99 (3): 435–52.

Lieberman, Evan S., Daniel N. Posner, and Lily Tsai. 2012. "Does Information Lead to More Active Citizenship?" An Evaluation of the Impact of the Uwezo Initiative in Kenya.

Mahoney, James. 2008. "Toward a Unified Theory of Causality." *Comparative Political Studies* 41 (4/5): 412–36.

———. 2010. "After KKV: The New Methodology of Qualitative Research." *World Politics* 62 (1): 120–47. doi:10.1017/S0043887109990220.

Mill, J.S. 1843. *A System of Logic*. London.

Pritchett, Lant, Salimah Samji, and Jeffrey Hammer. 2017. "It's All About MeE: Using Structured Experiential Learning ('e') to Crawl the Design Space." Center for Global Development Working Paper 322. Washington D.C. Accessed April 9. https://www.cgdev.org/publication/its-all-about-mee-using-structured-experiential-learning-e-crawl-design-space.

Seawright, Jason. 2016. *Multi-Method Social Science: Combining Qualitative and Quantitative Tools*. Cambridge University Press.

Shapiro, Ian. 2016. *Politics Against Domination*. Cambridge, MA: Harvard University Press.

Spencer, L., J. Ritchie, L. Lewis, and L. Dillon. 2003. *Quality in Qualitative Evaluation: A Framework for Assessing Research Evidence*. London: Government Chief Social Researcher's Office, Cabinet Office.

Stern, E., N. Stame, J. Mayne, K. Forss, R. Davies, and B. Befani. 2012. "Broadening the Range of Designs and Methods for Impact Evaluations." Working Paper 38, London: DFID.

Woolcock, Michael. 2013. "Using Case Studies to Explore the External Validity of 'Complex' Development Interventions." *Evaluation* 19 (3): 229–248.