External Validity and Policy Adaptation: From Impact Evaluation to Policy Design

Martin J. Williams

With the growing number of impact evaluations worldwide, the question of how to apply this evidence in policy making processes has arguably become the main challenge for evidence-based policy making. How can policy makers predict whether a policy will have the same impact in their context as it did elsewhere, and how should this influence the policy's design and implementation? This paper suggests that failures of external validity (both in transporting and scaling up policy) can be understood as arising from an interaction between a policy's theory of change and a dimension of the context in which it is being implemented. The paper surveys existing approaches to analyzing external validity, and suggests that there has been more focus on the generalizability of impact evaluation results than on the applicability of evidence to specific contexts. To help fill this gap, the study develops a method of "mechanism mapping" that maps a policy's theory of change against salient contextual assumptions to identify external validity problems and suggest appropriate policy adaptations. In deciding whether and how to adapt a policy, there is a fundamental informational trade-off between the strength of evidence on the policy from other contexts and the policy maker's information about the local context.

JEL codes: A12, B41, D04, O22
Keywords: external validity, impact evaluation, fidelity, adaptation, evidence-based policy.

In 2015, Zimbabwe's government rolled out a new Human Immunodeficiency Virus (HIV) treatment nationwide. The decision was evidence-based: A range of randomized control trials (RCTs) had shown the new treatment to be an improvement over the previous drug cocktail, and the World Health Organization (WHO) recommended that it be used as the standard treatment throughout sub-Saharan Africa. Yet after the new treatment was rolled out, "reports soon followed about people quitting it in

The World Bank Research Observer

doi: 10.1093/wbro/lky010 35:158–191

droves" (Nordling 2017, 20). It turned out that one of the drugs in the treatment, efavirenz, caused significant neuropsychiatric adverse effects (e.g., hallucinations, suicide ideation) in individuals with a particular genetic variant. This variant is rare worldwide, so these adverse effects were not deemed a major problem by international researchers, but it happens to be quite common in Zimbabwe (Masimirembwa, Dandara, and Leutscher 2016). Even though the Zimbabwean government and WHO had based their decisions on extensive and rigorous empirical evidence, the policy decision to include efavirenz had disastrous effects for a significant fraction of patients—an error that could have been avoided, since Zimbabwean scientists had previously identified this genetic variant and its interaction with efavirenz (Nyakutira et al. 2008). Relying on empirical evidence from elsewhere without also utilizing local information had led policy makers in Zimbabwe to make a costly mistake.

Policy makers worldwide face similar challenges in trying to apply evidence to policy decisions. With the recent boom in impact evaluations around the world, policy makers in many sectors now have at their disposal an overwhelming amount of evidence about "what works"—or at least what worked in a particular context. Following the example set by the medical sciences, the premise of evidence-informed policy making is that the design of policy can be based on evidence of what has worked in other contexts, rather than each policy maker having to start from scratch. Yet as impact evaluations have multiplied, it has become apparent that "the same" policy can have very different effects in different populations (Deaton 2010; Pritchett and Sandefur 2015; Vivalt 2017).[1] Similarly, policies shown to be effective in small trials have not always been as effective when implemented at scale, even in the same country (Banerjee et al. 2016a; Bold et al. 2018). This is the problem of the *external validity* of impact evaluations. Although there is widespread agreement on the importance of evidence for informing policy decisions, the limited external validity of impact evaluation evidence poses significant challenges for policy makers: How can one know if a policy will have the same effect in this implementation context as it did elsewhere? And to what extent should policy makers copy the design of policies that have worked elsewhere, rather than use local information to try to adapt them to fit the local context?

This paper begins by proposing a simple and flexible framework for thinking about these questions, and about external validity more broadly. A policy can have a different impact in a new context than it had in a previous context if *part of a policy's theory of change interacts with a difference in contexts*. A policy's theory of change is a set of logical steps spanning inputs to activities, outputs, intermediate outcomes, and final outcomes. Whether this mechanism works as intended depends at each step on the validity of a set of contextual assumptions. While these assumptions may have been true of the context in which a policy had previously been shown to work, whether the policy will have the same effects in a new context depends on whether these same contextual assumptions hold. Since context can include a wide range of factors—

location, target group, implementing organization, scale, time period, the existence of related policy interventions, and so forth—and the theory of change includes factors related to implementation as well as impact, this parsimonious framework encompasses the range of typologies of external validity failures discussed in existing literature (Deaton 2010; Cartwright and Hardie 2014; Muller 2015; Banerjee et al. 2016a) and applies equally to issues of scale-up as well as transporting successful policies.

Within the topic of external validity, one can further distinguish between the *generalizability* of evidence and the *applicability* of evidence. Questions of generalizability are about whether an impact evaluation's findings are likely to hold in general in other contexts (but without a specific destination context in mind), whereas questions of applicability are about whether evaluation results from one or more other contexts will hold in a specific destination context. While the generalizability of evidence is somewhat informative about whether a policy will work in a specific context, the multidimensionality of both policies and contexts means that evidence from elsewhere can never be fully determinative of a policy's impact in a new context as even apparently minor idiosyncratic details can have a major effect on policies' effectiveness. The problem is analogous to the "last mile" problem in public transportation and service delivery: Evidence from rigorous impact evaluations can get policy makers significantly closer to the "correct" policy choice, but there is still a need for structured guidance on how to bridge the gap from "what works in general" to "what will work in my context."

With this distinction in mind, this paper surveys the existing literature on external validity across a range of disciplines. Overall, the academic literature on external validity has focused mainly on questions of generalizability, whereas the question of applicability is the problem with which most policy makers are confronted when trying to use evidence to design policy. This is evident in the two main strands of work on external validity. First, one existing approach is to estimate the average effect of an intervention across different contexts by *aggregating* the results of multiple studies. This is the approach of replication, meta-analysis, and systematic review (Vivalt 2017). A second strand focuses on what evaluators can do to increase the external validity of a particular study, including a range of approaches such as formal theory and structural modeling (Deaton 2010), larger evaluations (Muralidharan and Niehaus 2017), various econometric extrapolation techniques (Angrist and Fernandez-Val 2010; Gechter 2016; Kowalski 2018), and integrating "structured speculation" on external validity into research papers (Banerjee, Chassang, and Snowberg 2016b). While both strands are informative about the potential applicability of evidence to specific contexts, their emphasis on generalizability offers policy makers little structured guidance on how to bridge the gap between evidence from other contexts and the inevitable particularities of specific contexts.

The paper then introduces a method that may be useful for policy makers to bridge the gap between the best evidence from other contexts—from impact evaluations,

meta-analyses, model-based extrapolations, and so forth—and the effective application of this evidence to policy design and adaptation in their own context. This method of *mechanism mapping* builds explicitly on the paper's proposed understanding of external validity as the interaction of mechanism with context by juxtaposing (1) the policy's theory of change with (2) the underlying contextual assumptions needed for each step of this mechanism to operate, and (3) comparing these assumptions to the actual characteristics of the policy maker's context. If a necessary assumption does not hold in the new context in the same way as it held in the old context, then the mechanism will be interrupted and the policy's impact will differ. The mechanism mapping process can also be applied to questions of policy scale-up, since implementing a policy at scale involves different contextual assumptions (e.g., implementation quality, resource requirements, general equilibrium effects, political economy) than a small pilot, even if the pilot was undertaken in the same geographical location.

Evidence—both local and from other contexts—plays a crucial role within the mechanism mapping process. Undertaking mechanism mapping would ideally consist of a systematic process of seeking empirical evidence to support contextual assumptions and understand actual contextual realities through descriptive statistics, qualitative data, and evidence from relevant impact evaluations. At the most rigorous extreme, one could undertake a series of "mechanism experiments" (Ludwig, Kling, and Mullainathan 2011) to validate each step of the theory of change and its underlying contextual assumptions. Where time or resource constraints make a more thorough process infeasible but a decision must nevertheless be made, even conducting a relatively brief mechanism mapping may help policy makers structure their judgment and avoid sole reliance on intuitions or prejudices. As Ravallion (2009) notes, learning from impact evaluations typically requires both theory and information from outside the evaluation; this simple and intuitive diagnostic process gives policy makers a flexible framework for marshalling all available empirical evidence from different sources and of different levels of rigor in a structured way in support of policy decisions. Whereas the lack of quantitative data has often hindered evidence-based policy making in data-poor contexts, mechanism mapping's ability to integrate less formal types of evidence makes it particularly well suited to such contexts—although of course the weaker the evidence, the lower the quality of the decision is likely to be. The third section discusses the application of mechanism mapping in more detail.

The process of mechanism mapping may also help identify appropriate policy adaptations, by highlighting specific aspects of the policy that are likely to work less well (or potentially better) than in the policy's original context. Policy adaptations thus flow directly from a diagnosis of the relationship between the policy context and the policy's theory of change, so that adaptations are based on a combination of local, context-specific information and evaluation evidence from other contexts. While this combination is a productive way to generate ideas for adaptation, it also

suggests a fundamental trade-off. Evaluation evidence on a policy's effectiveness in other contexts is likely to be more rigorous than available local information, but relying on this evidence from elsewhere requires strict fidelity to the original policy design. On the other hand, using mechanism mapping to identify potential adaptations makes efficient use of local information, but making these adaptations decreases the relevance of evaluation evidence from elsewhere. The optimal level of adaptation in each case will depend on the case-by-case balance between (1) the strength and relevance of evaluation evidence on the policy from other contexts and (2) the policy maker's information about the local context. This optimal level will thus vary not only by policy area and country, but also by the information set of the policy maker and the nature of the policy making process.

Of course, a key limitation of mechanism mapping as a tool is that it relies to an extent on the judgment of policy makers, as with any other process of policy design or adaptation. While this may provide opportunities for policy makers' biases or preferences to influence policy, numerous such opportunities already exist in policy design and implementation, and the structured process of comparing the theory of change, contextual assumptions, and contextual realities can arguably reduce these by making them explicit and structuring deliberation. Another limitation of mechanism mapping is that it yields only directional predictions of policy effectiveness rather than precise statistical point estimates of effect sizes and confidence intervals. In some cases mechanism mapping may not generate an unambiguous overall prediction for whether a policy is likely to be more or less effective than it had been in a previous context, since multiple contextual differences may shift effectiveness in different directions (i.e., be opposite-signed). Nonetheless, for many applied policy purposes—in particular for identifying aspects of a policy that may benefit from adaptation—a directional prediction may suffice, and the use of mechanism mapping does not preclude policy makers from also making use of more precise quantitative tools. A practical approach for policy makers would be to use evidence from impact evaluations and meta-analyses on what policies or interventions are likely to be most effective as a starting point from which to begin the process of mechanism mapping, marshalling context-specific evidence, and adapting policy. Finally, mechanism mapping as a method for analyzing external validity and adaptation is prone to all the same mental and political biases as any other decision process, and depends on the good faith, good judgment, and critical analysis skills of the policy maker(s) undertaking the process. A structured decision-making tool is a complement to, not a substitute for, these traits.

In its emphasis on understanding mechanism-context interactions, this paper is most similar to recent work in public health (Moore et al. 2015; Leviton 2017), economics (Bates and Glennerster 2017), philosophy (Cartwright and Hardie 2014), and public management (Barzelay 2007), and to "realist" approaches to evaluation in sociology (Pawson and Tilley 1997). The contribution of this paper is to (1) present

a flexible and parsimonious conceptual approach to understanding external validity, (2) survey the existing literature's strengths and limitations in helping policy makers analyze external validity issues in their own contexts, and (3) link this to a simple, practical, and intuitive framework for identifying likely external validity failures which (4) feeds directly into the policy adaptation process. Finally, mechanism mapping is related to adaptive policy making (Pritchett, Samji, and Hammer 2013; World Bank 2015) in emphasizing the use of local information to improve policy design; the fourth section discusses complementarities between the two approaches.

In order to focus on issues of external validity and policy transportation arising from real differences in context, this paper abstracts from the issues of the statistical or methodological accuracy of published impact evaluations that have been the focus of much of the literature on replication in the social sciences (Christensen and Miguel 2016). While these issues can also lead to differences in estimated policy impacts across contexts, they have been discussed extensively elsewhere and are conceptually distinct. For brevity, this therefore discusess impact evaluations as if they represent true causal estimates of the policy's impact in that context, even though policy makers should obviously interpret published findings through a critical lens.

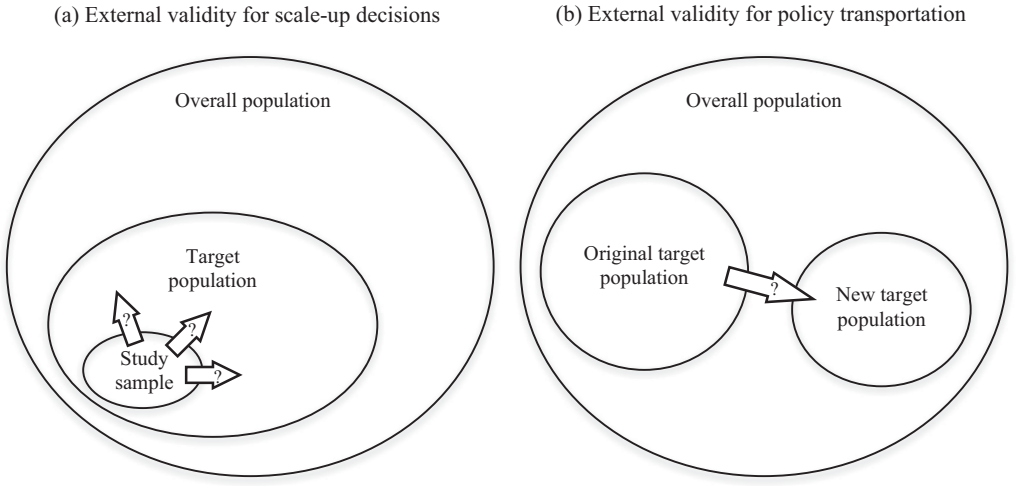
The remainder of this paper proceeds as follows. The next section defines external validity and elucidates the understanding of external validity failures as an interaction of context and theory of change. The following section discusses the limitations of existing approaches to external validity, largely arising from the high dimensionality of policies and contexts. The third section describes the process of mechanism mapping in more detail and gives examples and practical recommendations. The fourth section discusses adaptation and the fundamental informational trade-off, and a final section concludes with a discussion of implications for theory and research.

Understanding External Validity

Defining External Validity

An impact evaluation's *external validity* refers to the extrapolation of its findings beyond the study sample to another population. This contrasts with the *internal validity* of a study, which is established by the identification of a causal effect via comparison with a valid counterfactual. While academics may be concerned about establishing the extent to which a study has external validity in general—across all other hypothetical contexts—the policy maker's problem is whether the findings of a study conducted elsewhere would continue hold *in one specific context*. In Cartwright and Hardie's (2014) framing, an impact evaluation answers the question "did it work there?"; while policy makers are interested in the question "will it work here?" This section distinguishes between two different types of external validity: scaling up a policy within the same target population, and transporting a policy to a different target population.

Figure 1. Two Types of External Validity



Source: Author's elaboration.

First, one might be interested in extrapolating the findings from the *study sample* (the individuals or units who actually participated in the evaluation) up to the broader *target population* for whom the study's results are intended to be applicable. Will a policy have the same impact on the full target population as it did on a smaller pilot or trial group? This is the case of policy scale-up, as represented by panel (a) of fig. 1, and comprises two distinct aspects. One aspect of this can be achieved by having a study sample that is statistically representative of the study population (usually through random sampling). However, in some cases this is not possible. For example, many pharmaceutical trials are conducted on healthy individuals even though these are systematically different from the target population on important dimensions. Similarly, Henrich, Heine, and Norenzayan (2010) point out that most psychological studies are conducted on study samples that are Western, Educated, Industrialized, Rich, and Democratic (WEIRD)—often North American or European undergraduates—that are among the least representative sections of humanity.

In addition to these concerns about the representativeness of the study sample when scaling up a policy to the full target population, the second aspect of scale-up concerns the implementation and impact of the policy itself. Implementing at scale may mean implementing through a government bureaucracy that also implements many other policies rather than a small and closely supervised nongovernmental organization or academic research team, which could undermine effectiveness (Bold et al. 2018; Cameron and Shah 2017). Treating a higher fraction of the target population could also lead to higher or lower effects through spillovers (e.g., Miguel and Kremer 2004).

Second, one might care about whether a policy or intervention would have the same effect in a *different target population* than the original target population. This is the meaning of external validity with which policy transportation is usually concerned. Panel (b) of fig. 1 illustrates this meaning of external validity. External validity in this sense concerns the similarity of the two target populations on key covariates, both observable and unobservable.

Despite their distinctions, both types of external validity can be thought of as specific cases of the same underlying challenge: how to predict whether a policy will have the same effect(s) in a new implementation context as it did in a previous context. As discussed further below, the underlying factors that drive external validity failures of both types can be understood with the same framework, and the mechanism mapping approach to diagnosing them is equally applicable to both.

Failures of External Validity

Why might a policy be effective in one context but fail in another, or vice versa? This paper proposes a framework that builds on two key concepts. First, a policy is defined by its theory of change (also referred to as its mechanism, results chain, or logic model). This is a mapping of the policy's intended mechanism—how it is supposed to work. This begins with the specification of the intended *final outcomes* or ultimate goals of a policy. In order to achieve these final outcomes, a series of *intermediate outcomes* must occur, and a policy specifies *outputs* that the government will deliver in order to trigger these intermediate outcomes. To deliver these outputs, government plans to undertake a set of *activities*, which require certain *inputs* (e.g., financial or human resources, information).[2] The steps from the provision of inputs to the delivery of outputs comprise the *implementation* of the policy, while the link from these outputs to the policy's final outcomes via intermediate outcomes represents the *impact* of these outputs on society, or—in Bates and Glennerster's (2017) terminology—the behavioral response to the intervention.

Second, all policies are implemented in a particular *context*, and the characteristics of this context may affect a policy's effectiveness. Context here refers not just to location, but also to the full range of population and other variables that could affect the policy's implementation and impact. While the range of potentially relevant characteristics is effectively limitless, some particularly salient dimensions of context include:

- Location, polity, or society in which the policy is being implemented (e.g., Iceland or India), together with all the social, cultural, economic, geographic, and political characteristics that vary across locations;
- Target groups (e.g., working adults, single mothers, at-risk teens);
- Time the policy is being implemented, whether the year (e.g., 1965 vs. 2015), season, or duration since the policy began;

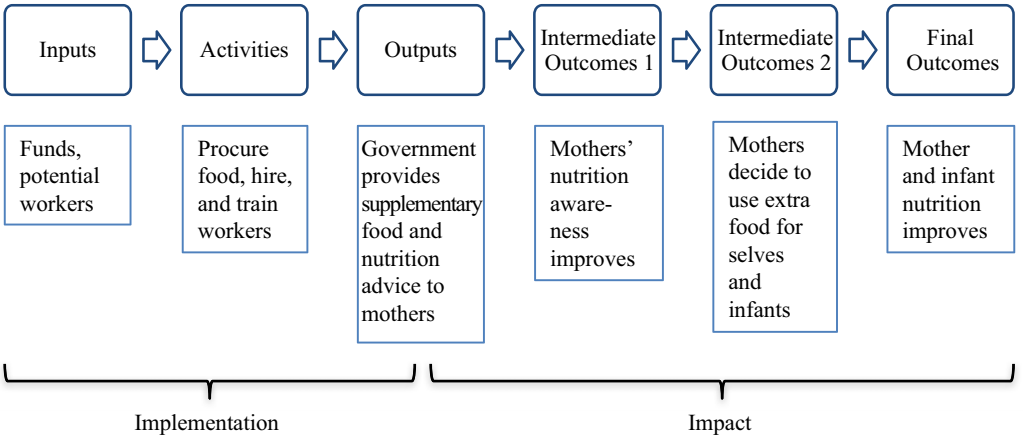
- Existence of related policy interventions, including spillovers from implementation of the policy in neighboring areas as well as availability of public services or infrastructure; and
- Implementing organization, including its competence, level of resources, and political constraints.

Combining these two concepts makes it clear how failures of external validity emerge. A policy's theory of change relies on each step actually occurring and leading to the next step as intended, both in terms of implementation and impact. Will the correct level of inputs be made available as intended? Will the activities needed to create outputs actually occur with the requisite quality and sequence? Will society react to these outputs as hypothesized? The answer to all of these questions may have been affirmative in the context in which a successful impact evaluation was undertaken, but transporting the policy to a new context requires making assumptions about the answers to these questions in the new context. The implementation and impact of a policy are thus a function of the combination of a policy's theory of change with the context in which it is being implemented. If it is then observed that a policy had one impact in one context but had a different impact in a different context, then it must be the case that the differences in context undermined one or more critical links in the policy's theory of change. Identifying these *interactions* of context and theory of change is critical to understanding external validity.

Two examples illustrate the diverse ways in which contextual differences can undermine policy impact. The first example comes from Cartwright and Hardie's (2014, 80–84) comparison of two World Bank–funded programmes: the Tamil Nadu Integrated Nutrition Programme (TINP), and the Bangladesh Integrated Nutrition Programme (BINP). The TINP project was implemented in the 1990s and sought to improve child nutrition in rural Tamil Nadu by simultaneously delivering two interventions: supplementary food for pregnant or nursing mothers and their children, and nutritional advice to mothers to correct a misperception that mothers should reduce rather than increase their food intake during pregnancy. A rigorous impact evaluation showed that the project was successful: mothers' nutritional knowledge improved, mothers and children consumed more food, and childrens' malnutrition and stunting decreased significantly. Figure 2 maps a simple version of this theory of change.

Following this evaluation, the program was copied and transported to rural Bangladesh, where the same problems existed. Yet under BINP, while mothers' nutritional knowledge improved, there was no impact on malnutrition. This was due to a key contextual difference: Whereas mothers were typically responsible for shopping and household food allocation decisions in rural Tamil Nadu, in rural Bangladesh men usually conducted the shopping and their mothers (the mothers-in-law of the pregnant or nursing women) controlled household food allocations (White 2005;

Figure 2. Theory of Change: Bangladesh Integrated Nutrition Programme



Source: Author's elaboration based on Save the Children (2003), White (2005), World Bank (2005a, b), Cartwright and Hardie (2014).

Cartwright and Hardie 2014). This difference in contexts *interacted* with a key link in the theory of change: the hypothesis that greater nutritional knowledge in mothers would lead them to decide to allocate the supplemental food to themselves and their children, rather than distributing it to other members of the household.

A second example comes from the Tools of the Mind early childhood education program, which aimed to improve executive function (e.g., resisting temptation, working memory). After a small but widely publicized RCT in New Jersey showed strong positive impacts (Diamond et al. 2007), a federally funded scale-up in other states actually found *negative* impacts relative to a control group. Evaluators explained that correctly implementing the curriculum—"the most complex we have ever seen"—required two years of training, ongoing in-classroom teacher coaching, and carefully sequenced implementation of the 60 activities that comprised the program (Farran and Wilson 2014, 21). Although teachers actually implemented the formal components of the program with relatively high fidelity (as measured by the number of activities implemented), the closely specified structure of the program did not fit well into the school day which—unlike the carefully controlled original RCT—also included many other nonprogram activities and demands on teachers' attention. While children undertook many of the structured parts of the Tools of the Mind curriculum, there was little time for them to undertake the kind of free play that would have allowed them to internalize the skills taught in the structured parts of the program (Farran and Wilson 2014). In this case the interaction between context and theory of change that undermined program effectiveness was quite subtle: Implementing the program in a "real-world" setting necessitated the compression of a program component that seemed unimportant but turned out to be crucial.

This framework for understanding external validity also makes it clear that contextual differences do not affect policy impact unless they interact with the policy's theory of change. Although contexts are characterized by an almost infinite number of dimensions and are thus all unique, this does not imply that all policies must be designed with a particular context in mind, since most of these contextual differences are irrelevant to the policy's mechanisms. In practice, of course, it can be difficult to identify salient contextual differences and judge their relevance—the specific interactions that undermined both BINP and the Tools of the Mind scale-up may not have been obvious *ex ante*. The third section below presents a structured approach to helping policy makers identify which dimensions of context are likely to affect a policy's impact.

Finally, although the examples presented here have been of negative interactions with contextual differences, these interactions could just as well be positive—leading a policy that was not effective in its original context to be effective in a new context. As the penultimate section discusses later, policy adaptations can aim not just to mitigate threats from transportation to a new context but also to improve their effectiveness.

Existing Approaches to External Validity

With this framework in mind, this section surveys a range of existing approaches to dealing with external validity. On the whole, it finds that the academic literature on external validity provides increasingly insightful answers to questions of the *generalizability* of impact evaluations or bodies of evidence—whether evaluation results from a specific context will hold in unspecified other contexts. However, it provides more limited insight into concerns about *applicability* of evidence—whether evaluation results from various other contexts will hold in the specific context in which a policy maker is working. Although these existing approaches can be very powerful and vary tremendously, from empirical to theoretical and formal to informal, their common limitation is their inability to analyze the heterogeneity of policies' impacts across more than a handful of dimensions. This contrasts with the high dimensionality, both of policies and of contexts (Pritchett 2017), and limits the overall usefulness of these approaches to policy makers grounded in specific contexts.

One empirically driven response to the variability of policy impacts across contexts is to *aggregate* numerous studies of the same policy. In its simplest form, this could be a simple replication in another context. As the policy is tried and evaluated in more contexts, it may become possible to aggregate these results further, through a systematic review or a meta-analysis. This empirically driven approach is perhaps most associated with the evidence-based policy movement, drawing its inspiration largely from medicine. Aggregation in this way can yield an average treatment effect across study samples (and if the samples are representative of their target population, across these populations) in which the policy has been studied. But this estimate is of an

average treatment effect in the *average* context in which the policy has been evaluated, which can differ from the policy's effect in a specific new context in two ways.[3]

First, the populations in which the policy has previously been tried and/or evaluated may differ systematically from the new context in important ways. For many social policy interventions, for example, there exist numerous studies from OECD countries but little or no evidence in developing countries, and Allcott (2015) has shown that policy experiments are often conducted first in the most favorable locations, leading to a site selection bias effect. Policy makers applying this evidence to their own contexts must therefore ask, "Is my context average?" Since contexts have many dimensions, all contexts are unique in some ways, and it is unclear how many and which of these dimensions of a context must be "average" in order for this average treatment effect to pertain. This is not to say that systematic reviews are uninformative: Under a normal distribution one would expect most contexts to be closer to the average than the extremes across most dimensions, and so absent any further information about the new context, an average treatment effect estimated from other contexts would be the best predictor of a policy's impact. But while this makes systematic reviews a useful starting point for policy makers, naïvely adopting a policy that has a positive "headline" average treatment effect in a systematic review is likely to backfire in many contexts.[4]

Second, there can be significant heterogeneity in policy impact across contexts, so that a policy that has a positive effect on average could have a negative effect in some contexts. The main empirical approach to dealing with heterogeneous effects is to employ subgroup analysis, which breaks down average treatment effects across important variables: age, gender, income, region, implementing authority, and conceivably any other observable variable on which data exists. Conducting subgroup analysis, either within a single study or in a meta-analysis, allows evaluators to answer the more nuanced question "what works *for whom*?" This allows policy makers to compare their contexts to others on these covariates, and provides some guidance about which dimensions of context might matter for a given policy.

While this information is useful, subgroup analysis is inherently limited in the number of variables along which it can disaggregate results. Individual studies are limited in the number of variables they can measure and collect, and subgroup analysis in meta-analysis is even further restricted by the limited set of variables that are common to all (or at least several) studies. Inevitably, there will be some contextual variables that mediate a policy's effectiveness—who controls household food allocations, the fit of a curriculum within the existing school day, the prevalence of particular genetic variants—that are either difficult to measure or that evaluators might not think to measure *ex ante*, and are thus unobserved. Even where a given covariate is present across studies, one might question whether this variable interacts with the policy in the same way across contexts. For instance, low income might undermine the effectiveness of a skill-upgrading intervention in rich countries because

individuals do not have time to attend the classes (e.g., if they are working multiple low-wage jobs, or cannot arrange childcare), but in a poor country income may not be correlated with time poverty in the same way, so the intervention might be more effective. While aggregation of evidence across multiple studies (and use of subgroup analysis for disaggregation within this) is indeed informative about the predicted success of an intervention in any given context—as Vivalt (2017) shows empirically—even for the most highly researched interventions there nonetheless remain numerous potential mechanism-context interactions that existing data cannot fully predict.

A second approach focuses not on aggregating evidence across more contexts, but on making out-of-sample extrapolations through structural modelling (Deaton and Cartwright 2016) or other empirical methods that can, in some circumstances, be used to extrapolate results from one study to other populations, by exploiting specific forms of selection and noncompliance within RCTs or by adjusting estimated impacts based on heterogeneity over observed covariates (Angrist and Fernandez-Val 2010; Gechter 2016; Kowalski 2018; Andrews and Oster 2018). Such theoretical or empirical extrapolation methods take advantage of variation within the study sample in order to better understand the underlying causal processes and extrapolate estimates from the actual study sample to other populations. These methods help researchers and policy makers further improve the informativeness of the existing evidence about the predicted impact of the intervention in a new context, but are also inherently limited in the number of variables and types of scenarios across which they can extrapolate. As Low and Meghir (2017, 34) write: "Structural economic models cannot possibly capture every aspect of reality…There will always be some economic choices left out of any particular model…". Yet as the examples of BINP, Tools of the Mind, and Zimbabwe's efavirenz rollout illustrate, the range of contextual factors that can influence policy impact is immense. While structural modeling can therefore provide powerful insights about the effect of *some* important contextual factors, even the best-judged model will only be able to incorporate a small handful of the numerous variables that policy makers must consider in policy design. As with the aggregation approach discussed above, this is certainly informative for policy makers in specific contexts, and is an improvement over simply having the results of an impact evaluation from another context without such extrapolation, but still falls short of taking into account all the potential mechanism-context interactions with which policy makers must concern themselves.

Similarly, the design of policy experiments can vary aspects of the policy that are important for understanding external validity, such as whether it is implemented by an NGO or government (Bold et al. 2018; Cameron and Shah 2017; Angrist 2017). Again, the limitation is that trials can only feasibly vary one or two dimensions of a policy without losing statistical power, while the number of dimensions of policy and context that could matter—combined with their interactions—is effectively infinite. Likewise, larger experiments would certainly improve external validity

(Muralidharan and Niehaus 2017), but applying the results will always require the consideration of contextual differences for any trial on a scale that is less than global. The proposal of Banerjee, Chassang, and Snowberg (2016b) for the inclusion of "structured speculation" on external validity in reports of impact evaluation results is perhaps the closest in spirit to the mechanism mapping approach developed in the next section of this paper, but is fundamentally a tool for evaluators, not policy designers, since such speculation is necessarily undertaken without a specific target context in mind.

The "realist evaluation" approach pioneered in sociology and social policy (Pawson and Tilley 1997) shares with the recent external validity literature in economics an emphasis on mechanisms and heterogeneity rather than simply establishing average treatment effects. In asking "why a program works for whom and in what circumstances" and seeing the objective of evaluation as the elaboration of Context-Mechanism-Outcome configurations (CMOCs), realist evaluation is also related to this paper's mechanism mapping approach, albeit from the perspective of the evaluator rather than the policy maker. However, the focus of realist evaluation is typically on how best to evaluate a program rather than how to use existing evidence to design a policy. This difference in target audiences has perhaps contributed to realist evaluation being perceived as deeply philosophical as well as unwieldy and time-intensive in practice (Marchal et al. 2012).

Finally, debates around external validity are perhaps most advanced in public health, where discussions of the interaction between mechanism and context have become central to thinking about the transportation and scaling of trial results (Moore et al. 2015; Leviton 2017), the complexity of interventions is widely acknowledged and is beginning to be explored empirically (Hawe 2015), systematic reviews routinely take realist approaches to unpacking mechanisms and heterogeneity (Greenhalgh et al. 2016), and a strong institutional architecture is seeking to establish reporting conventions and other steps to embed these new approaches in research (e.g., Wong et al. 2016).

Mechanism Mapping

Basic Process and Example

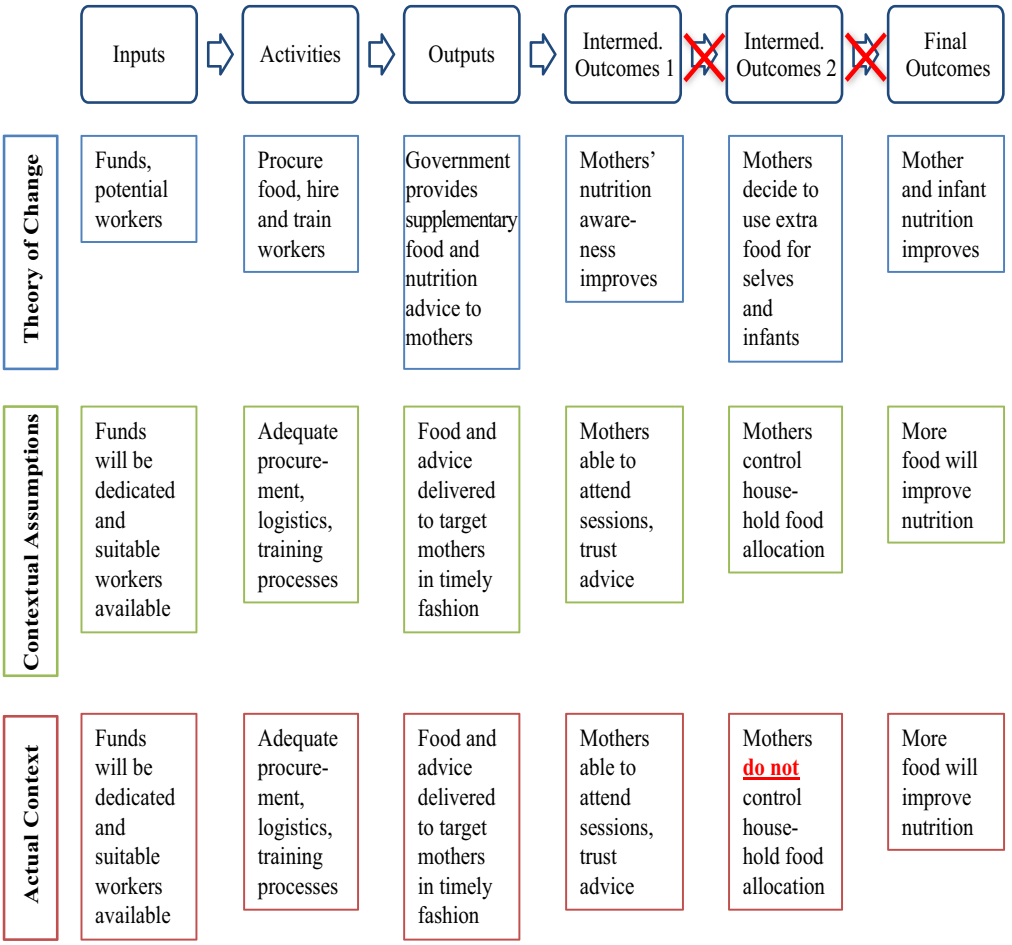
Although these techniques for assessing and improving the external validity of impact evaluation evidence are important to help guide policy makers towards policies and interventions that are more likely to be successful, evidence-based policy makers are still faced with the challenging task of diagnosing potential mechanism-context interactions that could influence the policy's effectiveness in their context. If external validity failures arise from interactions between a policy's theory of change and its context, then it follows that diagnosing such failures requires a way to

examine a policy's theory of change alongside its context. Furthermore, if contexts have numerous dimensions, for many of which hard data may not be available, then a useful framework also needs to be able to integrate high-quality, rigorous evidence ("observables") as well as softer, potentially tacit or local information ("unobservables"). This section introduces a *mechanism mapping* approach that fulfills both criteria. The paper introduces this method by first presenting the approach itself, then giving a brief example. It then discusses some practical and conceptual issues, including the role of empirical evidence, then suggests some ways in which this tool can be integrated into processes of policy making and evaluation in order to complement other approaches to evidence-based policy.

The first step of mechanism mapping is to lay out the *theory of change*, or mechanism, through which the policy had its measured impact in the previous context. As discussed above, this can be thought of as a causal chain leading from a policy's initial inputs to its intended final outcomes, via activities, outputs, and intermediate outcomes. The second step is to lay out the most important or salient *contextual assumptions* underpinning each step of this chain. These are the characteristics of the context that are required for the policy to actually function as the theory of change intends. If the policy in question had been shown to be successful in another context, then presumably these assumptions would have been valid in that context. The third and final step is to lay out the corresponding *actual contextual characteristics* for each step of the chain, highlighting any differences between actual contextual characteristics and the contextual assumptions necessary for the policy to function as intended. These differences in context—whether negative or positive—are what policy makers can use to predict whether the policy will have a similar, smaller, or larger impact on the final outcomes than it did in its previous context, as well as pinpointing the stage at which the theory of change is likely to be interrupted (and thus which aspects of the policy may need to be adapted, as the following section discusses later).

To illustrate the approach, consider BINP discussed above, which had an identical design to the World Bank's earlier TINP project in Tamil Nadu.[5] The intended final outcome of BINP was to improve mother and infant nutrition. To do so, government was to provide two main outputs: nutritional advice delivered to pregnant and nursing mothers, and the distribution of supplementary food to mothers to take home. These outputs would lead to the final outcome via two sets of intermediate outcomes: First, mothers' nutritional awareness would improve, alongside their receipt of the supplemental food; and second, mothers would then decide to use the supplemental food for themselves and their infants (as opposed to giving it to other family members, i.e., program "leakage"). In order to produce these outputs, the government required inputs of adequate financial resources to purchase the food and pay personnel, as well as a logistical system and potential pool of extension workers to deliver the food and nutritional advice. Key

Figure 3. Theory of Change: Bangladesh Integrated Nutrition Programme



Source: Author's elaboration based on Save the Children (2003), White (2005), World Bank (2005a, b), Cartwright and Hardie (2014).

activities for transforming inputs into outputs could include procuring the food, hiring and training workers, and conducting outreach to eligible mothers.

The contextual assumptions required for this theory of change to work are listed in the second row of fig. 3. Sound implementation requires that government do the following: dedicate adequate financial and human resources to the project; procure and distribute food and hire workers effectively, including quality assurance as well as prevention of excessive corruption, and train workers adequately; and deliver these outputs to a pool of eligible mothers predictably and in a timely fashion. Impact then requires that mothers be able to attend the sessions and trust the advice they are being

given; that mothers actually control household food allocation; and that the supplementary food, if consumed, actually lead to the desired improvement in nutrition. In the Tamil Nadu context, these assumptions were presumably valid—hence the impact evaluation finding that TINP significantly improved mother and infant nutrition (World Bank 2005b).

The third row of fig. 3 contrasts these contextual assumptions to the actual contextual characteristics of the new context, in this case rural Bangladesh. Recall that BINP succeeded in distributing food and nutritional advice to the mothers, and that mothers' nutritional awareness did actually improve as a result, but that the program failed to improve mother and infant nutrition because most of the supplementary food went to other family members. The key contextual assumption that did *not* hold in Bangladesh was that mothers controlled household food allocation, and would thus be able to act on their improved nutritional awareness. This broke the link between Intermediate Outcome 1 and Intermediate Outcome 2; since Intermediate Outcome 2 was not achieved, neither was the Final Outcome. If the designers of BINP had carried out a mechanism mapping when transporting the successful TINP program to Bangladesh, perhaps they would have uncovered this crucial but implicit assumption.

Practical and Conceptual Issues

Each of these three steps—the theory of change, contextual assumptions, and actual characteristics—are associated with important conceptual and practical challenges. The first is how to decompose a policy or program into a theory of change. In many cases, policy makers in new contexts can establish this based on the original impact evaluation or meta-analysis on which they are basing the transportation or scale-up of the policy (although many published impact evaluations do not fully specify their intended theory of change or evidence supporting this mechanism, leaving it partially to the policy maker in the new context to specify this causal chain). While this can be challenging, theories of change (and related tools such as logic models, results chains, and logframes) are an intuitive and commonly taught tool of policy making in many evaluation textbooks (Gertler et al. 2016), government manuals (HM Treasury 2011), and donor guidelines (DFID 2012).

Of course, some policies are more complicated than others. For the sake of clarity, mechanism mapping is illustrated here using the type of simple, linear theory of change that adequately characterizes many impact evaluations. Theories of change can, of course, be much more intricate, for example by mapping out multiple components of a multifaceted program. Appendix A1 presents an example of a policy whose theory of change contains 55 distinct steps across multiple interlocking components. Since each of these 55 steps is numbered and has its precursor steps clearly identified, the same basic procedure of juxtaposing the contextual assumptions against actual contextual characteristics for each step can be applied

(albeit perhaps in the form of a table rather than underneath each step in the theory of change, for reasons of visual clarity).

Similarly, mechanism mapping can be adapted to policies that are intended to lead to *multiple final outcomes* (e.g,. a cash transfer that is intended to increase consumption and improve child school attendance); to examine the likelihood of *negative outcomes* or side effects of the policy (Bonell et al. 2014); and to shed light on questions of *multiple competing mechanisms*. This involves laying out multiple parallel mechanism maps representing different potential pathways to the positive or negative outcomes, then analyzing each to identify which seem most plausible and which key contextual assumptions or characteristics seem most important. Appendix A2 presents an example theory of change from a Botswanan NGO that mapped out parallel theories of change against sets of contextual assumptions in order to examine competing mechanisms and scenarios leading to both positive and negative potential outcomes as part of the intervention adaptation process prior to transporting an intervention from Kenya to Botswana. Finally, some policies are not just complicated but also complex in that their implementation involves a significant degree of uncertainty, simultaneity, and/or feedback. While complex policies are inherently more difficult to create theories of change for (as well as to conduct impact evaluations on), Davies (2004), Rogers (2008), and De Silva et al. (2014) provide examples and guidance for mapping theories of change for complex interventions, which all share the common feature of breaking interventions down into a set of logical steps and are thus amenable to mechanism mapping.

A second conceptual challenge confronting policy makers is how to identify which are the most salient contextual assumptions to consider, since the high dimensionality of context makes it unfeasible to consider all such assumptions. Although this is ultimately a matter of judgment, four practical guidelines suggest themselves:

1. Results from subgroup analysis of impact evaluations or meta-analysis may give insights into common determinants of program effectiveness and thus shed light onto key assumptions. For example, a meta-analysis of food supplementation programs showed that programs where supplements were delivered at feeding centers found lower average leakage of food to other family members (15 percent) than when food was delivered at home (64 percent), suggesting that assumptions around intrahousehold food allocation are key to the success of such policies (Kristjansson et al. 2015).
2. Many dimensions of context are frequently salient and should be taken into consideration for almost any policy: demographic and socioeconomic characteristics of the target population; resource availability; political support and resistance; social and cultural norms; the effectiveness of implementing organizations; potential for corruption or resource diversion; geographic accessibility and other logistical issues; and so on. The UK government Magenta Book (HM Treasury 2011) and

Ravallion (2009) provide lists of aspects of context for policy makers to consider when transporting a program.

3. Important contextual factors specific to a particular policy are often implied by the policy's theory of change (Moore et al. 2015). For instance, laying out BINP's theory of change makes it clear that decision making over household food allocation is a key contextual assumption.
4. Participatory policy making processes, where input from affected stakeholders such as impacted populations (Parker et al. 2008) or implementing staff (Leviton and Trujillo 2017) is systematically sought during the policy design process, are especially well suited to identifying salient contextual assumptions, since directly affected or involved individuals are likely to be able to more accurately envision the practicalities of the policy's impact and implementation.

A third practical challenge—establishing actual contextual characteristics to compare to these assumptions—is another aspect of mechanism mapping where empirical evidence is crucial. In addition to compiling evidence on impact from existing impact evaluations or meta-analyses, policy makers can also gather new data by (to return to the BINP example) examining budget data and political context to shed light on resource availability, investigating the performance of the implementing agency's procurement processes, conducting a survey of eligible mothers' trust of the state and baseline level of nutritional knowledge, undertaking (or reading existing) qualitative research on household food allocation decisions in rural Bangladesh, and discussing with public health experts the prevalence of diseases that might inhibit infants from absorbing nutrients properly. Bates and Glennerster (2017) present several excellent examples of using simple descriptive data, some gathered in just two weeks, to validate contextual assumptions. Impact evaluation results from other contexts, and systematic reviews can enter into mechanism mapping via the contextual assumptions row, as policy makers can use the results of that evaluation to document the extent to which the contextual assumptions held in that context, and possibly even to investigate how variation in these contextual factors was related to the policy's effectiveness. When the mechanism mapping is being conducted for a scale-up of a policy that has already been trialed on a small scale in the same location, the mechanism mapper may even have quite detailed evidence on these issues, and so the search for new empirical evidence can focus on the aspects of context that are changing with the larger-scale implementation: the effectiveness of the implementing agency, general equilibrium or spillover effects, political economy issues, and so forth. One could even imagine policy makers conducting quick and cheap "mechanism experiments" (Ludwig, Kling, and Mullainathan 2011) to validate each step of the theory of change prior to beginning full-scale implementation, or (less ambitiously) follow Rigterink and Schomerus (2016) in compiling evidence from existing evaluations on the validity of some aspects of the policy's theory of change in cases

where no impact evaluation of the full theory of change has yet been conducted. In practice, of course, the available evidence on each of the contextual assumptions, and hence each step of the theory of change, is likely to vary in terms of rigor and depth. Mechanism mapping thus provides an integrative framework for policy makers to apply all available evidence—from RCTs to administrative data to qualitative research to expert judgment—to their policy decisions, and the same basic process can be scaled up or down in terms of detail in order to fit within policy makers' time, resource, and information constraints.

As a procedural matter, a practical way for policy makers to conduct mechanism mapping is in a *nested* manner. The analyst begins by identifying only the most salient steps in the policy's theory of change along with accompanying contextual assumptions and characteristics, following the guidance above. This presents a top-level picture of the overall fit of the policy's required assumptions with the context's actual characteristics. At this stage, it is likely that some steps of the theory of change will have a better fit than others (as in fig. 3). From this top-level view, each of these links in the mechanism can then be broken down and analyzed in more detail. Where the contextual assumptions seemed to fit well at the overall stage—for instance, the activities or outputs steps of fig. 3—breaking down the mechanism serves as a further plausibility check. For instance, an *ex ante* mechanism map of BINP could have thought in more detail about the steps involved in procuring and distributing food, in hiring and training workers, and in coordinating these two program elements, and what resources and bureaucratic processes and skills would be required to execute them. Where the contextual assumptions do *not* seem to fit well at the top-level stage—for instance in fig. 3's intermediate outcome of mothers and infants consuming the extra food themselves—going into more detail can help the analyst identify the root cause of the disjuncture. In the BINP case, this was that TINP's synergy between advice and food distribution to mothers would not apply in Bangladesh. Being more precise in pinpointing the problem simplifies the problem of adaptation discussed in the next section. Continuing this nested approach to mechanism mapping even further in detail could be especially useful for bureaucratic planning processes, by linking a program's theory of change to a detailed set of tasks to be performed and coordinated.

While mechanism mapping is intended primarily as a tool for policy makers to use prospectively to predict the impact of transporting or scaling up a policy that has been successful elsewhere, mechanism mapping is also of potential value to evaluators and to policy makers designing policy from scratch. First, it can be useful in the retrospective evaluation of policies by helping evaluators to show clearly the intended and actual mechanism(s) through which a policy had its impact (or nonimpact). Showing intended versus actual mechanism maps in this way can help evaluators clarify their own thinking and also make the evaluation more informative to readers and policy makers from other contexts. Second, prospective mechanism mapping (e.g., during

trial design or in a preanalysis plan, as in the example discussed in Appendix A2) can help evaluators ensure that they collect the data necessary to assess each of the contextual assumptions *ex post*, along with potential undesirable outcomes and the alternative mechanisms that might bring them about. This process can even be useful in cases where the policy being trialled is completely new, rather than transported from elsewhere or scaled up. Third, mechanism mapping can help evaluators make null results more informative, by specifying which aspects of the policy (if any) worked as intended, where the causal chain broke down, and what future trials should (or should not) considering adapting.

Of course, an important limitation of mechanism mapping for evaluation purposes is that mechanism mapping is only intended to yield directional predictions about overall policy impacts, unlike statistical methods that can yield point estimates and other more precise information. However, directional predictions are still useful for many purposes—in particular for optimizing policy design. Finally, the process of being explicit about a policy's theory of change and the fit between its contextual assumptions and actual characteristics is also important for policy makers who are designing policies from scratch, without the aid of a successful trial from elsewhere. While the lack of prior evidence makes this inherently more uncertain, particularly in terms of accurately specifying the theory of change, the same basic structure and concepts may be helpful nonetheless.

Policy Transportation and Adaptation

The external validity of impact evaluations is often framed as a question of "would the same policy work in another context?" In practice, however, it is usually necessary to make at least some adaptations to a policy in order for it to work in a new context. These can be superficial, as in the translation of program materials into a different language, or more substantive, for example by adapting the nutrition advice component of BINP to include not just mothers but their husbands and mothers-in-law. Even where such adaptation is not strictly necessary, appropriate adaptations may sometimes be able to optimize an already effective program. But the number of adaptations that could be made to a policy or intervention is nearly infinite—*which* aspects should be targeted for adaptation, and which left alone? And *how much* should a policy that was successful in another context be adapted, since adaptations risk changing aspects of the policy that make it effective?

Across social science disciplines, these questions of adaptation are even less well studied than questions of external validity, with only *ad hoc* (if any) discussion of how external validity concerns should be addressed through adaptations. The literature is most advanced in social policy and applied psychology, where there is a well-worn debate on the trade-offs between fidelity to evidence-backed interventions versus cultural adaptation of programs (Castro, Barrera, and Holleran Steiker 2010), although

even this literature offers little guidance to policy makers beyond conducting focus groups and small-scale pilots. A small literature examines this debate empirically, by comparing the effectiveness of various policies or programs according to whether they were newly designed ("homegrown"), transported but adapted, or transported without adaptation. The results are mixed: Hasson et al. (2014) compare 307 German and Swedish social interventions and find that novel and adapted programs are more effective than programs that are transported without adaptation; Leijten et al. (2016) find no difference on average between homegrown and transported parenting interventions across a range of countries; and Gardner, Montgomery, and Knerr (2015) find that several branded parenting interventions developed in the United States and Australia are at least as effective in non-Western countries, even with little adaptation. Of course, the challenge of trying to use meta-analytic methods to ascertain the optimal level of adaptation is that it is unclear exactly what changes in context the adaptations were responding to, or how appropriate the adaptations were. As with policy choice, knowing the average effectiveness of adapted policies is less useful for policy makers than guidance in identifying which adaptations are likely to be necessary and effective in their specific context.

The diagnostic aspect of mechanism mapping may help fill this gap. Since mechanism mapping as a diagnostic tool focuses on the interaction between a policy's theory of change and differences in context, the diagnosis of whether a policy is likely to be as effective in a new context as it was elsewhere inherently involves highlighting the aspects of the policy that should be targeted for adaptation. In the case of BINP, for example, fig. 3 makes it obvious that the key aspect where adaptation was necessary was the nutritional advice component, and specifically the individuals to whom this was targeted. The mechanism map alone is not sufficient to determine exactly what the adaptation should be—whether it is possible to simply include husbands and mothers-in-law through the existing delivery mechanism, for example, requires additional context-specific knowledge and feasibility investigations, as in any policy design process—but it can help identify which aspect of the policy is problematic, and why. Similarly, fig. 3 makes clear that the other steps in BINP's theory of change fit well with the contextual assumptions and previous context in which the program had been evaluated, suggesting that there is little need for adaptation in these respects. The design of the resulting adapted policy is thus informed both by evaluation evidence from other contexts—through the aspects of the original policy that were maintained in the new context, as well as by local, context-specific knowledge, through the aspects that were adapted.

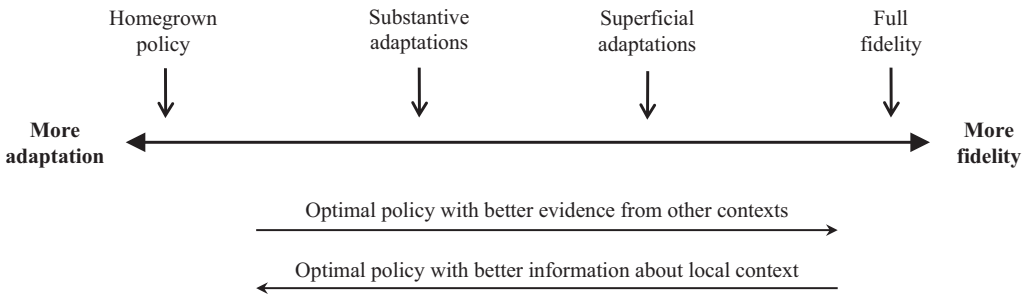
The same framework can also be used to identify adaptations that might be necessary (or potentially detrimental) in scaling up a policy that was successful in a small-scale trial. Most obviously, contextual assumptions that held in the trial may not hold when implementing at scale. For example, government agents may require different incentive or monitoring schemes than nongovernmental agents in order to elicit

similar effort levels (Cameron and Shah 2017), or incentive schemes that were effective for nongovernmental implementers in small trials may be unfeasible for political economy reasons when implemented at scale (Bold et al. 2018), necessitating adaptation of the policy for scale-up. Similarly, adaptations might be imposed by the scale-up process itself, as in the Tools of the Mind scale-up where implementation at scale in schools without close experimental control led unstructured-but-crucial components of the program to be crowded out by other demands on school time. Where such enforced adaptations or risks can be foreseen in advance, mechanism mapping can provide a framework for thinking through their potential consequences and possible mitigating measures.

Of course, making even apparently superficial adaptations to an intervention also creates new potential mismatches between the contextual assumptions for the adapted theory of change and contextual realities. In practice, then, policy makers need to repeat the diagnostic process outlined in the previous section on the adapted theory of change, then potentially readapt, and so on in an iterative process. In this way, mechanism mapping complements another recent innovation: adaptive policy making, which views policy design and evaluation as an iterative process of experimentation with tight feedback loops (Pritchett, Samji, and Hammer 2013; World Bank 2015). As a policy diagnostic tool that focuses on mechanism, context, and their interaction, mechanism mapping can be integrated with adaptive policy making processes to help connect experimentation to a more precise diagnosis of the effectiveness of previous iterations of a policy, thus adding precision to the experimental search process. Specifically, policy makers can use the weakest assumption link in the theory of change (as identified by mechanism mapping) as the starting point for adaptive experimentation. Similarly, since monitoring and data collection strategies are typically based a policy's theory of change or logframe, the data that organizations generate during adaptive policy making processes often closely aligns with the evidence required to make mechanism mapping more empirically rigorous.

With respect to the second question, on the optimal extent of adaptation, mechanism mapping's simultaneous use of evaluation evidence from other contexts and knowledge of the local context highlights a fundamental trade-off. On one hand, evaluation evidence on a policy's effectiveness in other contexts is likely to be more rigorous (especially in causal identification) than information about the local context. However, relying on this evidence requires strict fidelity to the original policy design, implying that policy makers should refrain from making adaptations as much as possible. On the other hand, using mechanism mapping to identify potential adaptations can make efficient use of local information, which (even if less rigorous) is specific to the context in question. However, using this local information to make adaptations decreases the relevance of evaluation evidence from elsewhere. There is therefore a trade-off between evidence from other contexts and local knowledge of the current

Figure 4. The Fidelity-Adaptation Spectrum



Source: Author's elaboration.

context, making it unclear how quick policy makers should be to make adaptations when they have identified a difference in context. Should adaptations be made in response only to major differences, or also to minor ones? And what constitutes a major or minor difference?

There is no universally optimal solution to these questions, because the characteristics of a specific context—however apparently minor or idiosyncratic they are—can undermine the effectiveness of even the most evidence-backed policy, yet policy makers' ability to foresee these interactions is limited (hence the need to evaluate policies and use evidence in the first place). That said, the respective roles of evidence and local information suggests that the optimal extent of adaptation will vary from case to case, depending on several factors. Figure 4 illustrates these trade-offs.

First, to the extent that available impact evaluation evidence on the policy is strong, consistent, and from similar contexts, policy makers should make fewer adaptations (all else equal). These factors reduce the uncertainty associated with a policy's impact in its current form, thus increasing the risk that adaptations could backfire. For example, Evans and Popova (2015), Pritchett (2017), and Vivalt (2017) show that some types of development interventions exhibit much greater variance across trials, suggesting that some interventions are more sensitive to contextual differences than others and thus presumably have a greater need for adaptation.

Second, the greater the policy maker's information about the local context—whether in the form of formal evidence and data, or simply familiarity and tacit knowledge—the more a policy maker should be willing to adapt a policy, since this knowledge allows for better identification of negative or positive context-mechanism interactions as well as suitable adaptations. This implies that the optimal level of adaptation in a specific case will vary not only by policy area and country, but also by the information set of the policy maker: *ceteris paribus*, a policy maker with less familiarity, information, or ability to gather information about the target context should

generally make fewer adaptations to a transported policy than a policy maker who either has or can obtain more detailed contextual information.

Third, the optimal level is also likely to vary according to the nature of the policy process. An extensive participatory policy design process that engages key stakeholders and beneficiaries will elicit more local information and is more likely to lead to useful adaptations than a quick decision made by an individual policy maker (Parker et al. 2008; Leviton and Trujillo 2017). For instance, although the World Bank based the design of BINP on its successful program in Tamil Nadu, the Bank's own evaluation found that the view that "project design and implementation should have sought to broaden the target audience for its nutritional messages…was expressed by BINP fieldworkers and women themselves in project areas during field visits" (World Bank 2005b, 9). When such participatory processes are not practically or politically feasible, or when policy makers do not have time to gather extensive data on actual contextual characteristics and are thus forced to rely on their own knowledge, policy makers should weight evidence from elsewhere relatively more heavily and thus usually make fewer adaptations.

Finally, making appropriate adaptations to a policy requires understanding not only of the context but also of the policy's mechanism, since what matters for impact is the interaction between mechanism and context. While better understanding of the mechanism is unrelated to the optimal level of adaptations to a policy (unlike better contextual information), one would expect it to lead to more successful adaptations. This is an area in which bureaucratic expertise and research—in particular high-quality evaluations or systematic reviews that are able to trace mechanisms—can be especially useful. Effective policy adaptation thus stems not just from contextual knowledge, but also from its combination with rigorous evidence and professional judgment.

Conclusion

As the harm caused by the neuropsychiatric interaction between the HIV drug efavirenz and the rare genetic variant common in Zimbabwe's population became evident, some of the same scientists who predicted the problem designed a strategy to address it. "[T]he current 'one size fits all' [efavirenz] dose strategy in sub-Saharan Africa needs to be carefully reevaluated by considering integration of an individualized therapeutic approach" that combines individualized testing, monitoring, and dosing adjustment (Masimirembwa, Dandara, and Leutscher 2016, 578). As one of the scientists, Collen Masimirembwa, stated: "It's not a bad drug. We just know it can be improved in Africa" (Nordling 2017, 20).

Just as the spread of precision medicine promises to move medical treatment beyond one-size-fits-all recommendations, so too is it necessary for impact evaluators and policy makers to find ways to make evidence-based policy making more

responsive to the particularities of specific contexts. While impact evaluations and systematic reviews provide excellent starting points for doctors and policy makers alike, even before the advent of precision medicine actual medical practice has always required doctors to combine rigorous research evidence with their individual expertise and judgment for each case (Deaton 2010). While the rapidly growing external validity literature has focused largely on the *generalizability* of a policy *from* a specific context, the relevant question for policy makers is the *applicability* of evidence *to* their specific context. This requires an understanding of the interactions between a policy's theory of change, as supported by contextual assumptions, and the actual characteristics of the context to which a policy is being transported. This article has introduced mechanism mapping as a flexible and conceptually simple diagnostic tool that may help policy makers assess the fit of evidence-based policies with their own contexts and identify the specific features of a policy that are most likely to need adaptation in the new context.

The rapid growth of evidence-based policy has created a rich pool of rigorous impact evaluations, and a growing literature on external validity has provided theoretical and empirical tools for generalizing these results beyond their original context. However, this paper has highlighted the paucity of guidance to policy makers on how to solve the "last mile" problems of policy transportation and scale-up: the need to bridge the inevitable gap between the best evidence available from other contexts and the particularities of their own context, and whether and how to adapt evidence-based policies to better fit these contextual idiosyncrasies. While the mechanism mapping framework is potentially useful in this regard, there is a need for more and better research on key questions such as how policy makers tend to make these contextual-fit assessments and adaptations in practice, how they update their beliefs and make adaptations based on this, and how research and institutional structures can improve their judgment in these regards. Questions like these represent the basis for an intellectually rich and policy-relevant agenda for researchers and practitioners alike.

Notes

Martin J. Williams is Associate Professor in Public Management, University of Oxford, Blavatnik School of Government, email: martin.williams@bsg.ox.ac.uk. The author is grateful for conversations and comments from Jon Ahlberg, Noam Angrist, Alex Baron, Maria Barron Rodriguez, Eleanor Carter, Suvojit Chattopadhay, David Evans, Flavia Galvani, Frances Gardner, Julie Hennegan, David Humphreys, Robert Klitgaard, Julien Labonne, Adauto Modesto, Aoife O'Higgins, Daniel Rogger, students at the Blavatnik School and the Escola Nacional de Administração Pública (Brazil), seminar audiences at the Blavatnik School and Oxford's Global Priorities Institute, and three anonymous referees. Any remaining errors are his own. A policy memo based on this paper with a five-step "how-to" guide is also available (Williams 2017).

1. This study uses the terms "policy," "intervention," and "program" interchangeably throughout, since the distinctions between them are not relevant for this paper's purposes.

2. While there are numerous different disciplinary and institutional approaches and terminologies associated with elaborating theories of change (e.g., DFID 2012; De Silva et al. 2014), the aim of this paper is not to adjudicate the debate between these various approaches, nor to suggest a best practice for doing so. Rather, this paper takes a simple approach in order to focus on the core concepts. This approach can equally be applied to theories of change written in different formats.

3. Meta-analysis is intended to capture real variation from differences in context as well as random statistical variation from chance; as discussed previously; the latter is outside the scope of this paper.

4. This point is not meant to caricature the views of authors of systematic reviews, most of whom have appropriately nuanced views of how systematic reviews should be used by policy makers, but simply to clarify the conceptual limitations of the "headline" average treatment effect that readers often focus on.

5. This example is based on Cartwright and Hardie's excellent exposition (2014, 80–84), as well as on Save the Children (2003), White (2005), and World Bank (2005a, b). This article's discussion of TINP and BINP and their contexts is, of course, simplified for clarity and brevity.

6. Noam Angrist, personal communication, February 19, 2018.

7. Angrist, personal communication, February 19, 2018.

References

Allcott, H. 2015. "Site Selection Bias in Program Evaluation." *Quarterly Journal of Economics* 13 (3): 1117–65.

Andrews, I., and E. Oster. 2018. "Weighting for External Validity." NBER Working Paper 23826, National Bureau of Economic Research, Cambridge, MA.

Angrist, N. 2017. "An Application of the Jump from Internal to External Validity: Transporting an HIV Prevention Intervention from Kenya to Botswana, and Testing Two Scale Models Across Key Parameters of Heterogeneity." Working Paper, Blavatnik School of Government, CSAE Research Workshops, Oxford University, Oxford, UK.

Angrist, J., and I. Fernandez-Val. 2010. "Extrapolating: External Validity and Overidentification in the LATE Framework." NBER Working Paper 16566, National Bureau of Economic Research, Cambridge, MA.

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shotland, and M. Walton. 2016a. "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application." NBER Working Paper 22931, National Bureau of Economic Research, Cambridge, MA.

Banerjee, A., S. Chassang, and E. Snowberg. 2016b. "Decision Theoretic Approaches to Experiment Design and External Validity." NBER Working Paper 22167, National Bureau of Economic Research, Cambridge, MA.

Barzelay, M. 2007. "Learning from Second-Hand Experience: Methodology for Extrapolation-Oriented Case Research." *Governance* 20 (3): 521–43.

Bates, M. A., and R. Glennerster. 2017. "The Generalizability Puzzle." *Stanford Social Innovation Review*, Summer (accessed June 22, 2017), https://ssir.org/articles/entry/the_generalizability_puzzle.

Bold, T., M. Kimenyi, G. Mwabu, A. Ng'ang'a, and J. Sandefur. 2018. "Experimental Evidence on Scaling Up Education Reforms in Kenya." *Journal of Public Economics* 168: 1–20.

Bonell, C., F. Jamal, G. J. Melendez-Torres, and S. Cummins. 2014. "'Dark Logic': Theorising the Harmful Consequences of Public Health Interventions." *Journal of Epidemiology and Community Health* 69 (1): 95–98.

Cameron, L., and M. Shah. 2017. "Scaling Up Sanitation: Evidence from an RCT in Indonesia." IZA Discussion Paper 10619, Institute of Labor Economics, Bonn, Germany.

Cartwright, N., and J. Hardie. 2014. *Evidence-Based Policy: A Practical Guide to Doing It Better.* Oxford: Oxford University Press.

Castro, F. G., M. Barrera, Jr, and L. K. Holleran Steiker. 2010. "Issues and Challenges in the Design of Culturally Adapted Evidence-Based Interventions." *Annual Review of Clinical Psychology* 6: 213–39.

Christensen, G., and E. Miguel. 2016. "Transparency, Reproducibility, and the Credibility of Economics Research." NBER Working Paper 22989, National Bureau of Economic Research, Cambridge, MA.

Davies, R. 2004. "Scale, Complexity, and the Representation of Theories of Change." *Evaluation* 10 (1): 101–21.

De Silva, M. J., E. Breuer, L. Lee, L. Asher, N. Chowdhary, C. Lund, and V. Patel. 2014. "Theory of Change: A Theory-Driven Approach to Enhance the Medical Research Council's Framework for Complex Interventions." *Trials* 15 (1): 267–78.

Deaton, A. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48 (2): 424–55.

Deaton, A., and N. Cartwright. 2016. "Understanding and Misunderstanding Randomized Controlled Trials." NBER Working Paper 22595, National Bureau of Economic Research, Cambridge, MA.

Department for International Development (DFID). 2012. "Review of the Use of 'Theory of Change' in International Development." Review Report. Department for International Development, London, UK.

Diamond, A., W. S. Barnett, J. Thomas, and S. Munro. 2007. "Preschool Program Improves Cognitive Control." *Science* 318 (5855): 1387–88.

Dupas, P. 2011. "Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya." *American Economic Journal: Applied Economics* 3: 1–34.

Evans, D., and A. Popova. 2015. "What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews." World Bank Policy Research Working Paper 7203, World Bank, Washington, DC.

Farran, D., and S. J. Wilson. 2014. "Achievement and Self-Regulation in Pre-Kindergarten Classrooms: Effects of the Tools of the Mind Curriculum." Working paper, Peabody Research Institute, Vanderbilt University, Nashville, TN.

Gardner, F., P. Montgomery, and W. Knerr. 2015. "Transporting Evidence-Based Parenting Programs for Child Problem Behavior (Age 3–10) Between Countries: Systematic Review and Meta-Analysis." *Journal of Clinical Child and Adolescent Psychology* 45 (6): 749–762.

Gechter, M. 2016. "Generalizing the Results from Social Experiments: Theory and Evidence from Mexico and India." Mimeo, Pennsylvania State University, University Park, PA.

Gertler, P., S. Martinez, P. Premand, L. Rawlings, and C. Vermeersch. 2016. *Impact Evaluation in Practice.* 2nd ed. Washington, DC: World Bank.

Greenhalgh, T., F. Macfarlane, L. Steed, and R. Walton. 2016. "What Works for Whom in Pharmacist-Led Smoking Cessation Support: Realist Review." *BMC Medicine* 14: 209–24.

Hasson, H., K. Sundell, A. Beelmann, and U. von Thiele Schwarz. 2014. "Novel Programs, International Adoptions, or Contextual Adaptations? Meta-analytical Results from German and Swedish Intervention Research." *BMC Health Services Research* 14 (Suppl. 2): O32.

Hawe, P. 2015. "Lessons from Complex Interventions to Improve Health." *Annual Review of Public Health* 36: 307–23.

Henrich, J., S. J. Heine, and A. Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33 (2-3): 61–135.

HM, Treasury. 2011. "The Magenta Book: Guidance for Evaluation." HM Treasury, London, UK.

Kowalski, A. E. 2018. "How to Examine External Validity Within an Experiment". NBER Working Paper 24834, National Bureau of Economic Research, Cambridge, MA.

Kristjansson, E., D. Francis, S. Liberato, M. B. Jandu, V. Welch, M. Batal, T. Greenhalgh, T. Rader, E. Noonan, B. Shea, L. Janzen, G. Wells, and M. Petticrew. 2015. "Food Supplementation for Improving the Physical and Psychosocial Health of Socio-economically Disadvantaged Children Aged Three Months to Five Years: A Systematic Review." *Campbell Systematic Reviews*, 4.

Leijten, P., G. J. Melendez-Torres, W. Knerr, and F. Gardner. 2016. "Transported Versus Homegrown Parenting Interventions for Reducing Disruptive Child Behavior: A Multilevel Meta-Regression Study." *Journal of the American Academy of Child and Adolescent Psychiatry* 55 (7): 610–17.

Leviton, L. C. 2017. "Generalizing about Public Health Interventions: A Mixed-Methods Approach to External Validity." *Annual Review of Public Health* 38: 371–91.

Leviton, L. C., and M. D. Trujillo. 2017. "Interaction of Theory and Practice to Assess External Validity." *Evaluation Review* 41 (5): 436–71.

Low, H., and C. Meghir. 2017. "The Use of Structural Models in Econometrics." *Journal of Economic Perspectives* 31 (2): 33–58.

Ludwig, J., J. Kling, and S. Mullainathan. 2011. "Mechanism Experiments and Policy Evaluations." *Journal of Economic Perspectives* 25 (3): 17–38.

Marchal, B., S. van Belle, J. van Olmen, T. Hoeré, and G. Kegels. 2012. "Is Realist Evaluation Keeping Its Promise? A Review of Published Empirical Studies in the Field of Health Systems Research." *Evaluation* 18 (2): 192–212.

Masimirembwa, C., C. Dandara, and P. D. Leutscher. 2016. "Rolling out Efavirenz for HIV Precision Medicine in Africa: Are We Ready for Pharmacovigilance and Tackling Neuropsychiatric Adverse Effects?" *OMICS: A Journal of Integrative Biology* 20 (10): 575–580.

Miguel, E., and M. Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72 (1): 159–217.

Moore, G. F., S. Audrey, M. Barker, L. Bond, C. Bonell, W. Hardeman, L. Moore, A. O'Cathain, T. Tinati, D. Wight, and J. Baird. 2015. "Process Evaluation of Complex Interventions: Medical Research Council Guidance." *British Medical Journal*: 350.

Muller, S. 2015. "Causal Interaction and External Validity: Obstacles to the Policy Relevance of Randomized Evaluations." *World Bank Economic Review* 29 (supp): S217–S225.

Muralidharan, K., and P. Niehaus. 2017. "Experimentation at Scale." *Journal of Economic Perspectives* 31 (4): 103–124.

Nordling, L. 2017. "Putting Genomes to Work in Africa." *Nature* 544: 20–22.

Nyakutira, C., D. Röshammar, E. Chigutsa, P. Chonzi, M. Ashton, C. Nhachi, and C. Masimirembwa. 2008. "High Prevalence of the CYP2B6 516G→T(*6) Variant and Effect on the Population Pharmacokinetics of Efavirenz in HIV/AIDS Outpatients in Zimbabwe." *European Journal of Clinical Pharmacology* 64 (4): 357–65.

Parker, E., B. Israel, T. Robins, G. Mentz, X. Lin, W. Brakefield-Caldwell, E. Ramirez, K. Edgren, M. Salinas, and T. Lewis. 2008. "Evaluation of Community Action Against Asthma: A Community Health Worker Intervention to Improve Children's Asthma-Related Health by Reducing Household Environmental Triggers for Asthma." *Health Education and Behavior* 35 (3): 376–95.

Pawson, R., and N. Tilley. 1997. *Realistic Evaluation*. London: SAGE Publications.

Petticrew, M., P. Tugwell, E. Kristjansson, S. Oliver, E. Ueffing, and V. Welch. 2012. "Damned if You Do, Damned if You Don't: Subgroup Analysis and Equity." *Journal of Epidemiology and Community Health* 66 (1): 95–98.

Pritchett, L. 2017. "'The Evidence' About 'What Works' in Education: Graphs to Illustrate External Validity and Construct Validity." CGD Notes, Center for Global Development, Washington, DC. RISE Insight, June.

Pritchett, L., and J. Sandefur. 2015. "Learning from Experiments when Context Matters." *American Economic Review: Papers and Proceedings* 105 (5): 471–5.

Pritchett, L., S. Samji, and J. Hammer. 2013. "It's All About MeE: Using Structured Experiential Learning ("e") to Crawl the Design Space." Working Paper 322, Center for Global Development, Washington, DC.

Rajman, I., L. Knapp, T. Morgan, and C. Masimirembwa. 2017. "African Genetic Diversity: Implications for Cytochrome P450-mediated Drug Metabolism and Drug Development." *EBioMedicine* 17: 67–74.

Ravallion, M. 2009. "Evaluation in the Practice of Development." *World Bank Research Observer* 24: 29–53.

Rigterink, A. S., and M. Schomerus. 2016. "Probing for Proof, Plausibility, Principle and Possibility: A New Approach to Assessing Evidence in a Systematic Evidence Review." *Development Policy Review* 34 (1): 5–27.

Rogers, P. 2008. "Using Programme Theory to Evaluate Complicated and Complex Aspects of Interventions." *Evaluation* 14 (1): 29–48.

Save the Children. 2003. "Thin on the Ground: Questioning the Evidence Behind World Bank-funded Community Nutrition Projects in Bangladesh, Ethiopia, and Uganda." Policy Report.

Vivalt, E. 2017. "How Much Can We Generalize from Impact Evaluations?" Mimeo, Stanford University, Palo Alto, CA.

White, H. 2005. "Comment on Contributions Regarding the Impact of the Bangladesh Integrated Nutrition Project." *Health Policy and Planning* 20 (6): 408–11.

Williams, M. J. 2017. "External Validity and Policy Adaptation: A Five-step Guide to Mechanism Mapping." Policy Memo, Blavatnik School of Government, Oxford University, Oxford, UK, https://www.bsg.ox.ac.uk/research/publications/external-validity-and-policy-adaptation-0.

Wong, G., G. Westhorp, A. Manzano, J. Greenhalgh, J. Jagosh, and T. Greenhalgh. 2016. "RAMESES II Reporting Standards for Realist Evaluations." *BMC Medicine* 14: 96–113.

World Bank. 2005a. "The Bangladesh Integrated Nutrition Project: Effectiveness and Lessons." Bangladesh Development Series Paper No. 8, World Bank, Washington, DC.

———. 2005b. "Project Performance Assessment Report: Bangladesh Integrated Nutrition Project." Report No. 32563, World Bank, Washington, DC.

———. 2015. "World Development Report 2015: Mind, Society, and Behavior." World Bank, Washington, DC.

Young, 1ove. 2016. "Theory of Change: Relative Risk Information Campaign." Young 1ove, Gaborone, Botswana.

Appendix A1: Example of a Complicated Theory of Change

Figure A1.1. Theory of Change: Supply Chains 4 Community Case Management (SC4CCM)

Source: Reproduced from "SC4CCM Project Theory of Change," http://sc4ccm.jsi.com/emerging-lessons/theory-of-change/.

Appendix A2: Example Theory of Change with Competing Mechanisms and Potential Negative Outcomes

One important consideration in the policy design process is the uncertainty of how a policy will work in practice—that is, the potential for multiple competing mechanisms. Some of these potential impact pathways may lead to positive outcomes while others lead to null or even negative outcomes. In its simplest form, the process of mechanism mapping can help policy makers understand the plausibility of their intended mechanism. However, the same logic can be applied to assess the plausibility of multiple mechanisms, some of which may lead to negative rather than positive outcomes.

To illustrate this, consider the theory of change constructed in 2016 by the Botswanan NGO Young 1ove for its planned transportation of the "sugar daddies" informational intervention on teenage HIV infection risk in Kenya (Dupas 2011) in figure A2.1. It considers the potential for eight different competing mechanisms, half of which would lead to unintended outcomes in terms of HIV rates, pregnancy incidence, and school attendance. Although this theory of change does not exactly follow the suggested mechanism mapping structure, it does explicitly consider the key assumptions underlying each step of the eight potential mechanisms. Many of these assumptions are framed as empirical questions on which additional descriptive data could be collected to help establish the plausibility of each of these mechanisms, and indeed the NGO did collect descriptive data on many of these assumptions prior to beginning the trial.[6] While Young 1ove's example shows that the basic concepts underlying mechanism mapping are intuitive and can be implemented even without explicit reference to the tool, the structured guidance in undertaking this process presented in this paper may nonetheless be useful for other organizations and policy makers.

This example also illustrates the feasibility of the mechanism mapping process for policy makers in the Global South and elsewhere. Young 1ove developed this theory of change through a six-month collaborative process led by its Batswana staff, which reportedly helped increase understanding of the program, surfaced key assumptions, and led to adaptations to the original design of the policy in Kenya.[7]

Figure A2.1. Theory of Change: Young 1ove

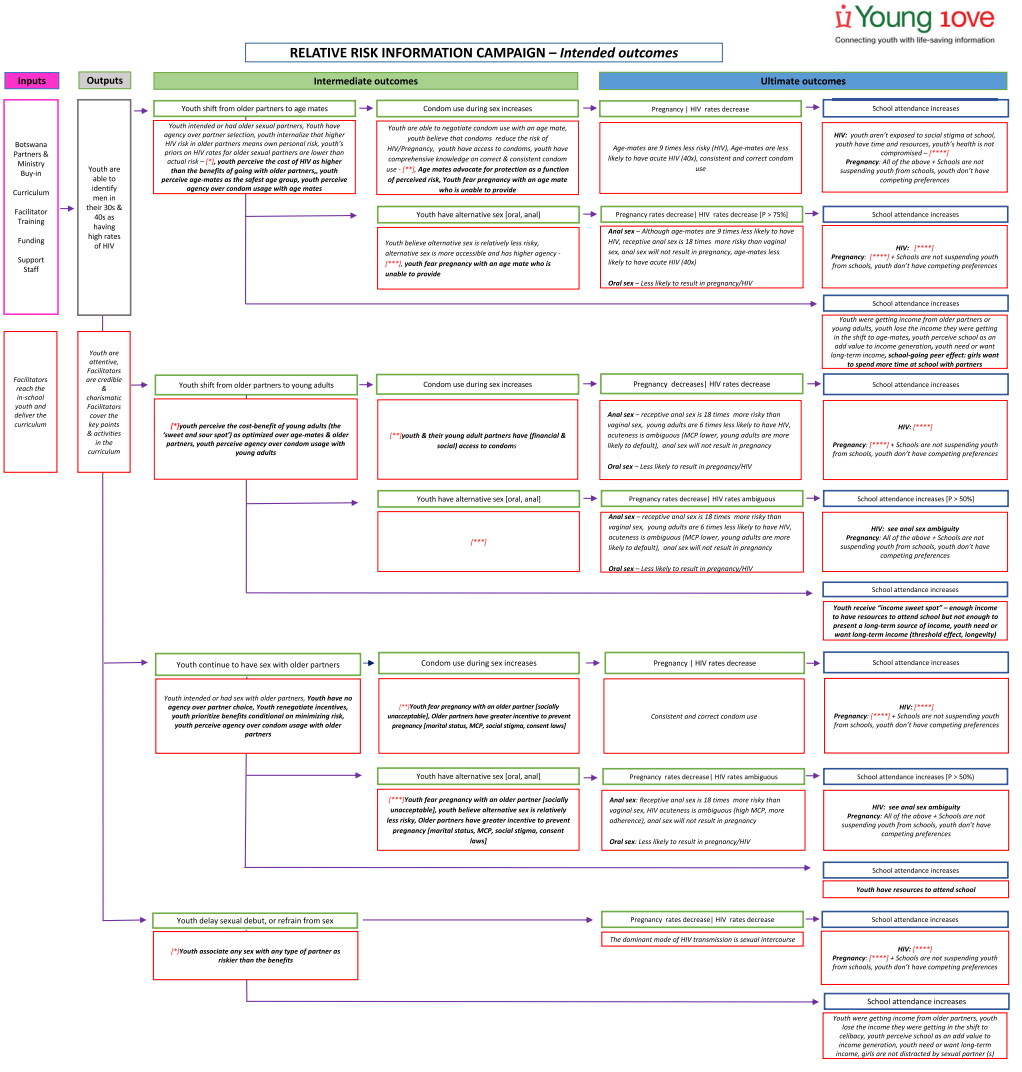
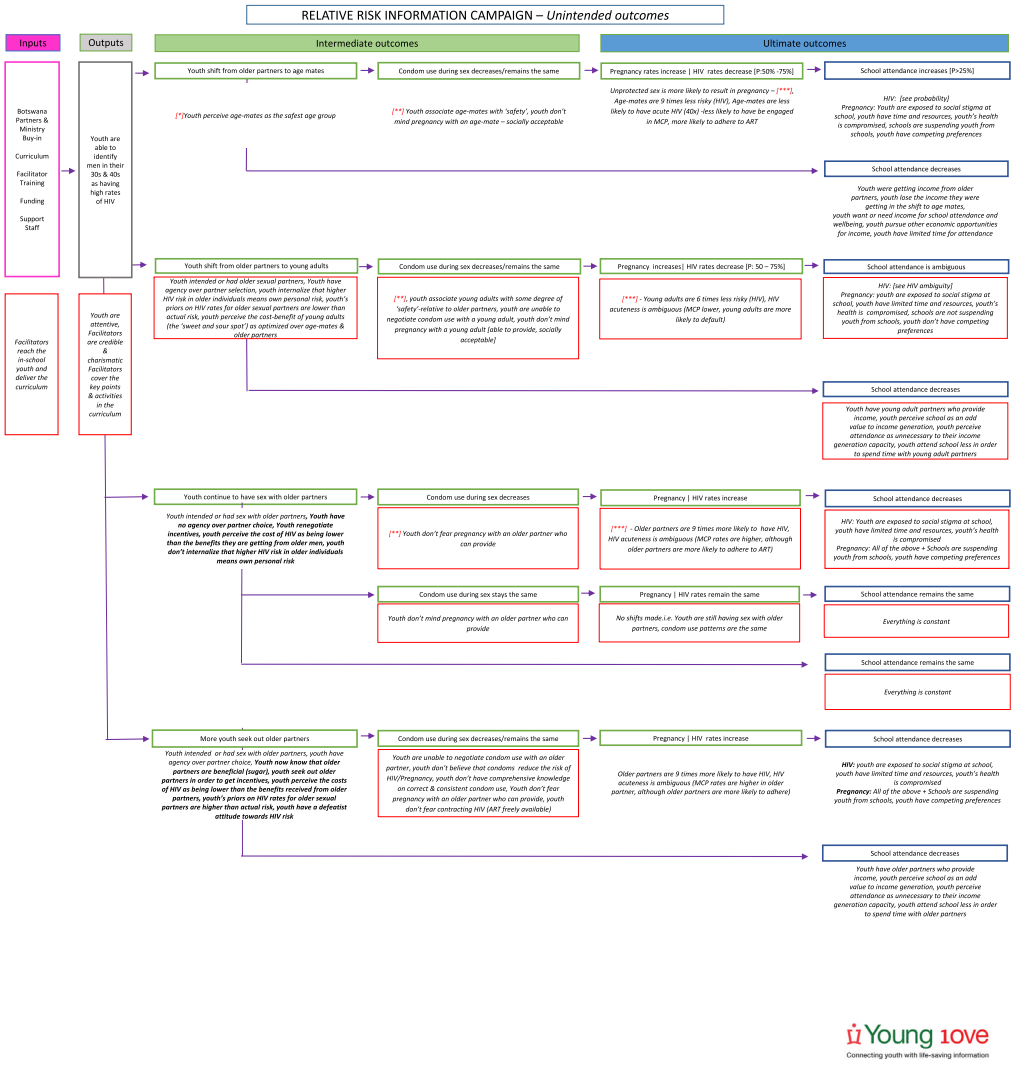


Figure A2.1. continued.



Source: Reproduced from Young 1ove (2016), "Theory of Change: Relative Risk Information Campaign," 25 April.
Notes: 1. Pregnancy fluctuations reflect pregnancy incidences; 2. Under all anal sex scenarios pregnancy rates will always decrease.