BSG Working Paper Series

Providing access to the latest policy-relevant research

BLAVATNIK SCHOOL OF GOVERNMENT

UNIVERSITY OF OXFORD

External validity and policy adaptation

From impact evaluation to policy design

BSG-WP-2017/019

July 2017

Martin J. Williams, Blavatnik School of Government, University of Oxford

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

Martin J. Williams *

July, 2017

Abstract

With the growing number of rigorous impact evaluations worldwide, the question of how best to apply this evidence to policymaking processes has arguably become the main challenge for evidence-based policymaking. How can policymakers predict whether a policy will have the same impact in their context as it did elsewhere, and how should this influence the design and implementation of policy? This paper introduces a simple and flexible framework to address these questions of external validity and policy adaptation. I show that all failures of external validity arise from an interaction between a policy's theory of change and a dimension of the context in which it is being implemented, and develop 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 in a new context, I show there is a fundamental informational trade-off between the strength and relevance of evidence on the policy from other contexts and the policymaker's knowledge of the local context. This trade-off can guide policymakers' judgments about whether policies should be copied exactly from elsewhere, adapted, or invented anew.

*Associate Professor in Public Management, University of Oxford, Blavatnik School of Government. Email: martin.williams@bsg.ox.ac.uk. I am grateful for conversations and comments from Alex Baron, Maria Barron Rodriguez, Eleanor Carter, David Evans, Flavia Galvani, Julie Hennegan, Robert Klitgaard, Julien Labonne, Aoife O'Higgins, Daniel Rogger, and students in the 2016-17 Policy Evaluation course at the Blavatnik School. Any remaining errors are my own. A policy memo based on this paper is available at http://www.martinjwilliams.com/research

1 Introduction

In 2015, Zimbabwe's government rolled out a new 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 droves" (Nordling 2017). 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 happens to be quite common in Zimbabwe (Masimirembwa et al 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 policymakers in Zimbabwe to make a costly mistake.

Policymakers worldwide face similar challenges in trying to apply evidence to policy decisions. With the recent boom in impact evaluations around the world, policymakers 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 policymaking is that the design of policy can be based on evidence of what has worked in other contexts, rather than each policymaker 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 (Rodrik 2009, Deaton 2010, Pritchett and Sandefur 2015, Vivalt 2016).[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 (Bold *et al* 2016, Banerjee *et al* 2016a). This is the problem of the *external validity* of impact evaluations. The limited external validity of impact evaluation evidence poses challenges for policymakers: 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 policymakers 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 proposes 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 mapping of its intended mechanism 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, etc. - and the theory of change includes factors related to implementation as well as impact, this parsimonious framework encompasses the variety of typologies of external validity failures discussed in existing literature (Deaton 2010, Cartwright and Hardie 2014, Banerjee et al 2016a).

This framework for understanding external validity failures can be used by policymakers to

[1] I use the terms policy, intervention, and program interchangeably throughout, since the distinctions between them are not relevant for this paper's purposes.

analyze whether a successful policy from another context can be transported to their context. Comparing the policy's theory of change against its underlying contextual assumptions - what this paper calls *mechanism mapping* - focuses policymakers' attention on the validity of these contextual assumptions. If a necessary assumption does not hold in the new context in the same way it held in the old context, then the mechanism will be interrupted and the policy's final 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.

Mechanism mapping ideally consists of a systematic process of seeking empirical evidence to support contextual assumptions 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 *et al* 2011) to validate each step of the theory of change and its underlying contextual assumptions. However, where time or resource constraints make this unfeasible, mechanism mapping can also be useful as a quick and informal desk exercise undertaken by a single policymaker. This simple and intuitive diagnostic process gives policymakers 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 data has often hindered evidence-based policymaking in data-poor contexts, mechanism mapping's ability to integrate less formal types of evidence makes it particularly well suited to such contexts. Section 4 discusses the application of mechanism mapping in more detail.

The process of mechanism mapping also feeds directly into policy adaptation, by identifying 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 diagnostic 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 (i.e. internally valid) 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 policymaker'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 policymaker and the nature of the policymaking process. This implies, for example, that an expatriate donor official should generally make fewer adaptations to a transported policy than a policymaker from the country with deeper local knowledge, all else equal.

The mechanism mapping approach suggested by this paper builds on and complements other approaches outlined in the small but growing literature on external validity. One existing approach is to try to determine the average effect of an intervention across different contexts and implementations, by comparing the results of multiple studies. This is, broadly speaking, the approach of meta-analysis and systematic review (What Works Network 2014, Vivalt 2016).[2] But no context is "average" across all dimensions, and while meta-analyses and systematic reviews can sometimes disaggregate average effectiveness across a handful of variables, the multi-dimensionality of contexts and policies will always mean that each context and each implementation of a policy - like the genetic makeup of each human - will be unique in some way. While these approaches usefully summarize and synthesize existing evidence and thus provide an excellent starting point for policy analysis and design, they offer little structured

[2]These methods can of course be used with greater nuance, as I discuss in Section 3.

guidance on what role this inevitable particularity of context should play in the transportation and adaptation of policies.

A second approach in the literature focuses on what evaluators can do to increase the external validity of a particular study. A range of methods have been proposed: formal theory and structural modeling (Deaton 2010), larger evaluations (Muralidharan and Niehaus 2016), and integrating "structured speculation" on external validity into research papers (Banerjee et al 2016b). While these approaches can shed light on the external validity of a particular study, their focus is on predicting the generalizability of a specific study to an unspecified context, rather than the application of evidence from other contexts to a specific context about which the policymaker has context-specific information. While the former is useful, the latter is the core of the policymaker's problem. Finally, mechanism mapping is related to adaptive policymaking (Pritchett *et al* 2012, World Bank 2015) in emphasizing the use of local information to improve policy design; Section 5 discusses complementarities between the two approaches.

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), 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 simple, flexible, and parsimonious conceptual approach to understanding external validity that is closely linked to (2) a practical framework for identifying likely external validity failures, and which (3) feeds directly into the policy adaptation process. Finally, this paper differs from much of the external validity literature in focusing primarily on the policymaker's problem of how to apply a body of evidence to a specific policy problem, rather than on the researcher's or evaluator's problem of how to evaluate a specific policy in order to contribute to a wider body of evidence. Although the policymaker's problem is practically important and intellectually challenging, it has received considerably less scholarly attention.

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. Throughout this paper I therefore discuss published impact evaluations as if they represented true causal estimates of the policy's impact in that context, even though policymakers should obviously interpret published findings through a skeptical lens. Similarly, while this paper mainly discusses evidence from randomized controlled trials, this is not meant to diminish the importance of non-experimental evidence or privilege experiments. Rather, the exposition of the key issues of the application of evidence are simpler to explain when abstracting from how the evidence is generated.

The remainder of this paper proceeds as follows. Section 2 defines external validity and elucidates the understanding of external validity failures as an interaction of context and theory of change. Section 3 discusses the limitations of existing approaches to external validity, largely arising from the high dimensionality of policies and contexts. Section 4 describes the process of mechanism mapping in more detail and gives examples, Section 5 discusses adaptation and the fundamental informational trade-off, and Section 6 concludes by discussing the use and limitations of mechanism mapping.

2 A Simple Approach to External Validity

2.1 Defining External Validity

An impact evaluation's *external validity* refers to the generalizability of its findings beyond the study sample. 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 policymaker'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 policymakers 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.[3]

First, one might be interested in the generalizability of findings from the *study sample* (the individuals or units who actually participated in the evaluation) to the *target population* (the larger set of individuals to 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 group? This is the case of policy scale-up, as represented by panel (a) of Figure 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).[4] However, in some cases this is not possible - for example, many pharmaceutical trials are conducted on healthy individuals even though these are systematically different than the target population on important dimensions. Similarly, Henrich *et al* (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 statistical 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 non-governmental organization or academic research team, which could undermine effectiveness (Bold *et al* 2016, Cameron and Shah 2017). Treating a higher fraction of the target population could also lead to higher or lower effects through spillovers (Miguel and Kremer 2004, Crépon *et al* 2013).

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 Figure 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.

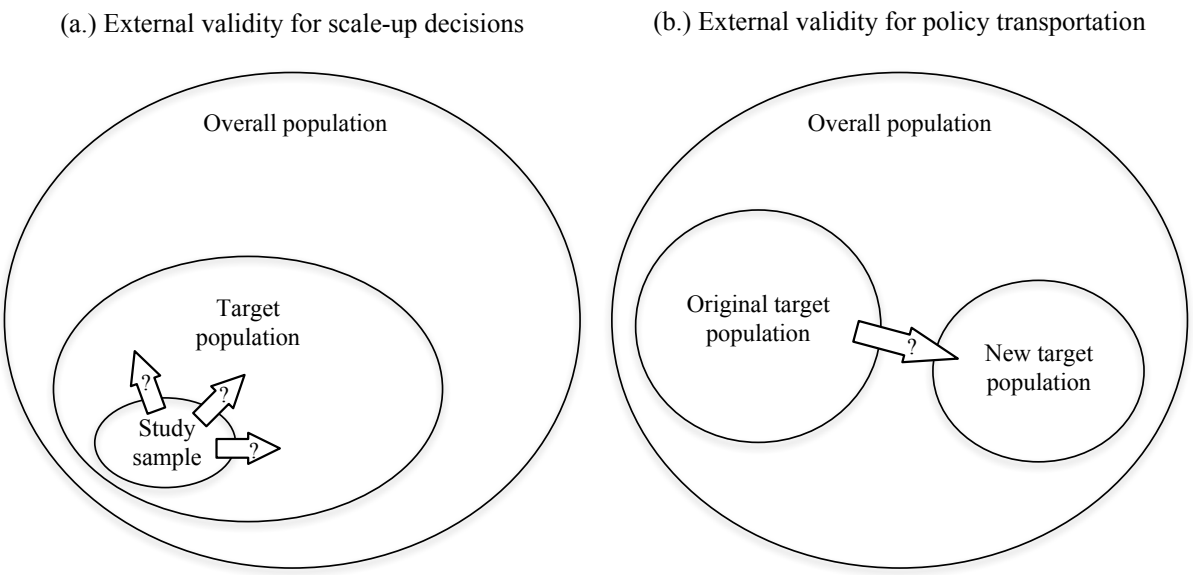
2.2 Failures of External Validity

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 policy specifies *outputs* that the government will deliver, which are hypothesized

[3]For a more technically precise discussion of these issues, see Deaton and Cartwright (2016).

[4]Random sampling is of course only sufficient for representativeness with large enough sample sizes, but as stated above, this article abstracts from this and other statistical issues.

Figure 1: Two Types of External Validity



to trigger a sequence of *intermediate outcomes* that lead to the final outcomes. To deliver these outputs, government plans to undertake a set of *activities*, which require certain *inputs* (e.g. financial or human resources, information).[5] 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 - Bates and Glennerster (2017) describe this as 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:

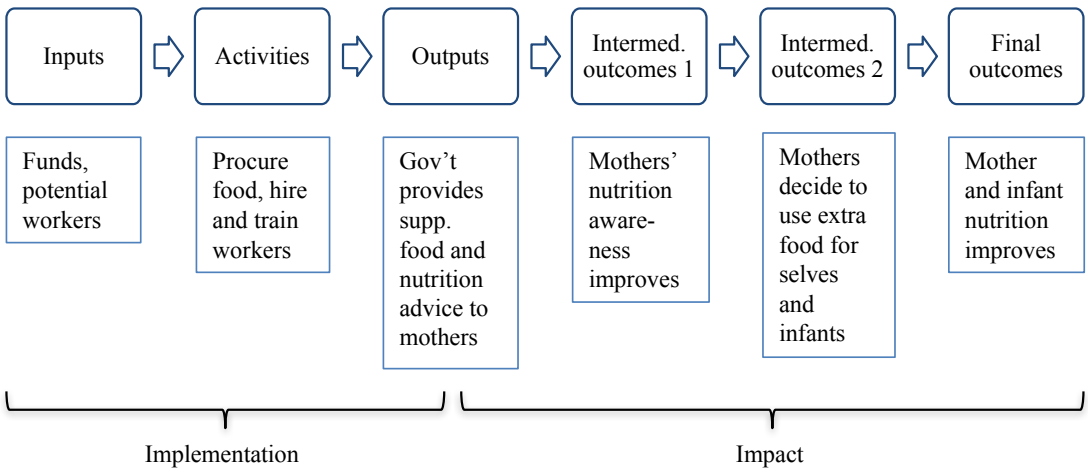
- Location, polity, or society in which the policy is being implemented (e.g. Iceland or India), together with all the social, cultural, economic, geographic, and political characteristics that vary across locations;
- Target groups (e.g. working adults, single mothers, at-risk teens);
- Time the policy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

[5]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, which can equally be applied to theories of change written in different formats.

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 we then observe that a policy had one impact in one context but the same policy 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 1990's 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 children malnutrition and stunting decreased significantly. Figure 2 maps a simple version of this theory of change.

Figure 2: Theory of Change: Bangladesh Integrated Nutrition Programme



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

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 allo-

cations (White 2005; 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 non-program 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. 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*. Section 4 below presents a structured approach to helping policymakers 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 Section 5 discusses later, policy adaptations can aim not just to mitigate threats from transportation to a new context but also to improve their effectiveness.

3 Existing Approaches to External Validity

With this framework in mind, this section briefly reviews existing approaches to dealing with external validity. While each is useful in some respects, they all have significant limitations in their ability to help policy designers to apply evidence to specific contexts. Although these approaches vary tremendously from empirical to theoretical and formal to informal, their common limitation is their inability to cope with the high dimensionality both of policies and of contexts. Policies are high dimensional in that even simple theories of change are composed of dozens, even hundreds of steps, tasks, and decisions, the failure of any of which can undermine the whole policy (as with the Tools of the Mind scale-up). Contexts are high dimensional in that they are defined by an almost infinite number of variables, each of which may or may not interact with a policy's theory of change. The existing approaches to external validity outlined below are only capable of dealing with heterogeneity along a handful of these dimensions, limiting their usefulness for policymakers.

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.[6] 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.[7]

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. Policymakers 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 policymakers, naïvely adopting a policy that has a positive average treatment effect in a systematic review is likely to backfire in many contexts.[8]

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 sub-group 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 sub-group 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 policymakers to compare their contexts to others on these covariates, and provides some guidance about which dimensions of context might matter for a given policy.[9]

While this information is useful, sub-group 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 sub-group analysis in meta-analyses 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 rare genetic variants - that are either difficult to measure or that evaluators would not think to measure *ex ante*, and are thus unobserved. Even where a given covariate

[6]The Cochrane and Campbell Collaborations and the UK government's What Works Network are the most prominent repositories of systematic reviews and meta-analyses.

[7]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.

[8]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 policymakers, but simply to clarify the conceptual limitations of the "headline" average treatment effect that readers often focus on.

[9]Economists have developed several other empirical methods to extrapolate results from one study to other populations, for example by exploiting selection and non-compliance within RCTs (Kowalski 2016) or by adjusting estimated impacts based on observed covariates (Angrist and Fernandez-Val 2010, Gechter 2016). However, for the purpose of policy design in a specific context, these methods share the same limitations as sub-group analysis.

is present across studies, one might question whether this variable interacts with the policy in the same way across contexts.[10] 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. The validity of proxies may also differ across contexts: for example, medical researchers sometimes use genetic data from African-American populations to extrapolate findings to African populations, even though African-Americans are not genetically representative of Africa as a whole (Rajman *et al* 2017).

As the explosion of impact evaluations in recent years has focused attention on questions of external validity, scholars have proposed various ways that authors of impact evaluations can improve the external validity of their studies. These proposals are implicitly directed at evaluators and academics, not policymakers: rather than discussing how to apply a body of evidence to a specific context, they aim to improve or clarify the generalizability of a particular study without a specific context in mind. While a full review of this growing literature is beyond the scope of this paper, it is important to note that - while welcome - none of these proposals fully address the policymaker's problem of predicting the impact of a policy in a specific context, due to the high dimensionality of contexts. For example, structural modeling can help evaluators understand behavior and mechanisms and make out-of-sample predictions (Deaton and Cartwright 2016), but such models can only incorporate a limited number of variables. As Low and Meghir (2017, 34) write:

> *"Structural economic models cannot possibly capture every aspect of reality, and any effort to do so would make them unwieldy for either theoretical insight or applied analysis. There will always be some economic choices left out of any particular model - the key question is how to judge what aspects to leave out without rendering the quantitative conclusions of the model irrelevant."*

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 policymakers must consider in policy design.

Similarly, the design of policy experiments has begun to deliberately 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* 2016, 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 2016), but applying the results will always require the consideration of contextual differences for any trial on a scale that is less than global. Banerjee *et al*'s (2016b) proposal 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 Section 4 of this paper, but is again 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

[10] More precisely, the correlation of observables with unobservables may differ across contexts. A further limitation of sub-group analysis is of course the issue of limited power and the risk of false positives arising from multiple testing (Petticrew *et al* 2012), but such statistical issues are beyond the scope of this paper.

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 policymaker. 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 and unwieldy and time-intensive in practice (Marchal *et al* 2012), whereas the mechanism mapping approach in this study is intended to be practical and simple enough for policymakers to incorporate into routine processes of policy design.[11]

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 (Jagosh *et al* 2015, 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). While this is a model for other social sciences and policy fields to follow in some respects, there is also a need for a simpler and less resource-intensive analytical framework that is more accessible for other fields and policy contexts.

4 Mechanism Mapping

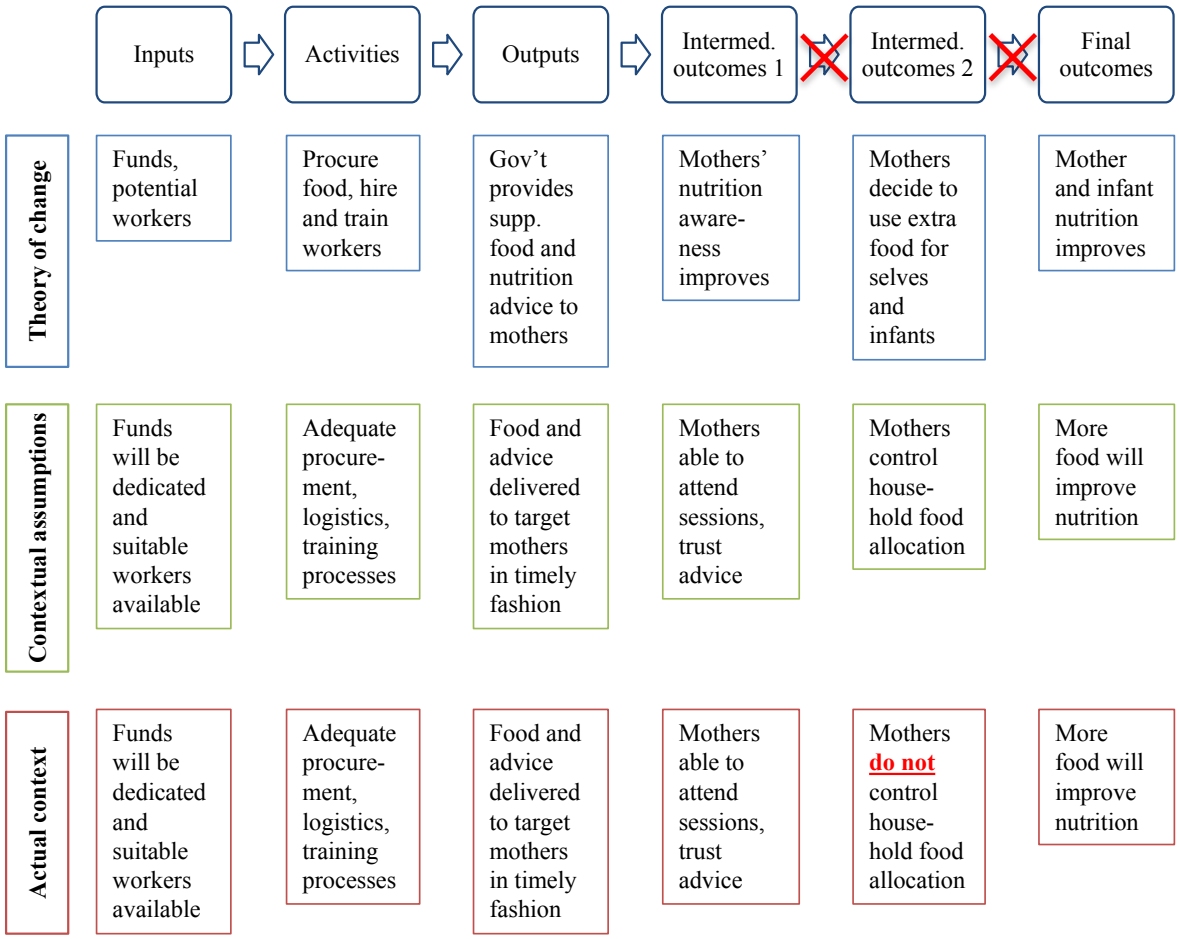
If external validity failures arise from interactions between a policy's theory of change and its context, then it follows that diagnosing such failures requires a way to examine theory of change alongside context. Furthermore, if contexts have numerous dimensions, for many of which hard data may not 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. I introduce this method by first presenting the approach itself, then giving an example. I then discuss some practical and conceptual issues as well as the role of empirical evidence, and finally suggest some ways in which mechanism mapping can be integrated into processes of policymaking (primarily) and evaluation (secondarily).

The first step of mechanism mapping is to lay out a policy's *theory of change*, or mechanism. As discussed in Section 2, 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 policymakers 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 Section 5 discusses later).

[11] Pawson and Manzano-Santaella (2012) criticize the quality of many evaluations that claim to be realist, particularly for their over-reliance on descriptive, qualitative data.

To illustrate the approach, consider a mechanism mapping for the Bangladesh Integrated Nutrition Programme (BINP) discussed above, which had an identical design to the World Bank's earlier TINP project in Tamil Nadu.[12] 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 activities for transforming inputs into outputs could include procuring the food, hiring and training workers, and conducting outreach to eligible mothers.

Figure 3: Theory of Change: Bangladesh Integrated Nutrition Programme



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

The contextual assumptions required for this theory of change to work are listed in the

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

second row of Figure 3. Sound implementation requires that government: 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 are 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, will 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 Figure 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.

For clarity, this article explains mechanism mapping using a simple, linear theory of change. Theories of change can of course be much more intricate, for example by showing complementarities or feedback mechanisms more explicitly or by mapping out multiple components of a multi-faceted program. The relevant point is that regardless of how simple or complicated a theory of change is, the same mechanism mapping approach - juxtaposing the causal linkages embedded in a policy against its contextual assumptions and the actual characteristics of the context - can be applied. Similarly, this article is agnostic on the precise definitional distinctions between the elements of the theory of change (what is the difference between an activity and an output? how many intermediate outcomes should there be between outputs and the final outcome? is there a conceptual difference between final outcomes and impact?). For the purposes of mechanism mapping, it matters less which particular "box" a causal link or contextual assumption is categorized under than that the relevant links, assumptions, and facts have been mapped out in a systematic fashion. Mechanism mapping can therefore be based on theories of change specified *ex ante* by the program designers, on research about similar policies implemented in other contexts, on the policymakers' own understanding, or some combination of the three.

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) simply by creating multiple mechanism maps, one for each outcome. The theory of change may be the same for each outcome or may differ slightly in emphasizing the aspects of the mechanism that are more salient. The key contextual assumptions and characteristics are likely to be different, however - if they were not (i.e. if the same mechanism and contextual assumptions were required for each outcome) then the outcomes could all be represented on the same mechanism map. The same procedure can also be used to assess the likelihood of *negative outcomes* or side effects of the policy (Bonell *et al* 2014), by placing these undesirable outcomes as the final outcome of the policy and assessing whether the policy mechanism and contextual characteristics and assumptions are likely to lead to these negative outcomes. Finally, mechanism mapping can also shed light on questions of *multiple competing mechanisms*, by constructing different mechanism maps and contextual assumption sets for each. These can then clarify which mechanism seems most plausible (during *ex ante* policy

design) or which specific empirical tests might discriminate among the competing mechanisms (during *ex post* policy evaluation).

One challenge is identifying which are the most salient contextual assumptions and characteristics to consider, since the high dimensionality of context makes it unfeasible to consider *all* aspects of context. Although this is ultimately a matter of judgment, two guidelines suggest themselves. First, 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. Second, important contextual factors specific to a particular policy are often suggested by the policy's theory of change. For instance, laying out BINP's theory of change makes it clear that decisionmaking over household food allocation is a key contextual assumption. Similarly, doing a mechanism map for the Tools of the Mind early childhood education program scale-up would have made it clear that assumptions were necessary about the fit of the program into the existing school day. While these two sets of factors may not always contain all relevant dimensions of context, they are likely to contain the most salient ones and thus serve as a good starting point for analysis.

As a procedural matter, one way to address the issue of which assumptions to consider would be to conduct mechanism mapping in a *nested* manner. The analyst would begin by identifying only the most salient steps in the policy's theory of change, along with accompanying contextual assumptions and characteristics. This would present 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 would have a better fit than others (as in Figure 3). From this top-level view, each of these links in the mechanism could 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 Figure 3 - breaking down the mechanism would serve 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 did *not* seem to fit well at the top-level stage - for instance in Figure 3's intermediate outcome of mothers and infants consuming the extra food themselves - going into more detail would help the analyst identify the root cause of the disjuncture. In the BINP case, this would be that TINP's synergy between training of 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.

Empirical evidence has an important role to play in mechanism mapping. Most obviously, the contextual characteristics in the crucial bottom row are questions to which empirical answers - or at least suggestive evidence - may well exist. A mechanism mapper could, for example, examine budget data and political context to shed light on resource availability, investigate the performance of the implementing agency's procurement processes, conduct a survey of eligible mothers' trust of the state and baseline level of nutritional knowledge, undertake (or read existing) qualitative research on household food allocation decisions in rural Bangladesh, and discuss with public health experts the prevalence of diseases that might inhibit infants from absorbing nutrients properly.[13] Similarly, impact evaluation results from other contexts and systematic reviews can enter into mechanism mapping via the contextual assumptions row,

[13]Bates and Glennerster (2017) present several excellent examples of using simple descriptive data to validate contextual assumptions.

as policymakers 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, etc. One could even imagine policymakers conducting quick and cheap "mechanism experiments" (Ludwig *et al* 2011) to validate each step of the theory of change prior to beginning full-scale implementation. 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 policymakers to apply all available evidence - from RCTs to administrative data to qualitative research to simple educated guesses - to their policy decisions.

While mechanism mapping is intended primarily as a tool for policymakers to use prospectively to predict the impact of a new policy, mechanism mapping is also of potential value to evaluators in two ways. 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 non-impact). 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 policymakers from other contexts. Second, prospective mechanism mapping (perhaps during trial design or in a pre-analysis plan) can also help evaluators design trials to 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. 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 - as the following section discusses.

Finally, the process of mechanism mapping is flexible enough that the same basic process can be undertaken either by an entire agency in systematic discussion with its stakeholders or by one analyst sitting at her desk for a short period, or any combination in between. Mechanism mapping seems especially well suited to participatory and collaborative policy design processes, since it is capable of integrating evidence of all varieties and the clients or beneficiaries of a policy may well be more aware of salient contextual assumptions or characteristics than policymakers. This contrasts with the top-down and largely technocratic approach to using quantitative evidence alone to inform policy choices.

5 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 often necessary to make 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 substantial, 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 successul in another context be adapted, since adaptations risk changing aspects of the policy that make it

effective? While the fidelity versus adaptation debate is well-worn in social policy (Castro *et al* 2010), mechanism mapping can provide a new perspective and a structured framework for analyzing these two questions.

With respect to the first question, 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, Figure 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 does identify which aspect of the policy is problematic, and why. Similarly, Figure 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.

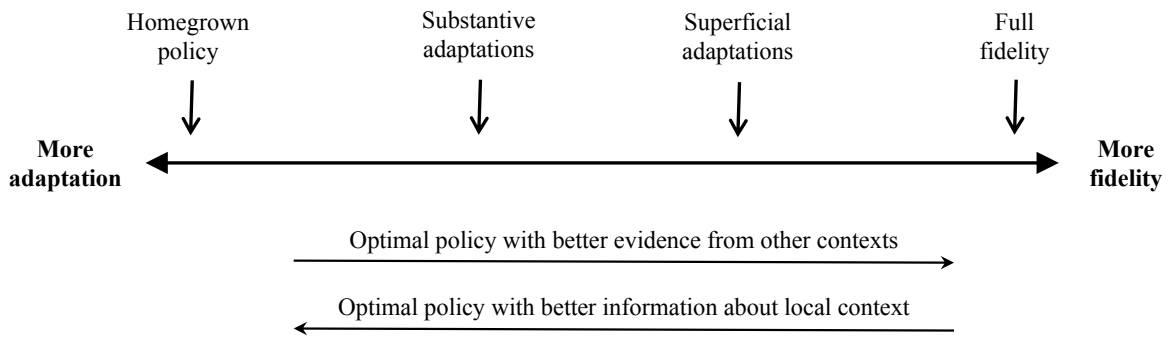
The same framework can also be used to identify adaptations that might be necessary - or potentially detrimental - in scaling up a policy that was successful in a small-scale trial. Most obviously, contextual assumptions that held in the trial may not hold when implementing at scale. For example, local government agents may require different incentive or monitoring schemes than non-governmental agents in order to elicit similar effort levels (Cameron and Shah 2017), or incentive schemes that were effective for non-governmental implementers in small trials may be unfeasible for political economy reasons when implemented at scale (Bold *et al* 2016), 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 provides a framework for thinking through their potential consequences and thus possible mitigating measures.

Mechanism mapping complements another recent innovation: adaptive policymaking, which views policy design and evaluation as an iterative process of experimentation with tight feedback loops (Pritchett *et al* 2012, World Bank 2015). As a policy diagnostic tool that focuses on mechanism, context, and their interaction, mechanism mapping seems especially promising to integrate with adaptive policymaking processes. Mechanism mapping could help connect experimentation to a more precise diagnosis of the barriers to the effectiveness of previous iterations of a policy, thus adding precision to the experimental search process. Specifically, policymakers could use the weakest assumption link in the theory of change (as identified by mechanism mapping) as the starting point for adaptive experimentation. Similarly, since mechanism mapping lines up closely with the policy's theory of change, and monitoring and data collection strategies are typically based a policy's theory of change or logframe, the data that organizations generate during adaptive policymaking processes often closely align 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 policymakers 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 context, making it unclear how quick policymakers 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?

Figure 4: The Fidelity-Adaptation Spectrum



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 policymakers' 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.

First, to the extent that available impact evaluation evidence on the policy is strong, consistent, and from similar contexts, policymakers 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) 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 policymaker's information about the local context - whether in the form of formal evidence and data, or simply familiarity and tacit knowledge - the more a policymaker 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 policymaker: *ceteris paribus*, an expatriate donor official should generally make fewer adaptations to a transported policy than a policymaker from the country with deeper local knowledge.

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 policymaker at her desk. 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 policymakers do not have time to gather extensive data on actual contextual characteristics and are thus forced to rely on their own knowledge, policymakers should weight evidence from elsewhere relatively more heavily and thus usually make fewer adaptations.

A small but growing body of research studies this issue empirically, by comparing the effectiveness of various policies or programs according to whether they were newly designed, transported but adapted, or transported without adaptation. The results are mixed (Castro *et al* 2010). For example, Hasson *et al* (2014) compare 307 German and Swedish social interventions and find that novel and adapted programs are more effective than non-adapted programs, while Leijten *et al* (2016) find no difference on average between homegrown and transported parenting interventions across a range of countries, and Gardner *et al* (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 for a policy is that it is unclear exactly what changes in context the adaptations were responding to, nor how appropriate the adaptations work. As with policies themselves, knowing the average effectiveness of adaptations is less useful for policymakers than knowing which adaptations will be necessary and effective in their specific context.

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 systematic reviews that are able to trace mechanisms - can be especially useful. Good policy adaptation thus stems not just from contextual knowledge, but from its combination with rigorous evidence and professional judgment.

6 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 *et al* 2016, 4). 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 policymakers to find ways to make evidence-based policymaking more responsive to the particularities of specific contexts. While systematic reviews and impact evaluations provide excellent starting points for doctors and policymakers 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 in economics has focused largely on the *generalizability* of a policy *from* a specific context, the relevant question for policymakers 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 tool to help policymakers assess the fit of evidence-based policies with their own contexts. Mechanism mapping is a flexible and conceptually simple diagnostic tool, which can be conducted either as a quick desk exercise by a lone policymaker or through a rigorous and collaborative process of evidence gathering, stakeholder participation, and experimentation. Similarly, the same basic procedure can be applied to simple policies with a linear theory of change and only one outcome of interest, or to complex policies with numerous potential outcomes and feedback loops. This diagnostic process also feeds directly into the identification of potential policy adaptations, by identifying the specific features of a policy that are likely to be problematic in the new context.

Of course, mechanism mapping as a tool for policymakers and evaluators also has important limitations. Most obviously, mechanism mapping yields only approximate directional predictions of policy effectiveness rather than precise statistical point estimates of effect sizes and confidence intervals. In some cases mechanism mapping may not even generate a clear overall prediction for whether a policy is likely to be more or less effective than it had been in a previous context, since different contextual differences may shift effectiveness in different directions (i.e., be opposite-signed). Without a formal model to make explicit the linkages between different components of a policy's theory of change and their interactions with context, there is no precise way to combine these directional predictions into an overall prediction. 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 policymakers from also making use of more precise quantitative tools.

Finally, mechanism mapping should be understood as a tool to help policymakers structure their judgment about policy transportation and adaptation, not a scientific procedure for determining whether or not a policy will work. It relies on policymakers' judgment in the identification of salient contextual assumptions, in deciding which potential interactions are significant enough to warrant adaptation, and in designing appropriate adaptations. While this may be frustrating for social scientists and evaluators seeking to give unequivocal guidelines for policymaking, some degree of judgment is inevitable in policy. By making causal links and contextual assumptions explicit, the more humble aim of mechanism mapping is to structure and improve policymakers' own judgment in the pursuit of better use of evidence in policymaking.

References

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *Quarterly Journal of Economics*: 1117-1165.

Angrist, Noam. 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, 26 April.

Angrist, Joshua, and Ivan Fernandez-Val. 2010. "Extrapolate-ing: External Validity and Overidentification in the LATE Framework." NBER Working Paper 16566, December.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. 2016a. "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application." Mimeo, September.

Banerjee, Abhijit, Sylvain Chassang, and Erik Snowberg. 2016b. "Decision Theoretic Approaches to Experiment Design and External Validity." Mimeo, July.

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

Bates, Mary Ann, and Rachel Glennerster. 2017. "The Generalizability Puzzle." *Stanford So-*

cial Innovation Review, Summer, ⟨https://ssir.org/articles/entry/the_generalizability_puzzle⟩, accessed 22 June, 2017.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur. 2016. "Experimental Evidence on Scaling Up Education Reforms in Kenya." Mimeo, November.

Bonell, Chris, Farah Jamal, G.J. Melendez-Torres, and Steven Cummins. 2014. "'Dark Logic': Theorising the Harmful Consequences of Public Health Interventions." *Journal of Epidemiology and Community Health* 69: 95-98.

Cameron, Lisa, and Manisha Shah. 2017. "Scaling Up Sanitation: Evidence from an RCT in Indonesia." IZA Discussion Paper 10619, March.

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

Castro, Felipe González, Manual Barrera Jr., and Lori 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, Garret, and Edward Miguel. 2016. "Transparency, Reproducibility, and the Credibility of Economics Research." NBER Working Paper 22989.

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. 2013. "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Random Experiment." Mimeo, November.

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, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48: 424-55.

Deaton, Angus, and Nancy Cartwright. 2016. "Understanding and Misunderstanding Randomized Controlled Trials." NBER Working Paper 22595, September.

Diamond, Adele, W. Steven Barnett, Jessica Thomas, and Sarah Munro. 2007. "Preschool Program Improves Cognitive Control." *Science* 318(5855): 1387-88.

DFID. 2012. "Review of the Use of 'Theory of Change' in International Development." Review Report, April.

Evans, David, and Anna 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.

Farran, Dale, and Sandra Jo Wilson. 2014. "Achievement and Self-Regulation in Pre-Kindergarten Classrooms: Effects of the Tools of the Mind Curriculum." Mimeo, Peabody Institute, July.

Gardner, Frances, and Paul Montgomery. 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*: 1-14.

Gechter, Michael. 2016. "Generalizing the Results from Social Experiments: Theory and Evidence from Mexico and India." Mimeo, December.

Greenhalgh, Trisha, Fraser Macfarlane, Liz Steed, and Robert Walton. 2016. "What Works for Whom in Pharmacist-Led Smoking Cessation Support: Realist Review." *BMC Medicine* 14:

209-24.

Hasson, Henna, Knut Sundell, Andreas Beelmann, and Ulrica 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, Penelope. 2015. "Lessons from Complex Interventions to Improve Health." *Annual Review of Public Health* 36: 307-23.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33: 61-135.

Jagosh, Justin, Paula L. Bush, Jon Salsberg, Ann C. Macaulay, Trish Greenhalgh, Geoff Wong, Margaret Cargo, Lawrence W. Green, Carol P. Herbert, and Pierre Pluye. 2015. "A Realist Evaluation of Community-Based Participatory Research: Partnership Synergy, Trust Building, and Related Ripple Effects." *BMC Public Health* 15: 725-35.

Kowalski, Amanda E. 2016. *How to Examine External Validity Within an Experiment.* Mimeo, August.

Leijten, Patty, G.J. Melendez-Torres, Wendy Knerr, and Frances 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, Laura C. 2017. "Generalizing about Public Health Interventions: A Mixed-Methods Approach to External Validity." *Annual Review of Public Health* 38: 371-91.

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

Ludwig, Jens, Jeffrey Kling, and Sendhil Mullainathan. 2011. "Mechanism Experiments and Policy Evaluations." *Journal of Economic Perspectives* 25(3): 17-38.

Marchal, Bruno, Sara van Belle, Josefien van Olmen, Tom Hoeré, and Guy 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, Collen, and Collet Dandara. 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): 1-6.

Miguel, Edward, and Michael 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 Evaluationg of Complex Interventions: Medical Research Council Guidance." *British Medical Journal*: 350.

Muralidharan, Karthik, and Paul Niehaus. 2016. "Experimentation at Scale." Mimeo, September.

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

Nyakutira, Christopher, Daniel Röshammar, Emmanuel Chigutsa, Prosper Chonzi, Michael Ashton, Charles Nhachi, and Collen Masimirembwa. 2008. "High Prevlance 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: 357-65.

Pawson, Ray, and Ana Manzano-Santaella. 2012. "A Realist Diagnostic Workshop." *Evaluation* 18(2): 176-91.

Pawson, Ray, and Nicholas Tilley. 1997. *Realistic Evaluation*. London: SAGE Publications.

Petticrew, Mark, Peter Tugwell, Elizabeth Kristjansson, Sandy Oliver, Erin Ueffing, and Vivian 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, Lant, Salimah Samji, and Jeffrey Hammer. 2013. "It's All About MeE: Using Structured Experiential Learning ("e") to Crawl the Design Space." Center for Global Development Working Paper 322, April.

Pritchett, Lant, and Justin Sandefur. 2015. "Learning from Experiments when Context Matters." *American Economic Review: Papers and Proceedings* 105(5): 471-75.

Rajman, Iris, Laura Knapp, Thomas Morgan, and Collen Masimirembwa. 2017. "African Genetic Diversity: Implications for Cytochrome P450-mediated Drug Metabolism and Drug Development." *EBioMedicine* 17: 67-74.

Rodrik, Dani. 2009. "The New Development Economics: We Shall Experiment, but How Shall We Learn?" In Cohen, Jessica, and William Easterly (eds), *What Works in Development?: Thinking Big and Thinking Small*, Washington DC, Brookings Institution Press: 24-47.

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, Eva. 2016. "How Much Can We Generalize from Impact Evaluations?" Mimeo, May.

What Works Network. 2014. "What Works? Evidence for Decision Makers." Presentation, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/378038/What_works_evidence_for_decision_makers.pdf, accessed 19 May, 2017.

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

Wong, Geoff, Gill Westhorp, Ana Manzano, Joanne Greenhalgh, Justin Jagosh, and Trish 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, December.

World Bank. 2005b. "Project Performance Assessment Report: Bangladesh Integrated Nutrition Project." Report No. 32563, 13 June.

World Bank. 2015. "World Development Report 2015: Mind, Society, and Behavior." Chapter 11.