

2016

Rapid Evidence Maps as Decision Tools for Evidence-based Policy

Mark Engelbert
SIT Graduate Institute

Follow this and additional works at: <https://digitalcollections.sit.edu/capstones>

 Part of the [Policy Design, Analysis, and Evaluation Commons](#)

Recommended Citation

Engelbert, Mark, "Rapid Evidence Maps as Decision Tools for Evidence-based Policy" (2016). *Capstone Collection*. 2942.
<https://digitalcollections.sit.edu/capstones/2942>

This Thesis (Open Access) is brought to you for free and open access by the SIT Graduate Institute at SIT Digital Collections. It has been accepted for inclusion in Capstone Collection by an authorized administrator of SIT Digital Collections. For more information, please contact digitalcollections@sit.edu.

Rapid Evidence Maps as Decision Tools for Evidence-based Policy

Mark Engelbert

A capstone paper submitted in partial fulfillment of the requirements for a Master of Arts in Sustainable Development at SIT Graduate Institute, DC Center in Washington, DC, USA

July 29, 2016

Advisor: Amy Jersild

Consent to Use

I hereby grant permission for World Learning to publish my capstone on its websites and in any of its digital/electronic collections, and to reproduce and transmit my CAPSTONE ELECTRONICALLY. I understand that World Learning's websites and digital collections are publicly available via the Internet. I agree that World Learning is NOT responsible for any unauthorized use of my capstone by any third party who might access it on the Internet or otherwise.

Student Name: Mark Engelbert

Date: July 29, 2016

Table of Contents

Abstract.....	1
1. Introduction.....	2
2. What is Evidence for Policymaking?	3
2.1 The Nature of Evidence	3
2.1.1 The Utility of Causal Knowledge.....	3
2.1.2 Causal Knowledge in Social Systems	5
2.1.3 External Validity, Generalizability, and Policy	8
2.2 Using Evidence for Policy	11
3. Review of Existing Evidence Tools.....	12
3.1 Purposes of Evidence Tools	12
3.2 Types of Evidence Tools	14
3.2.1 Literature Reviews/Narrative Reviews	14
3.2.2 Systematic Reviews.....	15
3.2.3 Rapid Evidence Assessments/Rapid Reviews.....	16
3.2.4 Evidence Gap Maps	17
4. The Rapid Evidence Map.....	21
4.1 What are REMs for?	22
4.2 Advantages of the REM Approach.....	23
4.3 Rapid Evidence Maps: Process and Product	25
5. Potential Use Case: Social Inclusion of Indigenous Groups in Peru	27
5.1 State of Indigenous Populations in Perú.....	28
5.2 MIDIS: Mission, Priorities, and Strategic Plan	28
5.3 Proposed REM Framework, Search Strategy, and Screening Criteria	30
6. Conclusion	32
References	33
Appendix: Indicators and Goals of Include to Grow	38

Abstract

The movement to make policy more evidence-based is gaining increasing influence in the field of international development. I argue that, properly understood and executed, evidence-based policy has the promise to make development efforts more effective. However, if policy decisions are to be evidence-based, there must be tools available to policymakers that provide evidence in usable formats. The research community has developed a variety of such tools, but there are some decision-making contexts for which none of the currently-available tools are ideally suited. This paper describes a new type of evidence tool, the rapid evidence map (REM), that provides users with an overview of the evidence base on a particular topic, while adopting a sufficiently narrow scope to allow for rapid completion and applicability to specific policy decisions. In this way, the REM seeks to accommodate both decision-makers' need for evidence and their real-world time constraints. After presenting the concept of the REM, I outline a pilot REM project for a Peruvian government agency seeking to improve its services for indigenous populations.

1. Introduction

The idea that policy and practice in international development should be “evidence-based” is increasingly prominent. This had led to both reflection and controversy within the development community about what “evidence” is, how it is generated, and what it means for decisions to be based on it. This capstone project explores these questions and defends the idea that large-scale, rigorous impact evaluations provide indispensable evidence about whether and how development programs achieve their intended effects. I also argue that if policy-makers and practitioners are to base their decisions on such evidence, they must have access to tools that collect, synthesize, and assess the evidence base. While a variety of such tools have emerged in the research world to address a variety of audiences and serve a variety of purposes, I outline the need for and an approach to a new type of evidence tool, the rapid evidence map (REM). The REM provides decision-makers with a user-friendly tool that offers both a high-level overview of evidence relevant to a policy question and quick access to available sources of evidence.

This work begins, in Section 2, with a discussion of what “evidence-based policy” means and why it is important. In the course of this discussion, I will argue that while evidence encompasses many kinds of information, certain rigorous methods for evaluation and research—namely, experimental and quasi-experimental methods—provide a distinctively useful kind of action-guiding information. This exploration of the nature of evidence-based policy will provide a foundation for my discussion of how certain tools can help decision-makers to make use of this sort of evidence. Section 3 provides a review of available tools for synthesizing, assessing, and mapping evidence, along with a discussion of their strengths and limitations. Section 4 describes the REM approach and how it fulfills a distinct need in the space of tools for evidence-based policy. Finally, in Section 5 I describe how a pilot REM project will be used to support the Peruvian Government’s Ministry of Social Development & Inclusion in identifying evidence-

based strategies for serving the country's indigenous populations, and detail the proposed framework and approach to carrying out this project.

2. What is Evidence for Policymaking?

2.1 The Nature of Evidence

A first step to understanding how evidence can support effective policy decisions is understanding what constitutes evidence and what it is about evidence that makes for better decision-making. I begin, therefore, by presenting the basic case for thinking of evidence as information provided by particular methods for research and evaluation—namely, those methods best suited to identifying genuine causal relationships in social settings. These arguments will likely be familiar to readers, but they are worth reviewing because they are essential to understanding what is at stake in recent discussions—and controversy—about evidence-based policy in development.

2.1.1 The Utility of Causal Knowledge

The push towards evidence-based policy is often framed as the idea that we should implement “what works.” But what, exactly, does it mean to say that something “works”? To say that something works is to say that it achieves, at least in part, its intended effect—in other words, something works when it *causes* a desired outcome. A common way of understanding causal relationships is in terms of *counterfactuals*: to say that Event *c* caused Event *e* is to say that if *c* had not occurred, then *e* would not have occurred either. To say that a tutoring program caused an increase in achievement is to say that had the tutoring program not existed, achievement would have been lower. This is often expressed by saying that the rise in

achievement *depends counterfactually* on the existence of the tutoring program. More colloquially, the tutoring program *makes a difference* to students' achievement.

As statisticians frequently remind us, causation is distinct from correlation, though the two often look very similar. But distinguishing genuinely causal relationships from mere correlations is not only an academic exercise; rather it essential for sound decision-making. Because causes *make a difference* to outcomes, while correlated factors do not, our actions can influence outcomes by manipulating causal factors but not by manipulating correlated ones (Cartwright, 1979; Northcott, 2006; Woodward, 2003). Determining which factors are causally related to outcomes we care about, then, is important because of its *practical value* in guiding our actions.

For example, suppose we observe that people in an area who take out microloans have higher incomes than those who do not. This relationship between microfinance and income may be causal or correlational, and the difference is in the counterfactual: if no one had had access to microfinance, would we still have observed the same patterns of income? Microfinance might be part of a causal chain linked to higher incomes: for example, with access to microfinance, people are able to start businesses that then generate higher incomes. This is to say that the use of microfinance is the *determining factor* for why some people have higher incomes than others. In contrast, the relationship may be correlational: perhaps the characteristics or circumstances of those who used microloans differed in some way from the circumstances of those who did not, and these other factors made the difference (e.g., those who took out loans had more formal education, which would have led to their earning higher incomes anyway).

The reason this matters for policymaking is that if the relationship is causal, we can influence incomes through the use of microfinance—i.e., we can increase people's incomes by

getting them to use microloans. If the relationship is correlational, however (e.g., if use of microfinance is correlated with education, which causes higher income), then manipulating the availability and use of microfinance will not bring about our goal of raising incomes. As Nancy Cartwright puts it, distinguishing between causal and merely correlational relationships is “needed to ground the distinction between effective strategies and ineffective ones” (1979, p. 420).

2.1.2 Causal Knowledge in Social Systems

One way of understanding the so-called “scientific method” is that it is a process designed to answer causal questions by mimicking counterfactuals. While we can never truly know what would have happened in some counterfactual scenario, we can, in principle, make reliable *inferences* about counterfactuals (and, therefore, causes) by observing a large number of outcomes under certain conditions. This is the idea behind the notion of an *experiment*: we observe a system multiple times while varying some aspects and holding others fixed, and look to see which variables make a difference and which do not.

However, this ideal of controlled investigation is, famously, rarely possible when it comes to investigating the kinds of *social* systems that are relevant to development policy and practice. For practical and ethical reasons, it is generally not possible to instantiate tightly controlled circumstances in which to methodically observe the real-world behavior of actual human beings and groups. However, several methods exist that aim to establish causal relationships even in the messy world of social phenomena. First, there is the randomized controlled trial (RCT) (Duflo, Glennerster, & Kremer, 2008). RCTs are called “experimental” studies because they approximate (in principle, at least) the nature of classical scientific experiments. In an RCT, the study population is randomly divided into two or more groups, with

each group exposed to a different set of conditions. In the simplest case, an RCT consists of two conditions: a treatment group that is exposed to an intervention and a control group that does not.¹ If assignment to treatment and control conditions is truly randomized, then we can be confident that the baseline characteristics of the two groups will be similar in all respects save the one we are investigating—i.e., exposure to the intervention. This equivalence between the treatment and control groups is what allows us to assess counterfactual conditions and draw causal inferences: if the treatment and control groups differ only in whether they are exposed to the intervention, then the trajectory of the control group provides a sort of simulation of what *would have happened* to the treatment group in the absence of the intervention. If we find that the two groups differ after the program, then our best guess is that it is the intervention that caused the difference; if the two groups do not differ despite some having received the program and others not, then this evaluation has not found evidence that implementing the program helped to achieve a desired outcome.

However, RCTs are notoriously difficult and expensive to conduct, and in many circumstances would be unethical even if feasible. Therefore, even if RCTs do offer the best protection against certain types of systematic error, they often do not represent the most appropriate method for evaluating the effects of a policy or intervention. In such cases, evaluators will often employ one or more *quasi-experimental* methods (Shadish, Cook, & Campbell, 2002). For purposes of the present discussion, the details of these methods are less important than the general structure they all share, which is attempting to mimic the effects of

¹ Alternative RCT designs include delivering alternative versions of an intervention to the different groups—e.g., tutoring provided by teachers versus tutoring provided by peers—or exposing all groups to an intervention but exposing some groups to *additional* interventions—e.g., all students in a sample receive tutoring but some receive tutoring plus free school supplies. In addition, note that the “intervention” need not necessarily be access to a program or service; for example, an intervention might be some form of encouragement to take advantage of some program to which the entire population has access (Duflo, Glennerster, & Kremer, 2008).

randomization by constructing treatment and control groups that are *equivalent*, even when randomization has not occurred. In other words, these methods, like RCTs, attempt to assess outcomes against a counterfactual.² In particular, quasi-experimental approaches are considered much more effective at controlling for unobserved factors—and hence, provide a much stronger basis for causal inference—than other popular social science methods like multivariate linear regression, which critics (e.g., Schrodt, 2014) argue is subject to severe limitations. Nevertheless, many proponents of RCTs argue that quasi-experimental methods are still vulnerable to certain forms of systematic error—which RCTs avoid—and therefore provide less reliable causal information than true experimental designs (Duflo et al., 2008; Karlan, Goldberg, & Copestake, 2009).³

When proponents of evidence-based policy advocate using “rigorous” approaches to impact evaluation and relying on “high-quality evidence” for policy decisions, they typically have in mind RCTs and quasi-experimental methods (and sometimes only the former). The use of such methods in evaluation and in shaping development policy has been the subject of both great enthusiasm and great controversy in recent years. In particular, there is considerable debate about the relative strength of different evaluation designs in allowing us to identify causal impacts of development policies and programs on development outcomes—i.e., to distinguish causal from correlational relationships (Cook, Shadish, & Wong, 2005, 2008; Deaton, 2010; Glazerman, Levy, & Myers, 2002; Heckman, 1991). However, most would be willing to grant

² There is difference of opinion regarding exactly which methods qualify as quasi-experimental, but in general the list includes differences-in-differences, regression discontinuity, instrumental variable estimation, and propensity score matching (for discussion of individual methods and the general logic of quasi-experimental approaches, see Austin, 2011; Bor, Moscoe, Mutevedzi, Newell, & Bärnighausen, 2014; Imbens & Wooldridge, 2008; Ravallion, 2007; Shadish, Cook, & Campbell, 2002).

³ This is not to say that these RCT proponents believe that all evaluations should be randomized; rather, they suggest (a) that one factor that evaluators should consider in choosing an evaluation design is the degree of confidence with which a method will allow us to draw causal inferences, and (b) that, all else equal, RCTs license greater confidence than other methods.

that *in principle*, rigorous designs provide better evidence about causes than do other approaches. The more pressing concern is one that goes to the heart of questions about the evidence-based policy movement: do these rigorous methods represent a practical approach to generating the kind of evidence that is useful for designing and implementing effective policies and programs in the real world? (Cartwright, 2007; Jones, 2009; Rodrik, 2008).

2.1.3 External Validity, Generalizability, and Policy

Perhaps the most prominent concern about the usefulness of rigorous evaluation designs is whether we can generalize results from one context to another. The most rigorously designed and flawlessly implemented RCT can perhaps establish that a particular program did or did not cause a change in some outcome for a particular population. But this information—while useful—falls far short of providing a definitive answer to the kinds of questions that confront policymakers: *what will happen if we implement this program here?* Cartwright and Hardie (2012) describe this as the problem of getting from “it worked there” to “it will work here.” This is the issue of “external validity” for research findings, that is, the extent to which the results from a study will apply to different populations in different contexts.⁴

How much can we generalize evidence from one setting to another, and how can we tell the difference between interventions that will work similarly in a new context, and ones that won't? These questions are partly conceptual and partly empirical. Conceptually, we would expect the generalizability of results to depend on how similar the contexts are in terms of the factors that play a causal role in determining how a program functions. And we would expect different degrees of similarity for different kinds of causal factors. When it comes to, say,

⁴ External validity is contrasted with internal validity, which refers to whether the results of an analysis accurately identify actual causal relationships *in the population studied* (i.e., in contrast to mistaking correlations for causes).

evidence about the effects of a drug on a disease, we might expect high generalizability. Given the basic biological similarities shared by all human beings, we can expect a certain level of consistency in how a particular chemical will interact with the components of a physiological system.⁵ The intuition behind using rigorous evaluations to drive development policy worldwide is that these evaluations are tapping into universal principles of human behavior that will determine how programs function in any setting. But when both the populations targeted and the surrounding institutional and cultural factors are highly dissimilar across contexts, it is easy to imagine that the same type of program will have very different effects when implemented in different times and places.

Empirical analyses of the research literature corroborate this intuition: as one might expect, when published findings of development impact evaluations are compared with fields like medicine, we see significantly more heterogeneity in the former (Vivalt, 2015). In addition, a number of recent papers have investigated the generalizability of particular studies, sometimes suggesting low external validity (Bold, Kimenyi, Mwabu, Ng'ang'a, & Sandefur, 2013; Dehejia, Pop-Eleches, & Samii, 2015; Gechter, 2015).

Thus, in practice, policymakers seeking the most evidence-based approaches will not find a tidy list of rigorous impact evaluations that conveniently sort interventions into “works” and “does not work” bins. A recent analysis by Pritchett and Sandefur (2013) highlights the challenges that confront real-world policymakers who turn to the research literature in search of the most “evidence-based” approaches. The authors consider several prominent studies in the literature on how class sizes and private school vouchers affect student achievement. These studies are a mixture of experimental and quasi-experimental designs, and were conducted in a

⁵ Of course, this kind of consistency is by no means guaranteed, and there is substantial individual- and population-level variation in the effects of many pharmaceutical and other medical interventions.

wide array of country contexts, including Colombia, India, Israel, the United States, and Kenya. If a group of policymakers wished to use this literature to estimate the likely effects in their country of reducing class sizes or providing school vouchers, how might they proceed? Pritchett and Sandefur ask which of the following approaches can be expected to provide more accurate estimates: (a) privileging the most methodologically rigorous (i.e., experimental) evidence gathered from very different contexts, or (b) relying primarily on evidence gathered in similar contexts but that employs methods somewhat more prone to error. Using econometric methods to assess the accuracy of the estimates given by these different approaches, Pritchett and Sandefur find that relying on the “less rigorous” quasi-experimental methods gives more accurate predictions—i.e., leads to better policy decisions—than the more rigorous but less context-specific evidence. They conclude:

These findings imply that the common practice of ranking evidence by its level of “rigor,” without respect to context, may produce misleading policy recommendations. In principle, this possibility is fairly obvious and well known, yet in practice appears to be heavily discounted in both academic and policy discussions... Given the current state of the field, the evidence here suggests that policymakers would do well to prioritize external validity over internal validity concerns when surveying the development economics literature. (Pritchett & Sandefur, 2013, p. 186)

In sum, the issue of generalizability establishes an essential lesson for evidence-based policymaking. Although the ability of rigorous evaluations to discern causal impacts is valuable, it would be a mistake to think that the strength of the evidence available for policymaking is determined solely by the degree to which existing evaluations approximate the “ideal” of randomized experiments. Rather, a decision-maker’s evidence about how a program will work in a particular setting comprises a wide array of considerations. The effects of similar programs in other contexts—as established by rigorous causal analyses—are but one component of the

evidence. As a number of writers on evidence-based policy have discussed, developing and testing robust *theories of change* is one of the most useful tools for identifying the particular factors that influence how people respond to a program in context, and for adapting program components to suit the needs of a particular population (Brown, 2016; Cartwright & Hardie, 2012; Gaarder, Glassman, & Todd, 2010).

2.2 Using Evidence for Policy

So far we have established that knowledge about the *causes* at work in social systems is indispensable for choosing actions that influence results, and that experimental and quasi-experimental methods of evaluation give us unique insight into the causal forces at work. However, we have also found that even if policymakers can be confident about what the causal effects of a program were in one context, they must draw on a much wider body of information in order to draw prudent conclusions about what approaches will be most effective in bringing about desired changes in the local context. What then can we say about the relevance of rigorous evaluation methods to designing and implementing policies that promote inclusive economic and social development?

I would argue that the most sensible conclusion is that while we should acknowledge that “evidence for policymaking” is a multifarious concept that extends well beyond the results of rigorous impact evaluations, such evaluations nevertheless provide an indispensable *kind* of evidence. As attested by many decades of unsatisfactory progress in global development, identifying promising approaches and predicting what will work are formidable challenges. Given the difficulty of these decision problems, knowing whether a program has worked elsewhere is a good starting place. And knowing whether something actually *has* worked

elsewhere requires distinguishing genuine causal links from deceptive correlations or impressionistic assessments—that is, it requires rigorous evaluation methods.

Enacting evidence-based policy, therefore, requires that policymakers and practitioners use—among many other sources—the results of such evaluations to guide decisions. Throughout the remainder of this paper, I explore ways that evaluators and other researchers can support this component of evidence-based policymaking, in particular by creating tools that make this type of evidence readily accessible to and usable by decision-makers. For simplicity’s sake, I will hereinafter use the term “evidence” in the narrow sense of information drawn from experimental and quasi-experimental impact evaluations. However, this should not be taken to imply that this is the only defensible understanding of what evidence is, or that only quantitatively rigorous impact evaluations provide useful, action-guiding information.

3. Review of Existing Evidence Tools

3.1 Purposes of Evidence Tools

Evidence tools are products that aid decision-makers in understanding the evidence base and leveraging it to improve development practice. Evidence tools exist to serve a variety of purposes, and particular types of tools may serve some purposes but not others. First, evidence tools *assemble* evidence. That is, they find evidence that is stored in various places (academic journals, the websites of organizations, working papers or other informal reports from researchers and practitioners), and catalog that evidence in one place. Thus, even absent any substantive analysis of the collected evidence, the list of sources compiled in an evidence tool serves a useful function in informing others where to look to find primary and secondary sources of evidence. Any evidence tool that provides a list of sources automatically fulfills—at least in a minimal way—this function.

Second, evidence tools can *synthesize* evidence. To synthesize evidence is to reduce the complexity inherent in the collected body of sources by grouping or categorizing them, extracting common themes, and identifying key lessons. Systematic reviews and meta-analyses commonly serve this function, with meta-analyses in particular seeking to ascertain and report, in a quantitatively rigorous way, an aggregated estimate of the effects found in particular studies.

Third, evidence tools *assess* evidence in terms of its overall reliability and its relevance to particular research or policy questions. Following recommendations in the Cochrane Collaboration's *Handbook for Systematic Reviews of Interventions* (Higgins & Green, 2008), assessing evidence often takes the form of evaluating the *risk of bias* in individual studies based on research design or other features of the study, such as the rate of attrition among participants.⁶ Evidence assessments can be particularly useful for decision-making, since they often categorize different courses of actions according to how strongly they are supported by evidence. An example is the *Steady, Ready, Go!* framework, which translates evidence assessments into clear policy recommendations (e.g., “Implement widely but continue to evaluate”; “More research and development still needed”) (Mavedzenge, Doyle, & Ross, 2010, pp. 4, 11).

Finally, evidence tools can *map* evidence. This involves categorizing evidence within a pre-established framework in a way that captures what is known about the evidence base. Importantly, *mapping* evidence is distinct from *synthesizing* evidence—while a particular tool can both map and synthesize evidence, the mapping process does not involve the kind of analysis characteristic of the synthesis process. Rather, mapping provides information about the *state of the evidence base*, while synthesis provides information about *what the evidence says* (this

⁶ Note that the term *bias* in this context does imply that the researchers purposefully skewed results to obtain a desired outcome. Rather, *bias* simply refers to any systematic error arising from methodological limitations of a study (limitations which may often be appropriate given the context of the evaluation).

distinction will be important as we consider the strengths and limitations of various evidence tools in the sections that follow).

3.2 Types of Evidence Tools

There now exist a wide variety of types of evidence tools, each addressing different audiences and making different tradeoffs among the various purposes that evidence tools can fulfill. A complete review of such tools is beyond the scope of this paper (see Grant & Booth, 2009 for such a review), but I will here discuss the tools most relevant to understanding the motivations for and methods of the new evidence tool explored in this capstone, the rapid evidence map.

3.2.1 Literature Reviews/Narrative Reviews

The “traditional” type of literature review is a narrative review. In this type of review, the author(s) identify components of the literature that they find most helpful or relevant to their research questions, and describe, in narrative form, the aspects of these works that bear on the question at hand. In completing a narrative review, authors generally have wide latitude to decide which sources from the literature to include, and which aspects of these sources to discuss and emphasize.

Narrative reviews can provide expansive coverage of the literature and can perform valuable analysis and synthesis. However, when it comes to serving the functions of an evidence tool, narrative reviews have several drawbacks. Chief among these is that authors of narrative reviews typically do not describe how they decided which parts of the literature to include and discuss in their reviews. Consequently, the reader has only limited information with which to assess how well the review achieves the goals of *comprehensiveness* and *freedom from bias*.

Were the sources included in the narrative review those that the authors came across first in their search? Were they the sources the authors liked best? Were they selected only from the handful of journals with which the authors are most familiar? Were they chosen because they fit the argument the authors wished to make? To the extent that any of these are true, a narrative review constitutes a very limited tool in terms of providing an accurate and complete picture of what scholarship on a topic has uncovered.⁷

3.2.2 Systematic Reviews

In light of the limitations to narrative reviews discussed above, scholars have increasingly promoted the *systematic review* (SR) as a method better suited to the needs of evidence-based decision-making (Boaz, Ashby, & Young, 2002; Gasteen, 2010; Snilstveit, 2012). SRs aim to be reliable tools for evidence-based policy by conducting exhaustive literature searches, including *all* and *only* those studies that meet pre-specified criteria, and documenting and reporting each stage of the process. In this way, well-executed systematic reviews are “transparent, rigorous and replicable” (Gasteen, 2010, p. 1). In addition to capturing as much of the research literature as possible, SRs also aim to *synthesize* the existing evidence base in order to identify questions on which the evidence is particularly strong or particularly weak, and to draw lessons for policy and practice. The evidence synthesis in an SR may be either qualitative (narrative synthesis) or quantitative (meta-analysis), depending on whether the included studies are sufficiently homogeneous (in terms of interventions and outcomes studied, study design, and so forth) to be pooled in a meta-analysis. The chief advantage of SRs in the context of evidence-based policy is

⁷ This is not to impugn the value of narrative reviews *in general* or the competence or motivations of those who produce them—I myself have written many such reviews, including the current section of this paper. Rather, the point is simply that narrative reviews, by their very nature, have considerable limitations with respect to the specific goals that evidence tools hope to achieve.

that they provide, in an easy-to-digest format, a more accurate picture of the research community's collective understanding of an issue than could be obtained from looking at individual studies or from a narrative review.

The *systematic* nature of SRs and other evidence tools bears an important relation to the earlier discussion of evidence-based policy. A key component of this systematicity is including only studies that are broadly comparable to one another, such that together they comprise a coherent body of evidence related to the research question. This requires applying consistent screening criteria, including methodological criteria. And this, in turn, requires being able to clearly categorize the methods employed in a candidate study. While limited in many ways, the rigorous impact evaluation methods discussed above provide a basis for identifying studies whose results are both likely to have adequate internal validity and sufficiently comparable to be combined and synthesized. It is for this reason that most evidence tools—including the rapid evidence map I propose—include only studies employing experimental or quasi-experimental designs.

3.2.3 Rapid Evidence Assessments/Rapid Reviews

While systematic reviews have been advocated as superior evidence tools, researchers have come to realize that the significant time and effort required to complete such reviews is not always optimal for the purposes of supporting real-world decisions. Consequently, in the last decade a new approach to evidence synthesis, known as the rapid evidence assessment (REA) or rapid review, has emerged. The REA aims to provide many of the benefits of systematic reviews, typically, their systematicity and transparency, in a shorter timeframe by sacrificing other advantages of SRs, typically comprehensiveness (Butler, Deaton, Hodgkinson, Holmes, & Marshall, 2005; Collins, Miller, Coughlin, & Kirk, 2015; Thomas, Newman, & Oliver, 2013).

Thus, compared with systematic reviews, REAs typically address narrower questions, search a narrower range of sources, and provide less in-depth analysis of included studies while still identifying key policy lessons.

REAs have the clear advantage of providing an evidence synthesis that is significantly less time- and resource-consuming to produce (typically 1-6 months, compared to a year or more for many systematic reviews). This makes REAs more feasible for many decision-making bodies with limited resources, and allows for decisions to be evidence-based even when they must be made quickly. However, their narrower scope and speed-oriented approach to screening mean that they are much more likely to miss relevant publications, leading to biased conclusions. As will be discussed in the section on rapid evidence maps below, there are some circumstances in which these trade-offs are appropriate given the intended use of the review, while in other circumstances the costs of REAs outweigh the benefits.

3.2.4 Evidence Gap Maps

The evidence gap map (EGM) is a recently-devised tool, pioneered by the International Initiative for Impact Evaluation (3ie), that applies systematic review methodology for the purpose of evidence *mapping* rather than evidence *synthesis*. In a paper outlining the purpose and methods of EGMs, Snilstveit et al. characterize these maps as evidence collections that “present a visual overview of existing systematic reviews or impact evaluations in a sector or subsector, schematically representing the types of interventions evaluated and outcomes reported” (Snilstveit, Vojtkova, Bhavsar, & Gaarder, 2013, p. 3). In addition to providing a visual presentation of the current evidence base, EGMs typically enumerate the included studies and offer brief summaries of each of them. EGMs take the same exhaustive and standardized approach to searching and screening the literature as systematic reviews, but because they

provide less depth of analysis, can typically adopt a broader scope and map the evidence base for an entire sector or topic area, rather than focusing more narrowly on a particular question.

EGMs thus serve two primary functions (Snilstveit et al., 2013; see also Miake-Lye, Hempel, Shanman, & Shekelle, 2016). The first is *access to evidence*: EGMs provide decision-makers with a user-friendly tool for exploring the evidence base and accessing occurrences of evidence. Using an EGM, decision-makers can quickly see where the evidence base is extensive and where it is thin, and can see which sources they can pursue to delve deeper into the evidence most relevant to the decisions at hand. I will explore the utility of this kind of access to evidence more fully in the section below on rapid evidence maps.

The second function is *identifying gaps*, supporting strategic decisions about how to allocate research resources: the identification of gaps in the evidence base helps funders and researchers decide where their resources will add the most value to what is known. And investing in research in understudied areas will, in turn, enhance the utility of future evidence syntheses by creating a more integrated and relevant network of findings to be synthesized.

The visual portion of an EGM can take the form of either a spreadsheet or an interactive online tool.⁸ Figure 1 shows an example of a gap map framework (comprising four interventions and four outcomes) that has been populated with studies, with each study examining a different combination of interventions and outcomes. In this example, four studies have been included: **Study A** measured the effects of Intervention 1 on Outcome 1 and Outcome 3; **Study B** measured the effects of Intervention 1 on Outcome 3 and the effects of Intervention 2 on Outcomes 2 and 3; **Study C** measured the effects of Interventions 3 and 4 on Outcome 4; finally, **Study D** measured the effects of Intervention 3 on Outcome 4.

⁸ For examples of such online tools and supporting documentation, see <http://www.3ieimpact.org/en/evidence/gap-maps/>.

Figure 1: Example of a Populated Gap Map

		Outcome Categories			
		Outcome 1	Outcome 2	Outcome 3	Outcome 4
Intervention Categories	Intervention 1	Study A		Study A Study B	N/A
	Intervention 2		Study B	Study B	
	Intervention 3		N/A	N/A	Study C Study D
	Intervention 4				Study C

If a cell contains no studies, that means the search/screening process did not identify any studies that met the screening criteria and measured that particular intervention-outcome pair. In some cases, we might not really expect a particular intervention to affect a particular outcome—for example, we would not necessarily expect to find evidence about the effects of a few small youth groups on national-level laws and policies.⁹ The cells corresponding to such intervention-outcome pairs are marked “N/A” to indicate that the absence of studies does not necessarily indicate a “gap” in the evidence. The cells that are entirely empty, however, represent gaps in the evidence base. So, this map tells us, for example, that with regard to the effects of Intervention 1, there is some evidence about its effects on Outcome 1, somewhat more evidence about its effects on Outcome 3, and no evidence about its effects on Outcome 2.

It may be tempting to read the map as *also* telling us that Intervention 1 is more effective than Intervention 2 at improving Outcome 3 (i.e., because there are *two* studies on the effects of Intervention 1 on Outcome 3, and only *one* study on the effects of Intervention 2 on Outcome 3).

⁹ This is not to say that such effects could not exist. But the causal chain would be very indirect, making it difficult for a rigorous impact evaluation to study the relationship.

Importantly, however, this is not the case; the gap map method does not code the *results* of the studies, only *what* they studied. It may be the case, for example, that both Study A and Study B found that Intervention 1 had no effect at all on Outcome 3. Nevertheless, even without capturing the evidence of effectiveness for different intervention-outcome pairs in the framework, the map still conveys useful information. For cells that are densely populated with studies, we know that there is a rich store of evidence that can be mined for policy insights—that is, we know that an epistemic foundation exists for the design of “evidence-based” policies.

Full descriptions of the process for creating EGMs are available in concept documents (Snilstveit et al., 2013) and in published reports from completed gap maps (e.g., Rankin et al., 2014), so I will not cover the details of implementation here. However, there are some aspects of the process that are worth reviewing, as they provide useful context for discussion that follows about the process for creating rapid evidence maps. Once the topic or sector for an EGM has been identified, the first step in the EGM process begins with a general scope of the literature on the topic area, along with consultations with expert researchers and practitioners. Based on what is learned through this initial exploration, the research team identifies the key interventions and outcomes on which high-quality evidence is needed, and constructs a matrix (as in Figure 1) accordingly. Once the framework is established, the project continues much in the same way as a systematic review, with the research team conducting comprehensive searches and screening sources for relevance and quality. Once the researchers have identified the list of studies to be included, each study is coded based on the interventions and outcomes investigated, and the matrix/framework is populated accordingly.

For an evidence gap map, it is important that the framework be established relatively early and that it remain relatively stable throughout the process. In particular, while it can be a

good idea to add additional interventions and outcomes as they are discovered during the search and screening process, researchers should think carefully before eliminating elements from the framework, even if they are rarely or never discussed in the literature. The reason for this is that the framework represents *where evidence is needed*, not merely where it exists—indeed, if the EGM is to fulfill its purpose of identifying *gaps* in the evidence base, it must include some interventions and/or outcomes about which little or no evidence exists. My own research team at 3ie, for example, has encountered this while conducting an evidence gap map on the topic of adolescent sexual and reproductive health in low- and middle-income countries. One of the important outcomes identified during the framework development stage was access to and utilization of abortion services. However, perhaps unsurprisingly, given the sensitivity and politicization of the topic, the team has uncovered very little evidence about this outcome. But because this outcome was added to the framework at the outset, the completed EGM will be able to reflect and call attention to this lack of evidence, which can perhaps then be remedied in future work.

4. The Rapid Evidence Map

The rapid evidence map (REM) may be considered a hybrid of several existing evidence tools, particularly the rapid evidence assessment and the evidence gap map. Specifically, the REM combines the short timeframe and restricted scope of REAs with the mapping process of EGMs. The result will be an evidence tool that is systematic in its coverage of a narrow range of sources, can be completed in a timeframe similar to (or even shorter than) an REA, and provides a high-level overview of the evidence base on a range of topics relevant to particular policy or programming decisions. This section outlines, in conceptual terms, the motivations for developing the REM methodology and how it differs from existing evidence tools. Since most of

the evidence tools discussed here are relatively new (and in the case of REMs, entirely new), their characteristics and boundaries are not yet entirely well-defined, and consequently, neither are the distinctions among them. However, it is nonetheless possible to distinguish among several likely uses for these tools and to identify scenarios in which one or another tool is most likely to serve decision-makers' needs.

4.1 What are REMs for?

The purpose of a rapid evidence map is to support evidence-based decision-making on a particular policy decision. That is, REMs assemble evidence related to specific decisions that confront particular decision-makers, rather than addressing a wide user base (as is intended for systematic reviews or evidence gap maps). Thus, those most likely to benefit from REMs are decision-making bodies such as government agencies or NGOs that are seeking to expand their understanding of the evidence base for the effects of particular interventions or for how best to achieve particular outcomes.

Recall that evidence gap maps serve two functions: *access to evidence* and *identifying gaps*. REMs focus almost entirely on the first of these functions rather than the second. This is because of the likely audience for using REMs, namely entities that exist to serve constituents by directing policy, implementing programs, and delivering services. While such decision-makers may on occasion prioritize building new knowledge in addition to serving constituents directly, in most cases their interest will be in *using* the evidence base rather than *building* it. The existence of an evidence gap is a very pertinent piece of information for researchers (and some funders), who are substantially—if not primarily—concerned with expanding the entire development community's understanding of what works. However, the existence of such gaps is

of less immediate concern to entities whose core mandate is shaping policy and delivering services to constituents.

On the other hand, the other core function of EGMs—providing quick and easy access to evidence—is highly pertinent to decision-makers in developing country contexts, and this is something that REMs are capable of providing. REMs provide decision-makers with a convenient compilation of sources that have been vetted for quality and relevance, and that have been classified according to the key interventions and outcomes of interest in the decision-making context. Again, because the REM is a mapping tool rather than a synthesis tool, the populated REM does not convey information about how *effective* the included interventions are at improving the included outcomes. Rather, the REM indicates what evidence exists and where it can be found.

4.2 Advantages of the REM Approach

REMs present several advantages as aids for decision-making, relative to other types of evidence tools. First, when compared with traditional systematic reviews and evidence gap maps, REMs can be completed with significantly less investment of time and other resources. The benefit of this shortened timescale is twofold. First, it increases the feasibility of the evidence map approach for resource-constrained organizations and institutions, thus allowing a wider array of decision-making bodies to take advantage of the insights that evidence maps can provide. Second, REMs can provide many of the benefits of evidence mapping tools even when the timeframe for making particular decisions is short, thus increasing the chances that policy and programming decisions—even those that must be made quickly—can still draw on evidence from rigorous impact evaluations.

The narrower scope of an REM is a further advantage that makes REMs more useful evidence-access resources than evidence gap maps. While an EGM may include upwards of 100 individual impact evaluations and systematic reviews, while an REM, focusing on a much more targeted question, may compile 10 or 20 such studies. Thus, it would not be feasible or productive for decision-makers to attempt to retrieve and digest all or most of the studies included in an EGM, but this could be done with an REM. An REM therefore allows decision-makers to engage in much deeper exploration of evidence related to the question at hand.

The scope of an REM can be narrowed in a number of ways, all driven by the needs of the decision-maker. As discussed in Section 2, drawing on evidence from similar contexts is a key feature of evidence-based policymaking. Thus, during the initial phase of the REM process, when the interventions and/or outcomes of interest are identified, end users and research teams can work together to determine which contextual factors are most relevant, and restrict the scope accordingly. For example, it may be appropriate to draw evidence primarily from only the recent literature, from the same geographic region, from countries with similar political and economic institutions, or from populations of similar age, socioeconomic status, or cultural tradition. Restricting the scope of the search process in such ways can ensure both that the REM project is tractable in a short timeframe and that the evidence identified will be maximally relevant to providing actionable guidance.

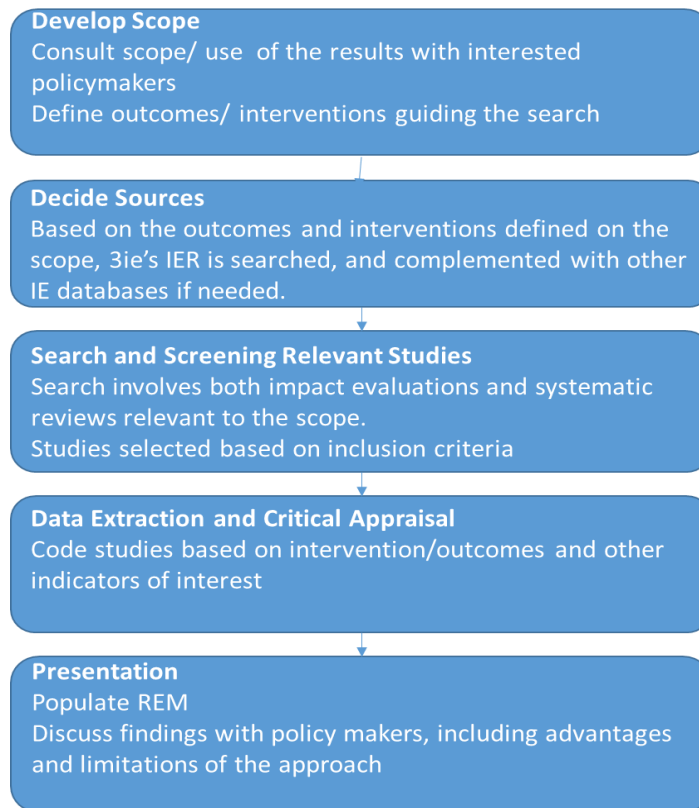
Of course, the advantages of a narrower scope and faster completion time also apply to rapid evidence assessments. However, REAs and REMs differ in several ways that make each of them suited to different decision-making contexts. One such difference is that REMs are better suited to addressing questions that are somewhat broader (though still fairly focused) or less clearly defined. Because REAs seek to *synthesize* evidence (i.e., provide greater depth of

analysis), they are best suited to addressing the evidence base for a smaller number of interventions and outcomes—e.g., the effects of one or two interventions on a handful of outcomes. REMs, meanwhile, aim to *map* the evidence (i.e., provide greater breadth of coverage), and so can include a range of interventions and outcomes pertinent to a group of related policy questions. Just as evidence gap maps provide a broader evidence overview with less depth of analysis than systematic reviews, so REMs provide a high-level picture of evidence in a domain while offering less specific answers than REAs about how a specific intervention might work. This lesser degree of specificity can itself be an advantage to REMs, since again REMs facilitate users' self-guided exploration of the literature. Thus, rather than having the evidence base distilled and filtered by researchers during a synthesis process, decision-makers can probe the relevant studies to extract whatever lessons they believe will aid their decisions, and can adapt their exploration process as their needs and their understanding of the issue evolve.

4.3 Rapid Evidence Maps: Process and Product

As with evidence gap maps, the final REM deliverable is a spreadsheet containing the populated matrix, with each cell containing references to all of the studies that provide evidence about that intervention-outcome pair. The process of developing an REM follows the same basic steps as an EGM (see Figure 2). First, the researchers, in consultation with other stakeholders, determine the scope, devise a framework, and decide which sources to search for literature. Next, researchers search these sources and screen the items retrieved. Those studies that meet screening criteria are coded according to the interventions and outcomes in the framework, and the map is populated accordingly. Finally, the REM's findings are presented to and discussed with stakeholders.

Figure 2: Process for Creating a Rapid Evidence Map



Source: International Initiative for Impact Evaluation (2016)

While the steps for completing an REM follow this general structure in most cases, it is important to note that this process can be adapted to fit the particular needs of the decision-maker. For instance, as mentioned above, in an evidence gap map it is important to develop the framework of interventions and outcomes at the outset, so that *gaps* can be identified where evidence is missing in the literature. However, because identifying gaps is not the focus of an REM, the development of the framework can be more open-ended and responsive to the needs of the audience and ongoing learning by the research team. For example, if a user has established a series of goals and is open to adopting a wide range of strategies to accomplish them, then it might be most appropriate to conduct the search and screening process based solely on a set of *outcomes* related to the user's goals, and allow the "interventions" axis of the framework to be filled in based on what is learned from the literature (i.e., which interventions targeting those

outcomes have been the subject of rigorous impact evaluations?). Alternatively, if an agency is planning to review its existing portfolio of programs to improve their implementation or redirect resources from less effective to more effective interventions, the framework can be driven by the interventions of interest, with outcomes being added to the framework later. This is the case, for example, with Mexico's National Council for Evaluation of Social Development Policy (CONEVAL), which wishes to review the evidence base pertaining to 12 of the Government's programs pertaining to food security (e.g., corn subsidies).¹⁰

5. Potential Use Case: Social Inclusion of Indigenous Groups in Peru

In conversations with 3ie, the Peruvian Government's Ministry of Development and Social Inclusion (MIDIS) has expressed an interest in an evidence tool to inform its initiatives to better serve and promote the interests of indigenous groups in the country's Amazonian areas. The need for such a tool within a short timeframe makes this a promising opportunity for applying the REM method. In particular, because the question of interest to MIDIS is a relatively broad one (*what strategies are effective in improving outcomes for indigenous groups?*), the broader picture provided by an REM will likely be more useful than an REA, which would only address the evidence for perhaps one or two particular outcomes. Working with 3ie staff, I have completed initial scoping work on MIDIS's current decision-making context and need for evidence, and outlined the scope of the project. This section describes this potential use case for an REM, and demonstrates how the needs of decision-makers can drive the development of an REM framework.

¹⁰ Mario Picon, personal communication.

5.1 State of Indigenous Populations in Perú

The Peruvian Housing and Population Census uses native language as the defining criterion for identifying indigenous populations (Planas & Vattuone, 2012, p. 11). Under this definition, there were approximately 4 million indigenous persons in Perú in 2012, representing about 16 percent of the total Peruvian population. Within the indigenous population, the most common native languages are Quechua (native language of 83 percent of indigenous Peruvians) and Aymara (11 percent). The Quechua-speaking population is concentrated in several regions, with the largest populations inhabiting the Cusco and Lima areas. The remaining six percent are speakers of various Amazonian languages.

Fifty-five percent of Perú's indigenous population is classified as living in poverty, with 26 percent living in extreme poverty. This is compared to rates of 28 percent and 6 percent for poverty and extreme poverty, respectively, among the general population (Davis & Edwards, 2013, p. 8). Poverty rates are especially high among Amazonian-language speakers, with 81 percent of this population experiencing poverty and 41 percent experiencing extreme poverty (Planas & Vattuone, 2012, p. 12). Levels of food insecurity and chronic child malnutrition are significantly higher among Perú's indigenous populations than in the country as a whole (Scaling Up Nutrition, 2015, pp. 9–11).

5.2 MIDIS: Mission, Priorities, and Strategic Plan

Established in 2011, MIDIS's mission is to “improve the quality of life of populations facing vulnerability and poverty, and to promote the exercise of their rights, their access to opportunities, and the development of their own capacities” (Planas & Vattuone, 2012, p. 3).¹¹ The ministry's overall approach attends to the short-term needs of constituents while building

¹¹ For this and other documents only available in Spanish, the translations are my own.

communities’ capacity and securing long-term progress through investment in early childhood development and education.

Currently, MIDIS’s flagship initiative is known as “Incluir para Crecer” (“Include to Grow”), which aims to promote positive outcomes across the lifespan by focusing on five periods of individual development (see Figure 3):

- Infant nutrition: 0-3 years
- Early childhood development: 0-5 years
- Integrated childhood and adolescent development: 6-17 years
- Economic inclusion: 18-64 years
- Protection for older adults: 65 years and beyond

Each of these goals is tied to a series of indicators, which are provided in the Appendix.

Figure 3: Goals and Intended Outcomes of Include to Grow

Goals	Goal 1: Infant Nutrition	Goal 2: Early Childhood Development	Goal 3: Comprehensive Child and Adolescent Development	Goal 4: Economic Inclusion	Goal 5: Protection for the Elderly	
Results	Reduce chronic malnutrition in children	Improve the physical, cognitive, motor, emotional and social development during early childhood	Improve skills for personal, educational and occupational development according to age	Increase autonomous household incomes	Improve welfare of the elderly	
0-5 YEARS		6-17 YEARS		18-64 YEARS		65 YEARS AND OVER

Source: Ministry of Development & Social Inclusion (2014, p. 29)

Also central to the mission of Include to Grow is its focus on individuals classified as part of the “population in the process of inclusion” (PEPI). This population includes rural households,

female-headed households, households in the bottom quintile of the national income distribution, and, most relevant for the REM proposed here, indigenous households (Ministry of Development & Social Inclusion, 2014, p. 15).

To address the needs of vulnerable citizens across the lifespan, MIDIS implements a number of core programs, including childhood nutrition programs, conditional cash transfers for education, infrastructure and enterprise investments, and a pension program for those 65 and older (Ministry of Development & Social Inclusion, 2014; Planas & Vattuone, 2012). However, a recent report identified a number of barriers that serve to limit access to many of these programs among indigenous groups (Planas & Vattuone, 2012, pp. 15–24). These barriers include inadequate information about program eligibility, varying local requirements, problems with disbursements, geographic dispersion of eligible households, and culturally irrelevant services. Thus, MIDIS is interested in using evidence from other settings to identify effective strategies for serving indigenous communities and adapt the implementation of existing programs to enhance their inclusiveness.

5.3 Proposed REM Framework, Search Strategy, and Screening Criteria

Because MIDIS’s interest in this case is in achieving particular outcomes across the lifespan, rather than assessing any particular policies related to these goals, this REM will begin with an outcomes-based framework, with interventions to be added based on the evidence discovered in the literature. Figure 4 depicts the eight key outcomes corresponding to MIDIS’s goals, grouped into four thematic categories.

Figure 4: Proposed Outcomes-based Framework: REM on Serving Indigenous Populations

EARLY NUTRITION AND HEALTH	CHILDHOOD AND ADOLESCENT DEVELOPMENT	ECONOMIC INCLUSION	PROTECTION OF THE ELDERLY
----------------------------	--------------------------------------	--------------------	---------------------------

Infant nutrition & physical development	Childhood nutrition & development	School enrollment	School attendance	Academic achievement	Household income	Health outcomes for the elderly	Social outcomes for the elderly
---	-----------------------------------	-------------------	-------------------	----------------------	------------------	---------------------------------	---------------------------------

As there is a large body of evidence regarding each of these outcomes, three further scope restrictions will serve to narrow the focus of the REM and increase the relevance of the collected evidence to MIDIS’s decision-making. First, studies must investigate the effects of an intervention on one of these outcomes *among indigenous populations*. Second, studies must examine a *government-implemented policy or program*. Finally, studies must investigate interventions undertaken *in Latin America*.

Consistent with the aims of a rapid evidence map, the search strategy for this REM will employ a limited number of databases and websites, but will aim to capture all relevant literature contained within those databases:

- The 3ie Impact Evaluation Repository¹²
- LARreferencia, a Latin American database of scientific articles, theses, and other sources¹³
- The Cochrane Library of systematic reviews¹⁴
- The Campbell Collaboration Library of Systematic Reviews¹⁵

Precise search terms are being developed in consultation with 3ie staff, but will include “indigenous,” or “native” plus terms related to the interventions and outcomes of the framework and terms to narrow the results by region and methodology. Following standard practice for 3ie’s evidence maps, in order to be included in the REM, a study must have been conducted in a country classified by the World Bank as low- or middle-income. In addition, the study must have used an experimental (i.e., randomized) or quasi-experimental (difference-in-differences,

¹² <http://www.3ieimpact.org/en/evidence/impact-evaluations/impact-evaluation-repository/>

¹³ <http://lareferencia.redclara.net/rfr/>

¹⁴ <http://www.cochranelibrary.com/>

¹⁵ <http://www.campbellcollaboration.org/lib/>

regression discontinuity, propensity score matching, or instrumental variable estimation) design to assess the impact of an intervention relative to a counterfactual. The goal of the search will be to retrieve all sources that describe impact evaluations or systematic reviews of the effects of government-run an intervention.

6. Conclusion

Using evidence to guide development policy is a worthy goal, with significant potential for improving human lives. Assessing the evidence for a potential policy is a holistic affair that should include the results of impact evaluations that provide causal knowledge, although context-relevance may often be more important than sheer methodological rigor. But policymakers cannot rely on this kind of evidence without tools that systematically gather and synthesize it. While there are many research tools available to support evidence-based policymaking, there is currently an absence of tools that combine the benefits of systematic evidence mapping with the benefits of a rapid completion timeline. The rapid evidence map seeks to fill this gap, and the pilot REM project outlined in this paper will serve as both a proof of concept and an opportunity to solidify and refine the REM methodology for other researchers to use. More importantly, it will aid a policymaking body in fulfilling its mission of promoting inclusive and sustainable development for its citizens.

References

- Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, 46(3), 399–424. <http://doi.org/10.1080/00273171.2011.568786>
- Boaz, A., Ashby, D., & Young, K. (2002). *Systematic reviews: what have they got to offer evidence based policy and practice?* (Working Paper No. 2). London: ESRC UK Centre for Evidence Based Policy and Practice. Retrieved from <http://www.kcl.ac.uk/sspp/departments/politicaeconomy/research/cep/pubs/papers/assets/wp2.pdf>
- Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2013). Scaling up what works: Experimental evidence on external validity in Kenyan education. *Center for Global Development Working Paper*, (321). Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2241240
- Bor, J., Moscoe, E., Mutevedzi, P., Newell, M.-L., & Bärnighausen, T. (2014). Regression discontinuity designs in epidemiology: Causal inference without randomized trials. *Epidemiology*, 25(5), 729–737. <http://doi.org/10.1097/EDE.0000000000000138>
- Brown, A. N. (2016, April 19). Let's bring back theory to theory of change. Retrieved from <http://blogs.3ieimpact.org/lets-bring-back-theory-to-theory-of-change/>
- Butler, G., Deaton, S., Hodgkinson, J., Holmes, E., & Marshall, S. (2005). *Quick but not dirty: rapid evidence assessments as a decision support tool in social policy*. London: Government Social Research Unit.
- Cartwright, N. (1979). Causal laws and effective strategies. *Noûs*, 13(4), 419–437. <http://doi.org/10.2307/2215337>
- Cartwright, N. (2007). Are RCTs the Gold Standard? *BioSocieties*, 2(1), 11–20. Retrieved from <http://search.proquest.com/reference.sit.edu:2048/pqcentral/docview/220825525/abstract>
- Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.
- Collins, A., Miller, J., Coughlin, D., & Kirk, S. (2015). *The production of quick scoping reviews and rapid evidence assessments: A how to guide*. Joint Water Evidence Group. Retrieved from <https://sbri.innovateuk.org/documents/3058188/3918930/The+Production+of+QSRs+and+REAs+How+to+guide.pdf/45975020-be7d-4788-b74b-f3b6ed32c73a>

- Cook, T. D., Shadish, W. R., & Wong, V. C. (2005). Within-study comparisons of experiments and non-experiments: What the findings imply for the validity of different kinds of observational study. In *French Econometric Society Meeting on Program Evaluation, Paris, France*. Retrieved from http://www.metheval.uni-jena.de/projekte/symposium2006/material/cook_shadish_and_wong.pdf
- Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, 27(4), 724–750. <http://doi.org/10.1002/pam.20375>
- Davis, M., & Edwards, S. (2013). *Killing us with hunger: Indigenous perspectives on nutrition, food aid and food sovereignty in the central Amazon of Peru*. London: Health Poverty Action. Retrieved from <https://www.healthpovertyaction.org/wp-content/uploads/downloads/2013/10/Killing-us-with-Hunger-web2.pdf>
- Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature*, 424–455. Retrieved from <http://www.jstor.org/stable/20778731>
- Dehejia, R., Pop-Eleches, C., & Samii, C. (2015). *From local to global: External validity in a fertility natural experiment* (IZA Discussion Paper No. 9300). Retrieved from <http://www.nber.org/papers/w21459>
- Duflo, E., Glennerster, R., & Kremer, M. (2008). Using randomization in development economics research: A toolkit. In T. P. Schultz & J. Strauss (Eds.), *Handbook of development economics* (Vol. 4). Elsevier.
- Gaarder, M. M., Glassman, A., & Todd, J. E. (2010). Conditional cash transfers and health: unpacking the causal chain. *Journal of Development Effectiveness*, 2(1), 6–50. <http://doi.org/10.1080/19439341003646188>
- Gasteen, M. (2010). *Systematic reviews and evidence-informed policy: Overview*. London: Department for International Development. Retrieved from http://r4d.dfid.gov.uk/PDF/Articles/What_is_a_systematic_review_external.pdf
- Gechter, M. (2015). *Generalizing the results from social experiments: Theory and evidence from Mexico and India*. Retrieved from http://www.economics.huji.ac.il/Uploads/Contents/documents/Gechter_Generalizing_Social_Experiments.pdf
- Glazerman, S., Levy, D., & Myers, D. (2002). *Nonexperimental replications of social experiments: a systematic review* (Interim Report). Washington, DC: Mathematica Policy

- Research. Retrieved from
<http://econpapers.repec.org/RePEc:mpr:mprres:9e8a2edf60ef4fa89ad9a5acf1a7948b>
- Grant, M. J., & Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, 26(2), 91–108.
<http://doi.org/10.1111/j.1471-1842.2009.00848.x>
- Heckman, J. J. (1991). *Randomization and social policy evaluation* (NBER Working Paper No. 107). Cambridge, MA: National Bureau of Economic Research. Retrieved from
<http://www.nber.org/papers/t0107>
- Higgins, J. P., & Green, S. (2008). *Cochrane handbook for systematic reviews of interventions*. The Cochrane Collaboration. Retrieved from <http://handbook.cochrane.org/>
- Imbens, G. M., & Wooldridge, J. M. (2008). *Recent developments in the econometrics of program evaluation*. National Bureau of Economic Research. Retrieved from
<http://www.nber.org/papers/w14251>
- International Initiative for Impact Evaluation. (2016). *Concept note: 3ie's Rapid Evidence Maps*. Washington, DC: International Initiative for Impact Evaluation.
- Jones, H. (2009). *The “gold standard” is not a silver bullet for evaluation* (ODI Opinion No. 127). London: Overseas Development Institute. Retrieved from
<http://www.alnap.org/pool/files/3695.pdf>
- Karlan, D., Goldberg, N., & Copstake, J. (2009). “Randomized control trials are the best way to measure impact of microfinance programmes and improve microfinance product designs.” *Enterprise Development and Microfinance*, 20(3), 167–176. <http://doi.org/10.3362/1755-1986.2009.017>
- Mavedzenge, S. M. N., Doyle, A. M., & Ross, D. A. (2010). HIV prevention in young people in sub-Saharan Africa: a systematic review. *Journal of Adolescent Health*, 49(6), 568–586. Retrieved from <http://www.sciencedirect.com/science/article/pii/S1054139X11000565>
- Miaka-Lye, I. M., Hempel, S., Shanman, R., & Shekelle, P. G. (2016). What is an evidence map? A systematic review of published evidence maps and their definitions, methods, and products. *Systematic Reviews*, 5, 28. <http://doi.org/10.1186/s13643-016-0204-x>
- Ministry of Development & Social Inclusion. (2014). *National strategy for development and social inclusion “Include to Grow.”* Lima: Ministry of Development & Social Inclusion. Retrieved from http://incluirparacrece.midis.gob.pe/descargas/include_to_grow_2014.pdf

- Northcott, R. (2006). Causal efficacy and the analysis of variance. *Biology & Philosophy*, 21(2), 253–276. <http://doi.org/10.1007/s10539-005-8304-3>
- Planas, M. E., & Vattuone, M. E. (2012). *Peru - Estrategia de pueblos indigenas* (No. IPP603) (pp. 1–49). The World Bank. Retrieved from <http://documents.worldbank.org/curated/en/2012/10/16901193/peru-pe-social-inclusion-tal-indigenous-peoples-plan-peru-estrategia-de-pueblos-indigenas>
- Pritchett, L., & Sandefur, J. (2013). Context matters for size: Why external validity claims and development practice do not mix. *Journal of Globalization and Development*, 4(2), 161–197. <http://doi.org/http://dx.doi.org/10.1515/jgd-2014-0004>
- Rankin, K., Cameron, D. B., Ingraham, K., Mishra, A., Burke, J., Picon, M., ... Brown, A. N. (2014). *Youth and transferable skills: An evidence gap map* (3ie Evidence Gap Map No. 2). New Delhi: International Initiative for Impact Evaluation (3ie). Retrieved from http://www.3ieimpact.org/media/filer_public/2015/09/01/egm2-youth_and_transferable_skills.pdf
- Ravallion, M. (2007). Evaluating anti-poverty programs. *Handbook of Development Economics*, 4, 3787–3846. Retrieved from <http://www.sciencedirect.com/science/article/pii/S1573447107040594>
- Rodrik, D. (2008). *The new development economics: we shall experiment, but how shall we learn?* Faculty Research Working Papers Series, Harvard University. Retrieved from http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=1296115
- Scaling Up Nutrition. (2015). *Strengthening the capacity of sun countries to scale up nutrition through learning routes: A pioneer project in peru final report*. Retrieved from http://scalingupnutrition.org/wp-content/uploads/2015/03/INFORME-FINAL-Ruta-Per%C3%BA_For-complete-ENG.pdf
- Schrodt, P. A. (2014). Seven deadly sins of contemporary quantitative political analysis. *Journal of Peace Research*, 51(2), 287–300. <http://doi.org/10.1177/0022343313499597>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage learning. Retrieved from <http://impact.cgiar.org/pdf/147.pdf>
- Snilstveit, B. (2012). Systematic reviews: from “bare bones” reviews to policy relevance. *Journal of Development Effectiveness*, 4(3), 388–408. <http://doi.org/10.1080/19439342.2012.709875>

- Snilstveit, B., Vojtkova, M., Bhavsar, A., & Gaarder, M. (2013). Evidence gap maps—a tool for promoting evidence-informed policy and prioritizing future research. *World Bank Policy Research Working Paper*, (6725). Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2367606
- Thomas, J., Newman, M., & Oliver, S. (2013). Rapid evidence assessments of research to inform social policy: taking stock and moving forward. *Evidence & Policy: A Journal of Research, Debate and Practice*, 9(1), 5–27. <http://doi.org/10.1332/174426413X662572>
- Vivalt, E. (2015). *How much can we generalize from impact evaluations?* Unpublished manuscript, Stanford University. Retrieved from <http://evavivalt.com/wp-content/uploads/2014/10/Vivalt-JMP-10.27.14.pdf>
- Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Appendix: Indicators and Goals of Include to Grow

The figures below provide the intermediate results and indicators for each of MIDIS's five strategic goals (see Ministry of Development & Social Inclusion, 2014, sec. 2.2).

Goal 1: Infant Nutrition

FINAL RESULT	Verifiable Indicator	Unit of measurement	National		PEPI		Verification Sources
			Baseline 2010	Goal for 2016	Baseline 2010	Goal for 2016	
Reduce prevalence of Chronic Malnutrition in Children under 3 years of age.	Rate of Chronic Malnutrition in Children according to WHO standard (children aged 0-5*)	%	23.2%	10.0%	50.0%	24.0%	ENDES
INTERMEDIATE RESULTS							
1. Reduce incidence of low birth weight	Percentage of children born with low weight (less than 2,5 kg)	%	7.2%	3.1%	12.5%	6.0%	ENDES
	Percentage of mothers who attend to 6 or more prenatal checkups	%	81.0%	90.0%	69.6%	85.0%	ENDES
2. Reduce index of morbidity through ARI and ADD in children under 36 months	Percentage of children under 36 months who had ARI in the last 15 days	%	18.3%	8.0%	19.3%	9.0%	ENDES
	Percentage of children under 36 months who had ADD in the last 15 days	%	16.6%	7.0%	17.3%	7.0%	ENDES
3. Diet quality (micronutrients) for children under 36 months	Percentage of children under 6 months who are exclusively breastfed	%	70.4%	90.0%	81.3%	95.0%	ENDES
	Percentage of children aged 6 to 59 months with anemia	%	37.7%	16.0%	46.2%	20.0%	ENDES
	Percentage of children aged 6 to 36 months who received Vitamin A doses	%	32.5%	50.0%	59.4%	65.0%	ENDES
	Percentage of children aged 6 to 36 months who received an iron supplement in the last 7 days	%	18.1%	50.0%	27.9%	55.0%	ENDES
	Food Insecurity Index MIDIS	VALUE	0.230	0.150	-	-	CENSO/ENAH0

Goal 2: Early Childhood Development

FINAL RESULT	Verifiable Indicator	Units of Measurement	National		PEPI		Verification Sources
			Baseline 2010	Goal by 2016	Baseline 2010	Goal by 2016	
Increase levels of physic, cognitive, motor, emotional, and social development in children aged 0 to 5.	Percentage of children aged 0 a 36 months who accomplish motor milestones (standing, running, going up and down the stairs) according to "Age and Stage" questionnaires	%					Survey of Early Childhood Health and Development
	Percentage of children aged 0 to 36 months who accomplish language milestones (a vocabulary of more than 50 words, ability to create a two-words phrase) according to "Age and Stage" test.	%	Not available yet	-	Not available yet	-	Survey of Early Childhood Health and Development
	Percentage of children aged 3 to 5 years who get the required score in the Picture Vocabulary Test "PPVT"	%					Survey of Early Childhood Health and Development
INTERMEDIATE RESULTS							
1. Improvement in Children's Health (chronic Malnutrition and infant morbid-mortality)	Rate of Chronic Malnutrition in children according to WHO standard (children aged 0-5)	%	23.2%	10.0%	50.0%	24.0%	ENDES
	Children mortality rate (in each thousand live births)	Value	17.0	13.0	18.0	14.0	ENDES
	Percentage of mothers who had institutional deliveries	%	84.4%	93.0%	51.4%	77.0%	ENDES
2. Quality of the environment where children develop	Number of families receiving orientation on children's care and development practices through Cuna Más	Number of families	-	89,720	-	-	Cuna Más
	Percentage of households with access to safe water.	%	76.8%	85.0%	40.1%	67.0%	ENAHO
	Percentage of households with drainage services	%	77.0%	85.0%	44.3%	70.0%	ENAHO
	Municipality member of the Project for Healthy Municipalities and Communities	Number of scheduled municipalities	249	500	-	-	Records of Project for Healthy Municipalities
4. Improvement in access to educative and quality early childhood services	Cuna Más* users: children aged 0 to 36 months	Number of users	64,040	240,000	-	-	Cuna Más
	Attendance of children aged 3 to 5 to Basic Regular Education.	%	73.8%	85.0%	60.9%	78.4%	ENAHO
	Number of users of the day care program Cuna Más	Value	6,061	6,900	-	-	Cuna Más

Goal 3: Comprehensive Childhood and Adolescent Development

FINAL RESULT	Verifiable Indicator	Unit of Measurement	National		PEPI		Verification Sources
			Baseline 2010	Goal by 2016	Baseline 2010	Goal by 2016	
Increase competencies for personal, educational and occupational development of children and adolescents according to age.	Percentage of elementary students who achieve the expected level in reading comprehension in the ECE.	%	28.7%	51.0%	-	-	ECE (Evaluación Censal de Estudiantes - Student Census Evaluation)
	Percentage of elementary students who achieve the expected level in math in the ECE.	%	13.8%	44.0%	-	-	ECE
	Percentage of high-school students who achieve the expected level (level 3) in reading comprehension of PISA. *	%	10.1%	32.0%	-	-	PISA
	Percentage of high-school students who achieve the expected level (level 3) in math of PISA. *	%	6.8%	31.0%	-	-	PISA
	Percentage of high-school students who achieve the expected level (level 3) in Science tests PISA. *	%	8.0%	31.0%	-	-	PISA
	Percentage of young people between 18 to 25 years who have finished high-school.	%	75.3%	85.0%	46.0%	60.0%	ENAH0
INTERMEDIATE RESULTS							
1. Improve quality, equality and supply of appropriate education for children and adolescents.	Percentage of children aged 6 to 11 who attend elementary school.	%	90.9%	96.0%	88.2%	89.0%	ENAH0
	Percentage of adolescents aged 12 to 17 who attend high-school.	%	69.5%	86.0%	57.3%	63.0%	ENAH0
2. Reduce incidence of risk behaviors in children and adolescents.	Incidence of pregnancy in adolescents aged 15 to 19	%	13.5%	9.0%	20.5%	16.0%	ENDES
	Incidence of illegal drug consumption by high-school students *	%	4.6%	4.0%	-	-	III National Survey: Prevention and Use of Drugs among secondary students
	Incidence of legal drug consumption by high-school students *	%	28.9%	27.0%	-	-	III National Survey: Prevention and Use of Drugs among secondary students
3. Reduce child and adolescent labor	Percentage of children aged 6 to 13 who work during school time.	%	18.8%	9.0%	47.8%	34.0%	ENAH0
	Percentage of adolescents aged 14 to 17 who work during school time.	%	38.8%	16.0%	60.6%	48.0%	ENAH0

Goal 4: Economic Inclusion

FINAL RESULT	Verifiable Indicator	Unit of Measurement	National		PEPI		Sources of verification
			Baseline 2010	Goal by 2016	Baseline 2010	Goal by 2016	
Increase household income	Incidence of extreme poverty based on income (population aged 18 to 64).	%	7.3%	4.0%	42.7%	22.0%	ENAH0
	Incidence of poverty based on income (population aged 18 to 64).	%					ENAH0
INTERMEDIATE RESULTS							
1. Increase provision of family's assets	Percentage of farming units (households) in districts with more than 50 % rate of extreme poverty that receive technical assistance (training, technical assistance or business advising).	%	n.d		n.d		Agricultural Survey 2012
	Amount of crops and livestock produced by households in process of extreme social inclusion (PEPIEx)	Number of crops and livestock	-	-	Crops:6.8; Livestock:3.5	Crops:7; Livestock:5	ENAH0
	Percentage of incomes from non-agricultural income sources for households not in extreme poverty (PEPIE no extremo)	%	-	-	39.0%	45.0%	ENAH0
	Hectares under irrigation (as % of total amount of households in extreme poverty)	Number of hectares	-	-	6.0%	20.0%	ENAH0
	Cattle value at baseline year 2011 prices (household average)	Value in soles; 2011 prices	-	-	575	1060	ENAH0
2. Increase productivity of family's assets.	Percentage of households accessing the integrated package of services	%	59.4%	70.0%	11.6%	46.1%	ENAH0
	Offices of financial institutions for every 100,000 adult inhabitants	Value	96.0	150.0	30.0	110.0	SBS
	Percentage of adult population with saving accounts.	%	42.0%	60.0%	n.d	-	SBS
	% farming units (households) in districts with more than 50 % of their population in poverty and extreme poverty that access farming information by phone, radio, television, internet, or written publications.	%	n.d	-	n.d	-	Agricultural Survey 2012
	Percentage of children aged 6 to 36 months who access day care services in prioritized districts.	%	-	-	11.5	16.0	Records of Cuna Más
	% of farming units in districts with more than 50 % extreme poverty who belong to an association, committee or cooperative of producers.	%	n.d		n.d		Agricultural Survey 2012

Goal 5: Protection for the Aged

FINAL RESULT	Verifiable Indicator	Unit of Measurement	National		PEPI		Sources of verification
			Baseline 2010	Goal by 2016	Baseline 2010	Goal by 2016	
Increase welfare and protection for the elderly	Extreme poverty rate of the elderly aged 65 or more	%	8.3%	7.0%	34.6%	21.0%	ENAH0
	Poverty gap in households with adult members age 65 and over	%	7.1%	5.0%	25.2%	15.0%	ENAH0
	Subjective poverty rate in households with adult members age 65 and over	%	48.2%	34.0%	77.6%	50.0%	ENAH0
INTERMEDIATE RESULTS							
1. Increase economic security for the elderly in poverty	Percentage of adults age 65 or more who receive a pension, including non contributory pension.	%	23.0%	40.0%	3.2%	30.0%	ENAH0
	Number of literate people, age 60 and over by literacy programs.	Number of people	n.d	11,000	-	-	MINEDU
2. Improve the access to quality health services for the elderly	Percentage of adults age 65 and over who have diseases, malaises, relapses and / or accidents and have not received medical treatment.	%	63.8%	40.0%	71.4%	45.0%	ENAH0
	Percentage of adults aged 65 or more who have health insurance.	%	69.29%	85.0%	70.5%	85.0%	ENAH0
	Older adult beneficiaries of the Oral Health Program "Vuelve a sonreir" (Smile again)	Number of people	n.d	56,000	-	-	MINSAs
	Number of homeless people age 60 and over who receive benefits of the National Program for Decent Life	Number of people	n.d	1,050	-	-	MIMP
3. Strengthen local networks of social support for the elderly.	Number of Local governments that implement Elderly Care Centers (CIAM)	Number of local governments	120	420	-	-	MIMP
	Number of people aged 60 and over who access to Elderly Care Centers (CIAM)	Number of people	n.d	110,000	-	-	MIMP
	Number of Elderly Centers (CAM) in Provincial Municipalities.	N° CAMs	112	125	-	-	ESSALUD