United Nations Research Institute for Social Development

Working Paper 2015-4

Methods of Measuring the Impacts of Social Policy in Political, Economic and Social Dimensions

Michael Samson, Sasha van Katwyk, Maarten Fröling, and Rumbidzai Ndoro, with Cara Meintjes, Lien Buts and Bryant Renaud

prepared for the UNRISD project on
Towards Universal Social Security in Emerging Economies: Process, Institutions and Actors

February 2015

The United Nations Research Institute for Social Development (UNRISD) is an autonomous research institute within the UN system that undertakes multidisciplinary research and policy analysis on the social dimensions of contemporary development issues. Through our work we aim to ensure that social equity, inclusion and justice are central to development thinking, policy and practice.

UNRISD, Palais des Nations
1211 Geneva 10, Switzerland

Tel: +41 (0)22 9173020
Fax: +41 (0)22 9170650
info@unrisd.org
www.unrisd.org

Contents

Acronyms

IV	Instrumental Variables
LEAP	Livelihood Empowerment Against Poverty
LEWIE	Local Economy-wide Impact Evaluations
NGO	Non-governmental Organization
PSM	Propensity Score Matching
QCA	Qualitative Comparative Analysis
RCT	Randomized Control Trial
RDD	Regression Discontinuity Design

Summary

Over the past decade the world has seen a sea change in the role of evidence informing social policy design and implementation. In the social protection sector in particular, rigorous quantitative impact assessments have changed how policy analysts and implementers measure success. Mainstream evaluations increasingly rely on experimental approaches, and sometimes quasi-experiments, requiring important compromises. Given the complexity of many critical policy questions, evaluation designers often face a trade-off between ensuring the most rigorous attribution of impact and illuminating the critical policy questions that policy makers are asking. An evaluation approach that balances the prerequisites for analytical rigour with the demands of policy makers for relevant answers is required to design and implement more effective social policies and strategies.

This paper reviews current impact assessment methods and builds on Amartya Sen's framework of comprehensive and culmination outcomes to identify elements of a comprehensive framework that enables a systems approach to the analysis of social policy. Discussing how mainstream evaluation methods have assessed the outcomes of social security systems, the underlying assumptions of these methods and the associated challenges for the realization of comprehensive outcomes, the paper argues that inclusion of the processes, institutions and actors of social policy interventions that importantly affect programme objectives—along with the actually realized outcomes—should be integrated into a comprehensive approach to better inform social policies.

The emerging framework recognizes the value of a cross-sectoral analysis that studies social, political and economic aspects across a variety of dimensions. It recognizes the importance of both short- and long-term analysis within a policy environment driven by a multiplicity of strategic objectives. In particular, the framework explicitly recognizes that the impact of one sectoral intervention on a specific outcome depends critically on the related interventions across a range of sectors. A comprehensive evaluation approach should inform the optimal balancing of multiple interventions to achieve a range of joint outcomes.

This paper is organized as follows: Section 1 introduces the concepts of culmination and comprehensive outcomes, following a framework proposed by Sen. Section 2 discusses the current mainstream impact evaluation methods for social protection. Several key aspects of a comprehensive evaluation that are not present in current mainstream methods are identified. Section 3 discusses in further depth the features of a comprehensive evaluation, and identifies steps that are already being taken toward a comprehensive evaluation. Emphasizing that inclusion of the processes, institutions and actors of social policy interventions that affect programme objectives in important ways—along with the actually realized outcomes—is central to alternative approaches to expand the scope of the assessment, section 4 concludes with the exposition of elements of a framework for comprehensive evaluation, and discusses future challenges and opportunities.

Michael Samson is Director of Research, Economic Policy Research Institute, South Africa. At the time of writing, Sasha van Katwyk, Maarten Fröling, Rumbidzai Ndoro, Cara Meintjes and Bryant Renaud were Fellows at EPRI and Lien Buts was an Intern there.

Introduction

Over the past decade the world has seen a sea change in the role of evidence informing social policy design and implementation. In the social protection sector in particular, rigorous quantitative impact assessments have changed how policy analysts and implementers measure success. However, a growing emphasis on methodological rigour has crowded out a more comprehensive approach to evidence-building, thereby creating gaps in the evidence base policy makers require to design and implement more effective strategies. Mainstream evaluations increasingly rely on experimental approaches, and sometimes quasi-experiments, requiring important compromises. An evaluation approach that better informs policy development would balance the prerequisites for analytical rigour with the demands of policy makers for relevant answers.

Amartya Sen has distinguished comprehensive outcomes from culmination outcomes. In culmination outcomes the analysis is confined to accounting for the consequences of determined inputs, with little consideration of the interactions, interests or unforeseen influence of actors and institutions throughout the process. Comprehensive outcomes, Sen (2009) contrasts, comprise the process, institutions and actors, as well as the outcomes of their actions. The application of this concept of comprehensive outcomes to the analysis of social security systems enables the expansion of the scope of assessment of social security systems.

This paper discusses how mainstream evaluation methods have assessed the outcomes of social security systems, exploring the underlying assumptions of these methods and how they create challenges for the realization of comprehensive outcomes. The processes, institutions and actors of social policy interventions affect programme objectives in important ways along with the actually realized outcomes. While a number of existing evaluation approaches offer potential to improve the measurement of comprehensive outcomes, more innovative work is required to better address the demands of policy makers and stakeholders who envision a proactive role for evidence in supporting better social policy.

In 1957 Campbell's foundational paper "Factors Relevant to the Validity of Experiments in Social Settings" (Campbell 1957) explored the concepts of internal and external validity as well as experimental and quasi-experimental design, laying the groundwork for the growth of randomized control trials (RCTs) that today some herald as the "gold standard" of social policy evaluation. Critics, however, point out that these quantitative impact assessments are often too limited, focusing on average, quantifiable effects while ignoring impact heterogeneity and important qualitative elements. Some of the most important developmental outcomes are difficult to quantify, and often evolve over a longer-term horizon than most RCTs encompass. Many important social policy questions cannot be reduced to experiments, and some experiments create conditions that policy makers reject for ethical, practical or political reasons.

This paper reviews current evaluation methods and builds on Sen's framework of comprehensive and culmination outcomes to identify elements of a comprehensive evaluation approach. The emerging framework recognizes the value of a cross-sectoral analysis which studies social, political and economic aspects across a variety of dimensions. It recognizes the importance of both short- and long-term analysis. The framework, for example, emphasizes linkages between social protection programme, sectors and institutions.

This paper is organized as follows: Section 1 introduces the concepts of culmination and comprehensive outcomes, following a framework proposed by Sen. Section 2 discusses the current mainstream impact evaluation methods for social protection. Several key aspects of a comprehensive evaluation that are not present in current mainstream methods are identified. Section 3 discusses in further depth the features of a comprehensive evaluation, and identifies steps that are already being taken toward a comprehensive evaluation. Section 4 concludes with the exposition of elements of a framework for comprehensive evaluation, and discusses future challenges and opportunities.

1. Comprehensive Versus Culmination Outcomes

Culmination outcomes rely on a clear itemization of inputs that, when applied to a given problem, produce certain measurable impacts (Arrow 2006). This is an attractive method for social policy design, as it enables policy makers to conceptualize fiscal, human and other resource inputs through a lens of applying a certain equation to a given problem. The increased demand for evidence-based social policy regularly reinforces this conceptualization by promoting more "scientifically rigorous" methods of policy study such as RCTs. RCTs are advantageous and often necessary methods for advancing clinical studies or for when the purpose of the study is constrained to the group under trial. These studies are often capable of attaining internal validity, and therein can empower critical impacts within the observed community (Deaton 2010). In the realm of social policy, however, an "internally valid" trial is only as informative as the borders of its own study. Without extensive—often very costly—replication of the trial across many communities, a social policy RCT cannot overcome the gap of external validity. Sen rejects the basis of reasoning on merely culmination outcomes, describing it as "ignoring the relevance of agencies, processes or relations" (Sen 2009: 217). Agents' processes are relevant because the act of choosing a certain pathway to an outcome adds greater relevance and understanding to that outcome that is often ignored. This extends beyond an individual institution being responsive to its own systems and influences: "of course the institutions themselves can sensibly count as part of their realizations that come through them, but they can hardly be the entirety of what we need to concentrate on, since people's lives are also involved"(Sen 2009: 82). This is Sen's argument for social realizations, in which it is our responsibility to recognize the consequential connections that "relate freedoms to obligations" (Sen 2009: 347).

The importance of social realization and identifying the impacts of choice to a process extends to all areas of evidence-based reasoning. Sen describes this reasoning of comprehensive outcomes as holding "extensive relevance to problems of economic, political and social behaviour whenever the act of choice has significance" (Sen 1997: 746).

A social policy impact evaluation necessarily falls within the realm of activities in which "acts of choice" have influenced pathways to an outcome. To frame this more concretely, this study observes the implementation and evaluation stages of a social policy programme as the critical stages in which the process informs comprehensive outcomes. In the case of social policy design, ignorance—either unintentional or wilful—of the social realizations comprising the full body of the policy's interaction with implementers and recipients alike, in favour of a study with a claim on "scientific rigour" simply on the basis of internal validity, means crucial interactions that either compounded or weakened a policy response are likely to be overlooked.

A close variant of the comprehensive evaluation is found in the realist evaluation approach. Realist evaluations seek to identify causal relationships by testing theories which detail how a certain programme initiates different mechanisms causing (desired) outcomes. Several systems are continuously at work, which interact with the other mechanisms and the context (or system) (Pawson and Tilley 1997). In order to do this, a realist evaluation takes into account how different underlying mechanisms are likely to interact with historical and cultural context, location, economic and political structures, participants and so on, to produce varying outcomes (White and Phillips 2012: 10).

A comprehensive evaluation looks at the effect of a policy at a cross-sectoral level. The comprehensive analysis of a policy should assess the impact of the policy on multiple outcomes and on the synergies between these. Several areas of economics already incorporate certain indirect effects such as reputation effects (Mui et al. 2002) but this does not make a comprehensive evaluation, since these indirect effects are still contextualized within the single frame of impacting a predetermined outcome. In reality, these indirect effects are the result of multiple interactions that remain not fully understood for the time being. Therefore, while a culmination evaluation can capture some of these indirect effects, the current methodology fails to address how to influence or monitor them.

Furthermore, a comprehensive evaluation recognizes that a given process occurred within a specific social, economic, political and cultural context that, while crucially informative, needs to be carefully and rigorously replicated under various contexts to determine the consistent impact pathways of a given intervention (Acemoglu 2010; Devereux and Roelen 2014: 1–2). This is a challenging reality to any policy design process, as it forces the stakeholders involved to accept that the effect of a policy will not be easily generalizable or clear.

In contrast to the comprehensive evaluation method, a culmination outcome evaluation is more confined to a linear description of input–output causality, enabling researcher and policy makers alike to see it as a mode of stimulate a more "scientifically rigorous" research process (Woolcock 2009). The result has been an over-emphasis on clinical study components including the importance of internal validity through tools such as randomization. Randomization should be promoted for a certain area of research. However, an emphasis on randomization in all studies is likely to fail to capture outcomes that are not easy to assess through (quasi-) experimental impact evaluations, such as women's empowerment (Sanson-Fisher et al. 2007). Randomized impact evaluations fail to explain the interactions and pathways that lead to an outcome and, by measuring averaged outcomes, once again ignore heterogeneous impacts (Olofsgard 2012: 1). Within any individual intervention, heterogeneous outcomes among individual agents are accepted as inevitable, yet are curiously ignored through the process of aggregated reporting (Olofsgard 2012: 1). In this way, RCTs may serve to reinforce a bias toward measuring outcomes that are immediately measurable and easily quantifiable.

2. Critique of Mainstream Evaluation Methods

In the field of impact evaluation RCTs and quasi-experimental methods such as propensity score matching (PSM), regression discontinuity design (RDD) and instrumental variables (IV) have become popular methods for researchers. They perceive such methods assign greater scientific credibility to their work and allow

policy makers to feel more confident about the evidence base guiding their decision making.

RCTs are perceived as a relatively fast, cost-effective way of evaluating a policy. They are particularly popular among non-governmental organizations (NGOs) seeking to evaluate the effectiveness of their programmes. Governments and international organizations are also increasingly making use of RCTs. Most notable among these is the World Bank which, after receiving criticism for being unable to base its programmes on social experiments, started using RCTs as a means of intervention evaluation (Deaton 2010: 426). The main goal of these evaluations is to construct a reliable counterfactual by finding a good comparison for the treatment group. Proponents of "experimental", rather than "non-experimental", methodologies see RCTs as requiring fewer assumptions to establish their own internal validity. Nevertheless, the common methodological approach for identification remains a problematic assumption under RCTs. Individuals make purposive choices about whether to comply with or enter into an intervention for a variety of latent factors unknown to the evaluators. In attempts to correct for unpredictable individual behaviour, the evaluators often institute a randomized assignment for programme participation as an instrumental variable. The result is a more anticipated average treatment effect (Ravallion 2009). However, this underlying corrective approach applies an exclusion restriction independent of actual choice by the individuals enrolled. Not only does a randomized instrumental variable fail to account for the individual's true motivations to comply with the intervention; it forces the assumption that the impact of the intervention will be the same for everyone. The failure to account for the possible heterogeneity of impact not only threatens the assumed internal validity of a randomized evaluation but it also undermines the external validity of the study (Ravallion 2009). Moreover, heterogeneous impacts are also linked to the inability of RCTs to explain the causal effect of intervention on the treatment group, as will be explained below.

Quasi-experimental methods are often referred to as "second-best" alternatives to experimental methods. The aim of the PSM approach, for instance, is to simulate the conditions of an experiment in which participants and non-participants are randomly assigned. The method consists of comparing treatment effects across participants and matched non-participants based on similar probability to participate in a programme. The propensity score to participate is calculated by means of a number of key variables that would determine programme eligibility. PSM assumes that all relevant difference between participants and non-participants can be captured by observable characteristics. Accordingly, selection bias is assumed to be based solely on observable characteristics and can therefore be corrected by PSM.

RDD methods consist of comparing participants and non-participants who are close to the eligibility criteria of the programme (slightly above or below the cut-off). As a consequence, this method assumes that any discontinuity at the cut-off can be attributed to the treatment.

With the IV approach, the selection bias on unobserved characteristics affecting the outcome is corrected by adding a new variable that is correlated with participation in the programmes but not correlated with unobserved characteristics. Limitations of quasi-experiments are mostly derived from their internal validity, as subjects assigned to the programme are selected according to predefined criteria. The validity of the results might then rest upon the quality of these criteria. However, RCTs present much higher internal validity as the subjects are randomly assigned.

External validity

There are considerable concerns with regard to the external validity of RCTs as the impact of an RCT is not necessarily replicable to a similar study in a different setting, or on a larger scale (see, for example, Sanson-Fisher et al. 2007). Since the impact of a programme on a larger scale is ultimately what policy makers would like to learn, the external validity, or lack thereof, should not be ignored (Ravallion 2009). Indeed, a typical RCT analysis produces results that are particular to the specific situation and conditions of the trial, but provides little evidence to support generalizing those results to, for example, an upscaled programme or another geographical area. Several features can be attributed to the methodology's lowered external validity: (i) RCTs focus on the impact of the treatment group while often neglecting the indirect effects on the control group, (ii) impact evaluations measure the impact of a policy in a partial equilibrium. When scaling up a policy, general equilibrium effects apply which could mean measured impacts on the scale of an RCT are offset when the policy is scaled up (Angelucci and De Giorgi 2009), and (iii) there is the possibility of diminishing returns, factor productivities and changes in prices. Other likely effects are composition effects and endogenous technology effects (Angelucci and De Giorgi 2009). The sum of these negative policy reactions has the potential to reduce or even to completely offset the measured positive impact of RCTs on a small scale in a relatively controlled environment.

Understanding the environment in which the respective RCT took place is therefore of vital importance for predicting the impact of a programme on a larger scale. When the environment is understood, and particularly those factors that influenced the policy's impact, the programme's effect in a different setting could be predicted. In support of this, Acemoglo (2010) argues that impact evaluations should make more use of economic theory, both in motivating and in explaining empirical work. He further argues for the estimation of so called "structural parameters"; that is, parameters to provide knowledge for understanding the effect of a policy within a given context. These structural parameters could provide external validly and "would thus be useful in testing theories on in policy analysis beyond the specific environment and sample from which they are derived" Acemoglo (2010: 2). Structural parameters can be useful guides in motivating and explaining empirical work. However, they too should be interpreted with caution, as policy making involves using key and robust implications from theoretical models rather than taking all the predictions of the model seriously.. An exogenous parameter in one evaluation, for example, may be inherently endogenous for another. To deal with the problem of external validity further, Acemoglo therefore argues for the use of cross-country variation and calibration of the results of previous studies. Several authors have, for instance, identified the long-term effects of historical institutions on current outcomes (Acemoglo 2010).

An absence of randomization in the selection of sites for the RCT is another threat to external validity. In practice, selection of the research location is often based on predetermined knowledge rather than truly being based on randomization. As a result, the claim of perfect randomization is questionable. If the sites where the pilot took place were randomly chosen, the theoretical external validity of RCTs would increase significantly and the impact might be more representative of the area in which the trials took place (Duflo et al. 2006: 72). Nonetheless, in some cases the locations where the pilots take place are subject to the willingness of participants rather than to chance. A possible solution to this problem would be to randomize at a larger scale, for example at

the village level rather than that of the household (Angelucci and De Giorgi 2009). However, despite these efforts, external validity is not guaranteed.

An important—but often neglected—aspect of external validity relates to the questions of who will be implementing the project when the programme is scaled up. Evidently, the implementation of the policy has a large influence on the actual impact, both at pilot level and in the case of a scale-up. Therefore, a comprehensive evaluation needs to take into account the implementation design, as is elaborated further in the next section. Nevertheless, even if the implementation design is accounted for in the impact evaluation, it does not necessarily mean that the effect of scaling up the project has been accounted for as well. When a project is expanded, it will often be implemented by a different actor (for example, the government) which might not have the same incentives, resources or knowledge as the original implementer (for example an NGO). Consequently, the programme's impact will, with other conditions remaining the same, also differ from the pilot to the national roll-out. Similarly, when a programme is scaled up, the targeted group as a whole may not have the same characteristics as the treatment group in the initial pilot. As a result, a randomized experiment might under- or overestimate the effect of the programme when implemented at a larger scale (Ravallion 2009).

Limited scope

Conventional, scientifically orientated evaluations tend to examine a relatively small number of variables in analysing the potential impact of an evaluation. Usually, these include more easily measureable quantitative indicators (Devereux and Roelen, 2014). More qualitative factors, such as improved social cohesion or beneficiary stigmatization, are often not measured unless they are the focus of the evaluation. However, these outcomes—positive or negative—may be very important to the feasibility and understanding of impact of the programme (Devereux and Roelen, 2014). Most quantitative variables in mainstream evaluations can be misleading as they only measure the average treatment effect on the treated, and the average effect of the "intention to treat", as people are not forced to accept treatment. Qualitative information may be valuable in interpreting these results when they overlook a positive impact for some beneficiaries, and a negative impact for others (Ravallion 2009). This phenomenon is related to the problem of heterogeneous impacts discussed earlier.

Short-term outputs or intermediate outcomes are also easier to evaluate than broader outcomes and objectives, and are therefore more readily evaluated. Examples of such outputs include the number of children going to school, or the number of hospitals per thousand inhabitants. These are very different from outcomes that are hard to measure, such as educational performance. For example, the number of enrolled children is only an intermediate output toward the final variable measuring education performance. Another example of an intermediate outcome is the amount of paid work created in the case of a public works programme. The objective of such interventions is, among other things, to make people better able to cope with shocks, to provide them with work experience and to promote their social inclusion. These outcomes are more complex in nature and thus require more complex methods for assessment. Additionally, negative and positive outcomes that might not be the immediate goal of the intervention should be evaluated (Devereux et al. 2013: 8).

RCTs often fail to evaluate the effect of policies on institutions. While well-functioning institutions are crucial for economic development, the change is often spread over a longer period of time, complicating the ability to quantify the outcomes. As a result, the

effect of policy on institutions is often not taken into account (Olofsgard 2012: 1). Consequently, many studies that ignore institutional and political factors, and thus general equilibrium effects, may come to the wrong conclusions (Olofsgard 2012: 1). Not only does this imply that part of the benefit (or cost) of a policy is ignored, but it also suggests that the treatment effect might be underestimated if the change in institutions affects the comparison groups. If, for example, certain communities are given cash transfers under a social protection programme, the government could decide to move resources to villages where these benefits are not distributed. An indirect effect of a transfer could thus be a shift in government resources from one area to another. If the area that benefits from this shift is used as a counterfactual, the treatment effect will be underestimated (Ravallion 2009). This example proves how important it is to monitor the effect of policy on other related government policies to obtain an unbiased treatment estimate.

In addition to the limited number of outcomes that mainstream evaluations aim to measure, they also tend only to examine the effect on the treated households or individuals. Both negative and positive effects on the non-treated are generally ignored, even though they are valuable to assess the overall impact of the programme (Angelucci and De Giorgi 2009). Social impacts are also often ignored in impact evaluation studies. This implies that an intervention, for example a public works programme, may be misinterpreted when the effect on a simple quantitative variable such as income is positive while local businesses fail to find labour because their wages are comparatively lower. Likewise, an intervention that does not show an immediate significant result on the respective main impact indicator might not be evaluated positively, even though this intervention contributes positively to, for example, lowering crime rates.

The last limitation is that RCTs are not only able to measure a limited number of outcomes. Each type of policy outcome will have different randomization requirements to guarantee the internal validity of the pilot. RCTs aimed at improving savings rates will require individuals who do or not save, whereas a pilot targeted at improving shock resilience for rural households will have a different set of requirements. The more complex a policy (that is, one incorporating different policy areas), the harder it becomes to guarantee absolute randomization of treated individuals or households (Olofsgard 2012: 10)

Causality

Ideally, RCTs measure outcome by comparing the effect of a policy on a treatment group and a control group. The treatment group receives the benefits stipulated in the policy and the control group does not. In approaches such as the Difference-in-Difference methodology, PSM or IV, a treatment effect is calculated by comparing the changes in the respective variable of the treatment group to the control group. These methods are designed to measure differences between the treatment and control groups but are unable to describe the causal pathways. Counterfactuals associate a single cause with a given effect, but cannot explain how the effect is produced. Similarly, these evaluation methods do not necessarily provide information on the social or political impact of a given intervention (Devereux et al. 2013.) Authors of quantitative-based studies often prove a possible explanation of causal relationship but, more often than not, fail to produce evidence to support their claim(s). Therefore, quantitative studies are ideally complemented by qualitative follow-ups to identify the causal relationship for differences between groups and causalities related to their environment. The understanding of a causal pathway contributes to interpreting the results, helps to

replicate the results as the necessary conditions can be identified and provides evidence to improve the existing policy. Devereux et al. (2014: 3) refer to this as the "learning loop", a process where evaluators and implementers amend programmes to reflect the results of holistic policy evaluations.

As discussed above, quantitative randomization evaluations generally only provide information about the average impact of a policy. Deaton (2010: 426) points out that this hides the distribution effects of a policy intervention across households. Using this basic method, only the distribution of the outcome value and the treatment and control groups are known. This could be particularly deceptive when a positive mean treatment effect is the result of a low number of positive outliers while the general trend is negative. Generally, RCTs are designed to give information about the mean treatment effect, but are not able provide information about more detailed features of the distribution which may be equally important in the policy-making process (Banerjee and Duflo 2009). A way of learning about the distribution of impact(s) is to stratify the sample from the start in order to estimate the impacts by sub-group(s). The sample of household (or individuals) can be divided into different groups (such as female and male) for which the effects are estimated separately. Such a methodology requires larger datasets in order to prove significant impacts on the defined subgroups[1]. Combining the treatment with other socioeconomic characteristics, and using quintile regression, are other ways of evaluating heterogeneous impacts (Daidone et al. 2012). In other words, by dividing the sample into large enough sub-samples by sociodemographic or socioeconomic characteristics increases the value of the analysis.

Interconnected interventions (that is, where the targeted populations benefit from diverse programmes) can be complex to evaluate as they should not evaluated separately. Examining the effect of a programme without taking into account the other benefits households or individuals receive would give a highly unrepresentative treatment effect. However, the reality is that "development interactions" (…) operate in a complex environment which makes isolating their impacts from the array of confounding factors highly challenging" (Devereux et al. 2013: 19) as the interaction between programmes leads to multiple causal relationships. Devereux et al. (2013: 18) use the concept of "recursive causality" to describe these connections and causalities. Because of this complexity, a comprehensive analysis should assess the impact of the intervention on multiple measures and the synergies between them (Stern et al. 2012).

Long-term effects

Mainstream impact evaluations prioritize short- (or medium-) term effects over longer-term effects in an intervention. Stakeholders involved in the evaluation process are under pressure to produce results that can be verified immediately due to limited time, finite financial resources and heightened political demands (Ravallion 2009). Consequently, the evaluations focus on outcomes that are readily quantifiable and observed, overlooking more structural developmental impacts that are more difficult to measure and evaluate (Ravallion 2009). These practices draw from the key assumption that interventions such as social protection programmes have a linear and successively increasing impact. However, interventions can produce different outcomes across their programme lifecycle. There is a wide range of possible impacts over time: the trajectory could be "J-shaped" (early, worsening outcomes before their gradual improvement), "S-shaped" (slow initial change, rapid change, then plateau) or it could follow a "step function" (extended period of no change before a sudden exogenous shock which

[1] Comprehensive outcomes should also take into account results that cannot be proven statistically significant, as will be discussed in section 4.

culminates into a rapid "snowballing" effect (Woolcock 2009). Due to these entrenched methodological priorities, current evaluation practices fail to grasp the larger context and the key structural impacts therein.

The implications of the impact evaluations' emphasis on short-term effects occlude a more rigorous approach that takes longer-term effects into consideration. These long-term effects are associated with more complex outcomes such as shifting gender norms, changing economic and political structures and institutional development (Woolcock 2009). While it is more difficult to quantify their causal effects, these long-term effects have important bearings on the success or failure of a social protection intervention. Evidence demonstrates that programmes aimed at empowering marginalized groups such as women only record positive results after years of initial community backlash and gradual institutional change as societal values evolve (characteristic of a "J-shaped" impact trajectory). RCTs conducted in early moments where the initial impact is negative can report inconclusive outcomes, wrongfully supporting the intervention's discontinuation with a premature impact (Woolcock 2009). Impact evaluations would clearly benefit from refocusing on broader institutional outcomes with increased sensitivity to impact trajectory and the appropriate reallocation resources with this methodological priority in mind.

Ethical considerations

Emerging literature has cast doubts on the ethical standards routinely employed and advocated for in the use of randomized experiments on human subjects. Unlike RCTs for most clinical programmes, where subjects are "blinded" so they are unaware of which group they have been randomly assigned to, participants and non-participants in social interventions are fully aware of their allocation to either control or treatment group (Barrett and Carter 2010). Denying the control group access to programme benefits can cause emotional distress, particularly for at-risk groups in need of the intervention. More importantly, interactions between control and treatment group can result in potentially negative group dynamics, affecting the differences in outcomes between the groups. Ethically, the methodology would violate the "do no harm" cornerstone principle of social research in stimulating adverse effects on the local population (Barrett and Carter 2010). Statistically, differences in outcome for the treatment and control group may be partially reflecting the change in behaviour caused in both groups, thus overestimating the difference in outcome. It may also lead to people participating in several programmes at the same time in order to increase their chances of being in the treatment group in at least one programme (Barrett and Carter 2010; Wolff 2000).

3. Toward Comprehensive Evaluation

Given the limitations of mainstream methods in social policy evaluation, as discussed above, it is important for policy analysts and implementers to broaden the measure of outcomes. Comprehensive outcomes evaluation empowers a study to map causal pathways that brought about culmination outcomes, while providing more complete information on the nature of heterogeneous impacts and crucial interactions that either compounded or weakened outcomes.

A comprehensive analysis should also include an analysis of the intervention's impacts across multiple sectors of interest. For instance, an education-based intervention will have outcomes that impact a variety of other institutional sectors and community

interactions, such as health outcomes, income security and the child's proclivity to engage in risky behaviour, to name only a few (Holmes et al. 2012). The interlinked pathways to multi-sectoral outcomes in these cases have been shown to create mutually reinforcing outcomes that can significantly influence the culminating outcomes of a policy intervention. Without the mapping of such multi-sectoral pathways, policy analysts and implementers are (i) falsely assigning too much credit of their immediate intervention to the final outcomes, and (ii) missing key opportunities to further promote the outcomes that, if identified, could be enhanced by reinforcing key sectoral linkages (Samson 2013).

The majority of current mainstream research focuses on the short term (Ravallion 2009), but some of the most substantive benefits—and costs—emerge over the long term (Lutz et al. 2014). The lack of a long-term lens on intervention outcomes further creates a distortion in the types of social policy interventions pursued. Some of the more significant social protection interventions, which would reap significant "returns on investment" in the long term through far-reaching and reverberating developmental outcomes, are overlooked due to limited assessment methods employed in culmination outcomes evaluations (Lutz et al. 2014).

As it currently stands, the majority of social protection policy evaluations originate from supply-side sources: the implementing or funding institutions themselves. However, comprehensive evaluations should be demand-driven rather than supply driven. While the funding or implementing institution—often a government department, international development partner, implementing NGO or a combination of these—is likely to have the greatest interest in an evaluation, the supply-driven nature of evaluation is also likely to limit its scope of measurable outcomes.

Increasingly, policy objectives are formulated so as to emphasize multi-sectoral collaboration, or interaction between the policy or intervention and other processes or sectors. This is a promising trend that should be extended so that evaluations focus not only on the process of engagement, but also on the methods of multi-sectoral outcome measurement and reporting (Devereux and Roelen 2014). The following subsections explore some of the methods that can better map these multi-sectoral linkages and can enable comprehensive outcomes evaluation.

Mixed methods research

Mixed methods research employs both quantitative and qualitative methods. Qualitative methods aim at understanding processes and behaviours as perceived by the participants and non-participants of a given programme. Through in-depth interviews, focus group discussions, observational reporting, various surveying techniques and other methods, qualitative reporting provides critical insights into the social, economic, political and cultural context within the observed space. Qualitative methods further enable more exploratory research that can describe detailed interactions between actors difficult to quantify, allow for in-depth study of anomalies in quantitative results, and offers a crucial tool for describing complex and interactive pathways based on individual experiences (White 2008). Qualitative methods are more inductive and subjective processes than quantitative methods and therefore have natural limitations to a complete outcomes analysis. For this reason, combining qualitative and quantitative methods can prove to be a valuable complementary method for evaluating comprehensive outcomes (Devereux and Roelen 2014).

Mixed methods research is already widely employed in the evaluation of social policy and interventions. Adato (2008) evaluates social protection programmes in Latin America, where both quantitative surveys and ethnographic models were used to evaluate conditional cash transfer programmes. The South African Child Support Grant Impact Assessment[2] employs non-experimental quantitative approaches combined with a multi-method qualitative component (DSD et al. 2012). While these and similar studies have been identified as cases of good practice (Devereux and Roelen 2014), there is still a methodological evaluation gap on how different complementary components interact through the research process and, in the case of small n studies, best verify findings.

There are commonly employed approaches to mixed methods research for social policy interventions. Carvalho and White (1997) identify three mixed methods approaches (and specifically integrating methods) for policy research: verification, triangulation and enrichment of findings. Greene, Caracelli and Graham (1997, in Stern et al. 2012) propose triangulation, complementarity, development, new start and expansion. Each of these approaches shares some common components but can uniquely service different objectives based on the intervention type and research context.

Quantitative research methods should be informed by insights obtained from qualitative methods and, contrariwise, qualitative tools should probe the reasons and causal pathways behind quantitative findings as they arise. Mixed methods research approaches are an indispensable tool to an evaluation of comprehensive outcomes.

Local economy-wide impact evaluations

In responding to the need for a broader set of outcomes for the evaluation of social programmes, researchers are applying the Local Economy-wide Impact Evaluations (LEWIE) methodology. This attempts to understand the full impact of an intervention on local economies. Generally, RCTs do not capture how interventions affect the control group, focusing exclusively on the treatment group and disregarding the "spill-over" effects into the community. Moreover, RCTs do not estimate what would happen if the project is scaled up to the larger community, confining the assessment to spatially static conclusions. The LEWIE methodology emphasizes the intervention's impact on the beneficiaries and the non-eligible and the process through which the observed effects occur (Taylor 2013). Daidone et al. (2012: 20) employ the LEWIE and mixed methods research approach to evaluate the impact of cash grants on household behaviour. Their study argues that a cash grant will affect households differently depending on the household composition and state of the local economy. Taylor (2013) uses the LEWIE methodology to demonstrate that cash transfers to beneficiaries in sub-Saharan Africa contribute to non-eligible households and largely rural economies by stimulating production activities for both control and treatment groups (Taylor 2013). By assessing both arms of the experimental groups and using a more comprehensive set of outcomes than RCTs, the LEWIE methodology is better able to answer questions regarding an intervention's impact on the wider environment across a range of programmes scales.

[2] Commissioned and funded by the Department of Social Development (DSD), the South African Social Security Agency (SASSA) and the United Nations Children's Fund (UNICEF) South Africa.

Lessons from impact evaluations with small samples (small n evaluations)

The focus of most mainstream impact evaluations is on outcomes that can be shown to be internally valid and statistically significant. Often, though, the outcome of interest has to be evaluated using a small sample (n) impact evaluation, for which traditional methods utilizing statistical significance are not useful (White and Phillips 2012). Comprehensive evaluation methods that evaluate a broad range of outcomes are more capable of accommodating multi-component analyses that cannot be evaluated using traditional methods most associated with large sample designs. Also, methods proposed for small sample impact evaluations can complement large sample impact evaluations.

Causality—Theory of change

Regardless of the number of treated households or individuals (*n*), it is very useful to draw on a theory of change. Such a framework can help to identify different (intermediate) outputs and outcomes. Moreover, a well-founded theory of change can help in identifying the causal relationships between these factors (White and Phillips 2012). The realist evaluation is noteworthy in this regard (Pawson and Tilley 1997). A realist evaluation seeks to identify causal relationships by testing theories that detail how a certain programme ignites different mechanisms which in turn lead to the (desired) outcome. As explained above, at any point of the implementation stage a number of programmes interact with each other: different mechanisms and socioeconomic and cultural environments. An example of such a realist evaluation is the study by Marchal et al. (2010: 787) on the impact of a number of different management practices on the performance of an urban district hospital in Ghana. The authors used a theory of change to predict causal pathways and outcomes, together with theories predicting outcomes given more specific contexts. The findings of this study were then also incorporated in a theory of change to create a loop which enables continuous improvement of the theory to test causal relationships.

A second approach to identifying causal pathways/theories of change is "process tracing". This qualitative research "attempts to identify the causal process—the causal chain and causal mechanism—between a potential cause or causes, e.g. an intervention, and an effect or outcome, e.g. changes in local government practise" (Georgen and Bennet 2005 in Hughes and Hutchings 2011: 7). It short it thus involves proving or disproving how the specific ways a particular cause produced a particular effect. The method was used, for example, by Oxfam to evaluate several interventions such as the "Fair Play for Africa" campaign (Hughes and Hutchings 2011). The process tracking approach was able to identify, among other findings, that the success of the campaign in Malawi was partially due to another campaign that was set up beforehand. The observed change was thus caused by the effect of an earlier policy, in this case a marketing campaign, rather than the researched policy. Researchers came to this conclusion by considering alternative, competing explanations until the explanation most supported by the data remained. Process tracing thus helps to identify the theory of change by eliminating less likely alternatives.

A third approach or theory is the "contribution analysis". This method encompasses creating a "plausible contribution story" that is based on a theory of change and which attempts to prove that the several steps set out in the described theory actually took place. A contribution analysis also tries to incorporate several causal pathways in one theory and to evaluate the relative contribution of each of these pathways (Mayne 2011).The approach thus focuses on identifying the contributions made by the intervention and other internal and external factors to the observed outcomes through

providing a better understanding of why the observed results have occurred or failed to occur. The approach is thus useful in shaping a theory of change and should not be used for uncovering implicit theories of change. Contribution analysis does not provide a definitive proof of a theory but adds to the body of evidence by examining the contribution of different factors to the impact.

The last approach to develop a theory of change is the Qualitative Comparative Analysis (QCA). QCA makes use of in-depth case studies, and tries to find causal patterns for different cases. The idea is that similar combinations of causes (or conditions) can lead to different outcomes, and that certain outcomes can result from different combinations of causes. QCA makes use of data matrices in which different possible causes are coded; these matrices are transferred into a "truth matrix". In the truth matrix, different combinations of causes that lead to the same outcome are summarized (Berg-Schlosser et al. 2009). It is a theory-driven approach and can reinforce existing theories of change but is unlikely to develop new ones.

Participatory evaluations

Participatory evaluations engage the key stakeholders of an intervention to actively develop the evaluation and implementation phases in partnership with one another. Project staff, community members and funders determine the indicators of the programme's evaluation; work jointly toward data collection and analysis; and prepare final reports. One example of a participatory approach is the "Most Significant Change" approach (Davies and Dart 2005). For this approach, several stories are collected by asking the participants questions about changes that occurred in their lives and how they perceived them. These questions are asked several times throughout the programme. From these stories the most significant ones are selected, through a process of passing the stories upwards (from fieldworkers to researchers), whereby the selection of significant stories becomes smaller at every level (Davies and Dart 2005).

Participatory evaluations aim to further develop indicators and outcomes that are not constricted by easily quantifiable outcome measures primarily employed by RCTs. Use of local voices in the research process underlies the importance of a mixed method approach where quantitative and qualitative methodologies lead to a more integrated evaluation. Tapping into the local reservoir of knowledge, the methodology allows reflection on the beneficiaries' programme experience and for a more in-depth study of issues underscored by involved stakeholders. Moreover, participatory evaluations shy away from average impact estimates which are routinely utilized by RCTs. Such a limited measure fails to capture the complex institutional measures highlighted by participants. Rather, participatory evaluations are better able to estimate distribution of impacts by answering questions about the programme effects on beneficiaries. Progress can be captured in real time through data verification from key stakeholders, establishing possible causal pathways.

Combined effects

This paper has already described the value of mapping causal pathways across agent interactions to better evaluate outcomes across various sectors. The subsequent policy step is to use better knowledge on multi-sectoral pathways to create cross-institutional interventions. These interventions would be designed to take advantage of mutually reinforcing pathways to produce combined effects. There are already examples of such interventions, such as the Livelihood Empowerment Against Poverty (LEAP) programme in Ghana. LEAP beneficiaries are provided both an unconditional cash grant

as well as access to free health insurance, covered by the National Social Health Insurance Scheme. This follows from LEAP's "integrated social development approach" (Handa et al. 2013).

Handa et al. (2013) describe the multi-sectoral outcomes of the LEAP project. While neither a cash grant of the target scope or value currently provided, nor the existence of health insurance would be sufficient to significantly assist households in coping with health shocks, the combination of these two policies has the possibility of offering long-term social protection to some of Ghana's most vulnerable. Furthermore, the cross-institutional approach has allowed each of the relevant implementing agencies to provide this social protection programme that, were they each to utilize a silo approach, would not be cost effective (Handa et al. 2013). Poverty is addressed not only by removing the monetary constraints that poor households face, but by helping to mitigate health shocks. By evaluating not only the effects of both programmes, but also how the two complement one another and thus offer a variety of impacts to the population, combined effects programmes such as this one are able to address a variety of overlapping socioeconomic issues within one study.

4. Proposing a Framework

This paper has described the state of social policy evaluation by presenting two archetypes of evidence reasoning: culmination and comprehensive outcomes. It has thus far described some of the key conceptual and methodological limitations of a culmination approach including the limited scope of outcome measures, the misleading and equivocal results that arise from aggregative and averaged impact assessments, and the fundamental concern that much of a social policy's impacts cannot be clearly described through clinical study design methodologies.

This paper has also expanded on the opportunities for expanding the evaluation approach within some of the culmination outcomes methods, while also introducing increasingly important examples of comprehensive outcome evaluation approaches. This section proposes a guiding framework for comprehensive evaluations that more closely appraises Sen's "social realizations" by enabling multi-sectoral mapping of causal pathways. Based on the previous sections, some essential design components of comprehensive evaluation can guide future evaluations:

A comprehensive evaluation should evaluate the linkages and interactions between different policies, actors and institutions.

A comprehensive evaluation should be able to map both immediate as well as reverberating outcomes that both compound and weaken the impacts of a given intervention.

A comprehensive evaluation should incorporate an assessment of the social, economic, political and cultural context within a given intervention area, and attempt to portray how this context may influence the mapped outcomes.

A comprehensive evaluation should be, ideally, demand-driven and should accommodate both short- and long-term evaluation methodologies. These methods should promote fiscal analyses and "value for money" evaluations that consider all social policies to have multi-sectoral and long-term "returns on investment" that emerge through far-reaching and reverberating developmental outcomes.

While comprehensive outcomes evaluations can and should be applied to a wide variety of policy strategies, this paper has focused on social protection interventions. The effectiveness and relevance of a comprehensive outcomes evaluation approach can be realized most fully if the various social protection strategies for a given country are integrated into a broader planning framework in which multi-sectoral institutions are already engaged in the kinds of interactions that will allow easier identification and monitoring of interacting outcomes (Samson 2013).

Based on the previous sections, two essential methodological strategies stand out as promising tools for future comprehensive outcomes evaluation:

1. Mixed methods research and integrated evaluation. The current emphasis on RCTs (experimental evaluations) needs to be further expanded to broader acceptance of mixed methods research and integrated evaluation. As discussed in the section above, critiquing mainstream methods, RCTs face many limitations and the literature critiquing randomization as the gold standard is extensive and growing (see for example Ravallion 2009; Deaton 2010; Acemoglu 2010). Mixed methods research is gaining widespread acceptance as a more complete approach for relevant evidence-building (see, for example, Adato 2008; Devereux et al. 2013). Complementing quantitative methodologies with qualitative approaches enables a better understanding of both the outcomes and the pathways leading to them (Devereux et al. 2013). Using mixed methods contributes to making an impact study more policy relevant and contextualized (White 2008), and, ultimately, more broadly useful to future multi-sectoral interventions.

2. Multi-sectoral integrated approaches to evaluation. Policy makers are increasingly recognizing the importance of multi-sectoral interventions as promising paths to more efficient and effective social policy impacts. However, complex evaluation frameworks that can identify and assess the most efficient and effective combinations of interventions are still emerging. A more robust comprehensive outcomes evaluation methodology that can assess the multi-sectoral pathways by which outcomes are strengthened or weakened is required to provide the evidence for multi-sectoral policy design.

Several linkages between different impact areas demonstrate promising mutually reinforcing properties. A cross-sectoral cost–benefit analysis includes benefits from these different sectors, such as health care and schooling. If allocation of resources results from fragmented single-sector evaluations, more effective and long-term investment solutions are less likely to be explored and, ultimately, funded (Lutz et al. 2014).

Donors and key stakeholder institutions can be reluctant to use the cross-sectoral cost–benefit analysis since it could mean that they have to finance part of a project that they would not have financed following a silo analysis. However, there is growing evidence to show that implementing cross-institutional approaches such as co-financing do have positive returns, using either cost-effectiveness (Remme et al. 2014) or "return on investment" (Claxton et al. 2007) models.

The macro framework for evaluation

A move toward more integrated and comprehensive evaluation requires methodological change at both macro and micro levels. The global social protection sector demonstrates

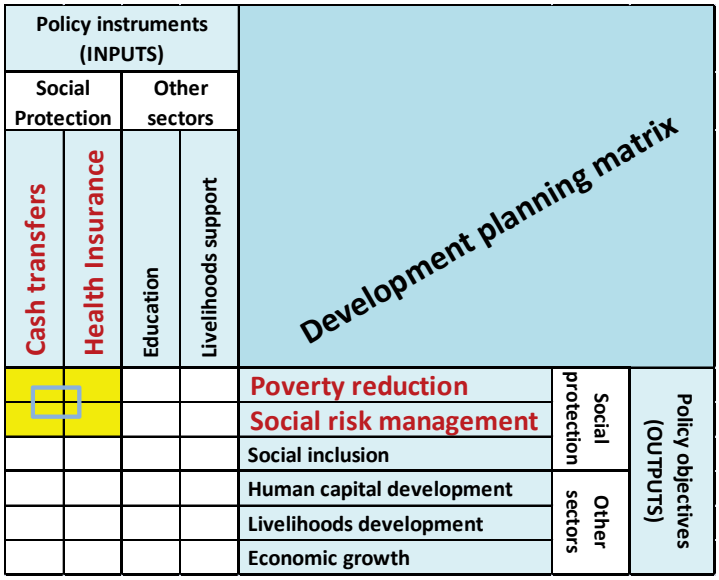
a macro trend toward more integrated and comprehensive approaches. Increasingly, developing country government ministries managing socioeconomic planning processes take responsibility for integrating comprehensive social protection responses into national development plans. This holistic approach to development planning recognizes that policies strengthening social protection's natural tendency to promote livelihoods and foster pro-poor and inclusive economic growth and development yield the greatest impact when coordinated within a larger planning framework (Samson 2013: 77).

Brazil demonstrated early leadership by incorporating inter-ministerial initiatives into the *Bolsa Familia* programme (MSD 2007). Brazil's social policy development support to a number of countries influenced policy development around the world (OECD 2013: Chapter 6). For example, Ghana's LEAP programme links a cash transfer programme with access to social health insurance benefits, providing more comprehensive social protection and strengthening the impacts of both interventions; see FAO (2013: 1-2). Mozambique has further developed its social protection system by integrating inter-ministerial initiatives to promote livelihoods and employment by "considering broader macro-economic areas for social investments to raise overall living standards (such as in agriculture, food security and employment-generating activities)" (UNICEF 2011). The development planning approach is strengthening multi-sectoral interventions and reinforcing multidimensional impacts in Bangladesh, Cambodia, Indonesia, Nepal, Rwanda, South Africa, Tanzania, Uganda and other countries (Samson et al. 2011; Samson 2012).

The framework supports both planning and impact assessment—recognizing the importance of both *ex ante* and *ex post* evaluation. Planning agents within governments employ the framework to balance social and economic investment priorities to jointly achieve the primary policy objectives, which. These can include the promotion of equity, reduction of poverty, increasing employment, and the strengthening of pro-poor and inclusive economic growth and development. The planning process builds linkages within specific sectors (such as the mutual strengthening of cash transfers and social health insurance within the social protection sector, as in Ghana) and across sectors (such as designing social protection programmes to maximize education, health, nutrition and livelihoods impacts). The matrix shown in Figure 1 illustrates a stylized model of the planning process, drawn from the development planning process Uganda employed in 2010.[3]

[3] The figure is only illustrative—Uganda's actual matrix, for example, had hundreds of columns and dozens of rows.

Figure 1. An illustration of the development planning approach



The framework reinterprets the input–output matrices employed in the development planning models of the 1960s, departing from the classic model by defining "inputs" as the set of public (and sometimes private) instruments, programmes and policies that enable the government to achieve priority "outputs", which are defined in this approach as the achievement of national policy objectives. The framework's innovation is the emphasis of the importance of "intra-sectoral" and "inter-sectoral" linkages. For example, intra-sectoral linkages within social protection reflect the mutual strengthening that results when cash transfers finance the contributions of otherwise destitute households for social health insurance. This in turn protects the members from the catastrophic health shocks from which social transfers alone provide inadequate protection. The combination of cash transfers and social health insurance works much better to reduce poverty and vulnerability than the individual instruments on their own (Samson 2013).

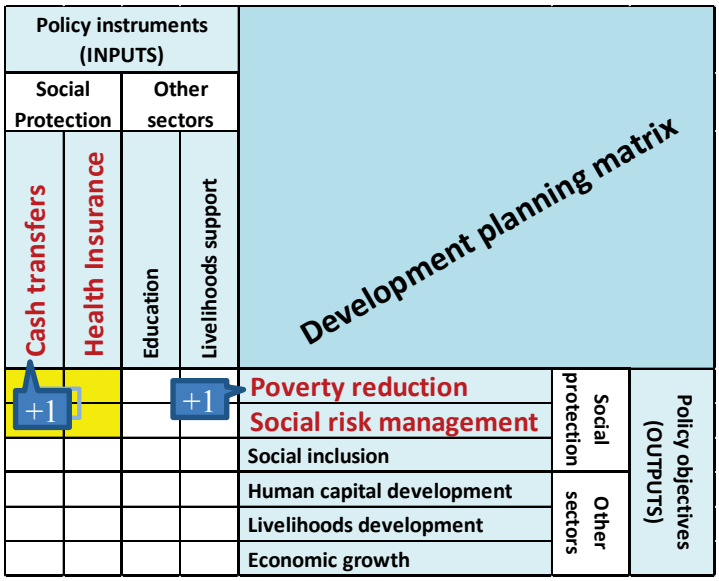
The matrix also illustrates inter-sectoral linkages. For example, social protection instruments strengthen outcomes outside their own sector, deepening human capital, strengthening livelihoods development and broadly promoting pro-poor and inclusive economic growth. Many social protection instruments work by expanding poor people's access to markets. Cash transfers expand the effective demand for market goods and services, enabling people to meet their basic needs while strengthening livelihoods engagement and stimulating economic activity. Likewise, a whole range of policy sectors can effectively strengthen the achievement of social protection objectives: education builds human capital and effective tackles the intergenerational transmission of poverty, livelihoods programmes strengthen household resilience and reduce vulnerability, and economic reforms can create opportunities that enable poor households to lift themselves out of poverty.

The essential element of the development planning framework for social protection is a national coordinating mechanism that plans, prioritizes and integrates the set of public policies and practices (including those of the social protection sector). This mechanism increasingly includes a national development plan coordinated by a national planning institution (such as a National Planning Commission). These comprehensive and integrated planning approaches strengthen both impact and efficiency by increasing the likelihood of achieving the priority policy objectives while reducing costs and risks. In

addition, practical development plans reinforce credibility in the government's strategy, enabling the government to expand its policy options.

The main gap in this macro evaluation framework is the measurement of conditional rates of returns. Consider the investment of a billion dollars in a social cash transfer programme. In a single-sector/single-outcome evaluation, reflecting the optimal assignment of cash transfers to the poverty reduction objective, a culmination evaluation would quantify the future impact on poverty reduction. The matrix can illustrate the hypothetical benefit of, for example, USD1 billion in terms of poverty reduction (Figure 2). The single-sector intervention's return is relatively low because, for example, cash transfers on their own have limited impact on the management of catastrophic health shocks. The conventional rate of return analysis discounts the future billion dollars in benefits, so the social rate of return is negative. A similar analysis holds for the investment of 1 billion dollars in a social health insurance programme. One billion dollars invested today might hypothetically yield a 1 billion dollar future return in managing the major risk that poor and near-poor households face in many developing countries—a catastrophic health shock. Again, the single-sector intervention has limited impact because very poor households often fail to access social health insurance programmes, and then discounting the billion dollar future return generates a negative social rate of return.

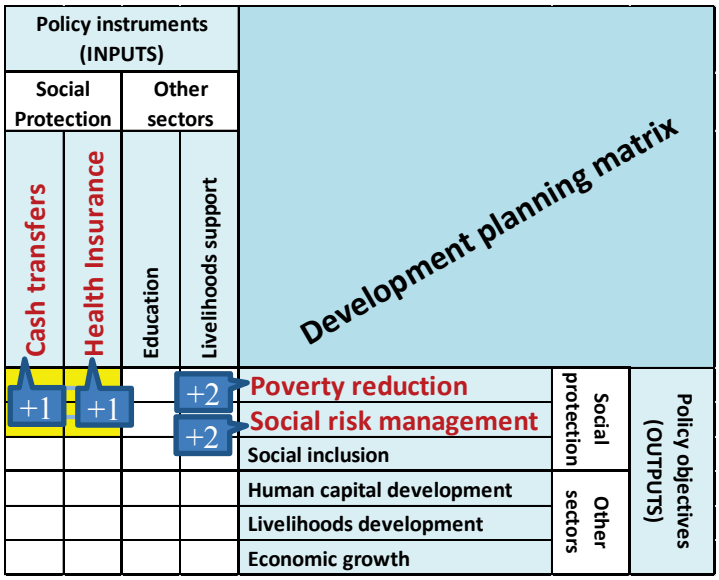
Figure 2. An illustration of a single-vector intervention



A multi-sectoral approach may be more effective and more efficient. Combining cash transfers with social health interventions improves the effectiveness of both interventions (as illustrated in Figure 3). Cash transfer initiatives better enable poor households to access social health insurance programmes, and the health benefits protect poor households from the catastrophic shocks which are too great for cash transfer interventions to provide adequate protection. The result is a greater impact on both of the priority outcomes. The rate of return to a cash transfer programme depends on the level of investment in the social health insurance initiative, as well as on the pattern of investments in the entire range of the government's social and economic investments. The same is true for the rate of return to a social health insurance investment. In order to analyse a manageable problem, conventional evaluation approaches generally assume the investments in complementary interventions are unchanged. As a result, these frameworks usually are unable to provide evidence

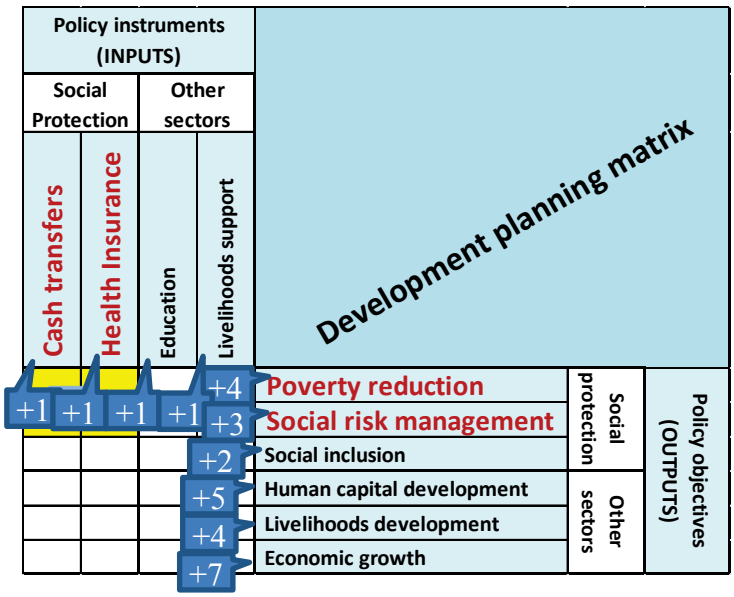
enabling the government to optimize the mix of investments and benefit from synergies across sectors.

Figure 3. An illustration of a multi-sector approach



The problem generalizes beyond the two-by-two case (two instruments and two outcomes) illustrated in Figure 2. Policy makers demand evidence on the trade-offs in returns to multi-sectoral complex interventions, understanding that the incremental return to an additional investment in one sector depends on the pattern of investment in every other sector (as illustrated in Figure 4). Simple culmination evaluations cannot answer this question, since they usually focus on a single intervention in one sector, or at most a few complementary or alternative treatments. Policy makers requires evidence on the impact of different states of a socioeconomic system, and this defies RCT: one cannot randomize countries. The benefits to complex evaluation are twofold: (i) when complex interventions benefit from substantial interaction effects across sectors, only a complex evaluation can measure the additional synergy benefits and illuminate the complex trade-offs in combining the multiple initiatives required to generate the optimal impact, and (ii) complex evaluations capture the entire spectrum of benefits, avoiding the undercounting associated with single-sector/single-outcome trials.

Figure 4. An illustration of a mix of investments and synergies across sectors



The micro framework for evaluation

The shift toward more integrated and comprehensive evaluation also requires methodological change at a microlevel. The conventional micro- evaluation approach of an RCT—or, in social policy terms, an experimental pilot—tends to bias toward single-sector/single-outcome trials. The conventional evaluation approach is best suited to trials where implementers know the precise parameters of the interventions in advance, and these can be implemented homogenously across the treatment group(s). RCTs often fail if the treatment must be adjusted during the course of the trial. RCTs are better suited to the objectives of culmination evaluation rather than the broader process of evidence-building that often requires a learning-by-doing approach. In contrast to an experimental evaluation, an evidence-building evaluation allows for greater flexibility in both design and implementation. The evidence-building framework sacrifices some risk of internal validity for greater potential variability in the intervention and stronger policy relevance. Table 1 contrasts key characteristics of experimental and evidence-building evaluations. Both types of evaluation share common tools. For example, both can use randomization techniques to map out the counterfactual. However, experimental evaluations are generally designed to maximize internal validity, often constraining the intervention to a limited and constant set of outcomes and treatments, given the typical resource constraints that limit any evaluation initiative.

Most RCTs, or experimental pilots, have simple interventions (treatments) testing simple single-sector outcomes. These interventions are usually constant (static) over the course of the trial. When the questions policy makers ask can be answered with simple static interventions, RCTs provide robust and cost-effective evidence.

Evidence-building evaluations aim not only to test whether an initiative works, but also to help design the optimal combination of interventions. Evidence-building evaluations tend to have more broadly represented samples, to increase the external validity of the assessment. The treatment in an evidence-building pilot can change over time, as learning-by-doing leads to an improvement in the technology for delivering impact. The result is an increased variability in the effectiveness of the intervention, which may reduce statistical power.

Table 1. Experimental versus evidence-building evaluations.

Characteristic	Experimental evaluation	Evidence-building evaluation
Objective	Simple	Complex
Treatment type	Simple	Complex
Treatment dynamic	Static	Dynamic
Validity focus	Internal (rigour)	External (relevance)
Evaluation type	Culmination	Comprehensive
Quantitative method	Simple or conventional multiple treatment RCT	Mixed methods including choice of non-experimental approaches or layered RCTs
Ethical approach	Unethical not to experiment	Do no harm

There are also differences in ethical approaches that experimental evaluations and evidence-building evaluations adopt. Howard White, Executive Director of the International Initiative for Impact Evaluation (3iE), makes the case that "the really unethical thing is not the withholding the program, perhaps temporarily, from some group. The really unethical thing is the spending billions of dollars each year on programs that don't work. And without rigorous impact studies, including RCTs, we won't know if they work or not. The sacrifice of having a control group is a small one compared to the benefits of building an evidence base about effective development programs" (White 2013: 10). Essentially, it is unethical not to experiment if one does not have rigorous evidence on programme effectiveness. For example, in a 3ie/IFPRI Seminar Series on Impact Evaluation, Renos Vakis of the World Bank reported evidence on the impact of a cash transfer programme on early childhood cognitive development, based on an RCT that randomized benefits to a sample of infants (3ie and IFPRI 2012).

Evidence-building pilots generally adopt a "do no harm" ethical approach, adopting quasi-experimental methodologies (such as PSM or regression discontinuity techniques) or else providing treatments with unambiguous net benefits (such as cash transfers) to both treatment and control groups while randomizing treatments with uncertain costs and benefits. For example, evaluations of South Africa's Child Support Grant, a rights-based cash transfer programme, have successfully employed PSM techniques that have not required the withholding of benefits to any child. An RCT of savings and investment linkages to the Child Support Grant employed treatment and control groups of adolescents who all received the cash transfer, but randomized the assignment of treatments that involved costs and benefits to participating households and an uncertain net benefit (in clinical terms, satisfying the condition of "equipoise").

Conclusions

While experimental methodologies in social policy research have contributed substantially to the social policy evidence base, these "gold standard" evaluation designs are limited in their explanatory power. Amartya Sen makes an important distinction between culmination and comprehensive outcomes, emphasizing the difference in instrumental and descriptive

significance between the two. In culmination outcomes the analysis is confined to accounting for the consequences of determined inputs, with little consideration of the interactions, interests or unforeseen influence of actors and institutions throughout the process. Comprehensive outcomes, Sen contrasts, comprise the process, institutions and actors, as well as the outcomes of their actions. This analytical distinction offers a dichotomous framework from which this paper can compare methodologies and interrogate the effect of study design on the final interpretation of the outcomes.

Standard experimental and quasi-experimental methodologies are usually designed only to capture culmination outcomes of the evaluated intervention. These methodologies have important merits; however, they are characteristically guided by corrective mechanisms that prioritize internal validity over policy relevance. While these methods can facilitate clinical evaluations, they are limited in social policy settings where community context and individual behaviour vary unpredictably, and in particular where the specific elements of the required intervention are unknown.

A number of methodological approaches can contribute to a broader understanding of policy impact: emphasizing the importance of mixed methods research; incorporating *ex ante* evaluation approaches, including theories of change into evaluation design; mapping causal pathways through a diversity of analysis mechanisms and "points of analysis"; designing evaluations to measure combined effects; and integrating participatory evaluations.

A comprehensive evaluation approach will include a number of key features:

- Evaluating the linkages and interactions between different policies, actors and institutions.
- Mapping both immediate as well as reverberating outcomes that both compound and weaken the impacts of a given intervention.
- Assessing the social, economic, political and cultural context within a given intervention area, and attempt to portray how this context may influence the mapped outcomes.
- Responding to policy makers' questions (demand-driven); should accommodate both short-and long-term evaluation methodologies.
- Recognizing that the impact of one sectoral intervention on a specific outcome depends critically on the related interventions across a range of sectors.
- Informing the optimal balancing of multiple interventions to achieve a range of joint outcomes.

A move toward more integrated and comprehensive evaluation requires methodological change at both macro and micro levels. The macro evaluation framework reinterprets a traditional input–output matrix to describe the importance of both "intra-sectoral" and "inter-sectoral" linkages, wherein the inputs are the set of instruments, programmes and policies that enable the government to achieve national priorities objectives (outputs). This macro framework is a means of visualizing the complex interactions between multiple interventions in a manner that allows for better planning and description of policy priorities.

The micro framework expands the model for conventional experimental evaluations, proposing an alternative trial model that rebalances the trade-off between internal and external validity toward a greater emphasis on policy relevance. An "evidence-building pilot" that employs learning-by-doing processes to identify and evaluate appropriate and effective combinations of interventions can expand the methodological options available for social policy analysis. Given the complexity of many critical policy questions, evaluation designers can place a greater emphasis on illuminating the

increasingly complex questions that policy makers are asking. The approach can incorporate intra-trial innovation and support the testing of a wider range of interventions in a more cost-effective manner.

Both the micro and macro evaluation frameworks mutually inform each other. The micro evidence-building pilot approach enables policy makers to understand the resulting impacts of interactions of multiple interventions on a set of important policy outcomes. The macro framework integrates the evidence from multiple micro assessments into a macro policy evaluation matrix. This framework recognizes how the magnitude of any specific intervention's impact on an outcome of interest depends critically on the complementary interventions that also affect this outcome. By integrating complex micro assessments, the macro evaluation framework offers two important enhancements to conventional evaluation frameworks:

1. The conditional impact of an intervention (conditioned on the complementary interventions) can be measured. This better enables policy makers to choose optimal combinations of interventions to maximize cost-effective impact.
2. The full range of impacts across policy sectors can be more comprehensively measured. This provides a fuller evaluation of an intervention's true impact.

An example of a complex problem in a specific policy context illustrates the innovation potential of this approach. Policy makers in South Africa face challenges of high youth unemployment, rising HIV incidence among adolescents, challenges in delivering human capital services and a fiscal crisis from slowing economic growth. The country invests more in social protection than almost any other developing country—two out of three South Africans live in a household receiving social cash transfers. How can more comprehensive evaluation better inform policy choices?

The government is implementing an evidence-building pilot that integrates multiple interventions, including the Child Support Grant, financial inclusion initiatives, youth development programmes and other elements. The outcomes of interest include educational attainment, financial savings, youth employment, HIV prevention and larger outcomes in terms of social cohesion and inclusive economic growth.

The pilot aims to identify a cost-effective combination of interventions that achieve a range of policy outcomes, each one on its own insufficient to warrant the expected cost of the policy reforms. The required set of interventions was not known prior to the start of the pilot, so treatment changes dynamically in response to the evidence the trial provides. For example, use of cell phones and bank accounts unexpectedly exposes trial participants to severe private sector financial abuses, so consumer protection interventions are incorporated during the trial.

Comprehensive evaluation in this case provides two additional benefits compared with traditional approaches:

1. It enables the government to identify a complex set of interventions not identifiable in advance to cost-effectively achieve a range of diverse but critical policy outcomes. The comprehensive approach maximizes the likelihood of success while providing a cost-effective solution.
2. It facilitates a full benefit analysis across a spectrum of policy sectors, expanding the range of stakeholders who value the joint intervention and strengthening political support for the necessary reforms and innovations.

References

3ie/IFPRI (International Initiative for Impact Evaluation/International Food Policy Research Institute) (2012), *Cash Transfers, Behavioral Changes, and Cognitive Development in Early Childhood: Evidence from a Randomized Experiment.* Seminar Series on Impact Evaluation. 6 September 2012. www.managingforimpact.org/resource/3ieifpri-cash-transfers-behavioral-changes-and-cognitive-development-early-childhood-eviden, accessed 24 November 2014.

Acemoglu, A. (2010), Theory, General Equilibrium, and Political Economy in Development Economics, *Journal of Economic Perspectives*, 24 (3), Summer 2010, 17–32.

Adato, M. (2008), Combining Survey and Ethnographic Methods to Improve Evaluation of Conditional Cash Transfer Programmes, *International Journal of Multiple Research Approaches*, 2 (2), 222–236.

Angelucci, M. and De Giorgi, G. (2009), Indirect Effects of an Aid Programme: How do Cash Transfers Affect Ineligibles Consumption? *American Economic Review* 99, 486–508.

Arrow, K. (2006), Freedom and Social Choice: Notes in the Margin, *Utilitas*, 18(1), 52–60.

Banerjee, A. V. & Duflo, E. (2009, The experimental approach to development economics. *Annual Review of Economics* 1, 151–178.

Barrett, C.B. and Carter, M.R. (2010), The Power and Pitfalls of Experiments in Development Economics: Some Non-random Reflections, *Applied Economic Perspectives and Policy* 32(4), 515–548.

Berg-Schlosser et al. (2009), Qualitative Comparative Analysis (QCA) As an Approach, in Rihoux, B. and Ragin, C. (eds.) *Configurational Comparative Methods, Qualitative Comparative Analysis (QCA) and Related Techniques*, Thousand Oaks, CA, and London: Sage.

Campbell, D.T. (1957), Factors Relevant to the Validity of Experiments in Social Settings, *Psychological Bulletin*, 54(4), 297–312.

Carvalho, S. and White, H. (1997), *Combining the Quantitative and Qualitative Approaches to Poverty Measurement and Analysis. The Practise and Potential*, The World Bank, Washington, DC.

Claxton K. et al. (2007), *Mark versus Luke? Appropriate Methods for the Evaluation of Public Health Interventions*, CHE Research Paper 31. York, UK: The University of York.

Daidone, S. et al. (2012), *Analytical Framework for Evaluating the Productive Impact of Cash Transfer Programmes on Household Behaviour*, Methodological Guidelines for the From Protection to Production (PtoP) project, FAO.

Davies, R. and Dart, J. (2005), *The Most Significant Change (MSC) Technique: A Guide to Its Use*, April, www.mande.co.uk/docs/MSCGuide.pdf, accessed 24 November 2014.

Deaton, A. (2010), Instruments, Randomisation and Learning about Development, *Journal of Economic Literature* 48(2), 424–455.

Devereux, S. and Roelen, K. (2014), *Evaluating Outside the Box: Mixing Methods in Analysing Social Protection Programmes*, Centre for Development Impact, Practice Paper, www.ids.ac.uk/cdi.

Devereux, S. et al. (2013), *Evaluating Outside the Box: An Alternative Framework for Analysing Social Protection Programmes*, IDS working paper 2013(431).

DSD et al. (2012), *The South African Child Support Grant Impact Assessment: Evidence From a Survey of Children, Adolescents and Their Households*. Pretoria: UNICEF South Africa.

Duflo, E. et al. (2006), Using Randomisation in Development Economics. In T.P. Schultz and J. Strauss, (eds.), *Handbook of Development Economics*, Amsterdam: Elsevier, pp. 3895–3962.

FAO (Food and Agriculture Organization of the United Nations) (2013), *Qualitative Research and Analysis of the Economic Impacts of Cash Transfer Programmes in Sub-Saharan Africa: Ghana Country Care Study Report*. Oxfam Policy Management. FAO, Rome.

Handa, S. et al. (2013), *Livelihood Empowerment Against Poverty Impact Evaluation*, Carolina Population Center, University of North Carolina.

Holmes, R. et al. (2012), *The Potential for Cash Transfers in Nigeria*. Overseas Development Institute, London.

Hughes, K. and Hutchings, C. (2011), *Can We Obtain the Required Rigour without Randomisation? Oxfam GB's Non-experimental Global Performance Framework*, 3ie Working Paper No. 13.

Lutz, B. et al. (2014), Financing Structural Interventions: Going beyond HIV-Only Value for Money Assessments, *AIDS* 28, 425–434.

Marchal, B., Dedzo, M. and Kegels, G. (2010), Turning around an Ailing District Hospital: A Realist Evaluation of Strategic Changes at Ho Municipal Hospital (Ghana), *BMC Public Health* 10, p. 787.

Mayne, J. (2011), *Contribution Analysis: Addressing Cause and Effect, Evaluating the Complex*, New Brunswick, NJ: Transaction Publishers.

Mui, L. et al. (2002), *Notions of Reputation in Multi-Agents Systems: A Review, AAMAS '02 Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1.* New York, AAMAS, pp. 280–287.

Olofsgard, A. (2012), *The Politics of Aid Effectiveness: Why Better Tools can Make for Worse Outcomes*, Stockholm Institute of Transition Economics, Working Paper.

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

Ravallion, M. (2009), Should the Randomistas Rule?, *Economists' Voice*, 6(2), 1–5.

Remme, M. et al. (2014), Financing Structural Interventions: Going beyond HIV-Only Value for Money Assessments, *AIDS,* 28, 425–434.

Samson, M. (2012), *Exit or Developmental Impact? The Role of 'Graduation' in Social Protection Programs*, Research Report commissioned by the Australian Agency for International Development (AusAID), AusAid, Canberra.

Samson, M. (2013), How are Countries Using Social Protection to Benefit the Poor?, In OECD, *Development Co-operation Report 2013: Ending Poverty*. OECD Publishing. http://www.oecd-ilibrary.org/docserver/download/4313111ec010.pdf?expires=1424072194&id=id&accname=guest&checksum=957539DFBDEF76E4DF9F84A64ABB2670 , accessed 16 February, 2015.

Samson, M., I. van Niekerk and K. Mac Quene (2011), *Designing and Implementing Social Transfer Programmes*, Second Edition, EPRI Press, Cape Town.

Sanson-Fisher, R.W. et al. (2007), Limitations of the Randomized Controlled Trial in Evaluating Population-Based Health Interventions, *American Journal of Preventive Medicine* 33(2), 155–161.

Sen, A. (1997), Maximization and the Act of Choice, *Econometrica*, 65(4), 745–780.

Sen, A. (2009), *The Idea of Justice*, Harvard University Press, Cambridge, MA.

Stern, E. et al. (2012), *Broadening the Range of Designs and Methods for Impact Evaluations*, Working Paper 38, April. Department for International Development, London.

Taylor, J.E. (2013), *A Methodology for Local Economy-Wide Impact Evaluation (LEWIE) of Cash Transfers, Methodological Guidelines for the From Protection To Production Project*, Department of Agricultural and Resource Economics, University of California, Davis, California.

UNICEF (United Nations Children's Fund) (2011), Strengthening and Scaling Up the Social Protection System. http://www.unicef.org/mozambique/media_9455.html, accessed January 2015.

White, H. (2008), Of Probits and Participation: The Use of Mixed Methods in Quantitative Impact Evaluation, *IDS Bulletin* 39(1), 98–109.

White, H. (2013), An Introduction to the Use of Randomized Control Trials to Evaluate Development Interventions, *Journal of Development Effectiveness*, 5(1)

White, H. and Phillips, D. (2012), A*ddressing Attribution of Cause and Effect in Small Impact Evaluations: Towards An Integrated Framework*, International Initiative for Impact Evaluation, Working Paper 15, http://www.3ieimpact.org/media/filer_public/2012/06/29/working_paper_15.pdf, accessed 2 February 2015.

Wolff, N. (2000), Using Randomized Controlled Trials to Evaluate Socially Complex Services: Problems, Challenges and Recommendation, *The Journal of Mental Health Policy and Economics* 3, 97–109.

Woolcock, M. (2009), *Towards a Plurality of Methods in Project Evaluation: A Contextualised Approach to Understanding Impact Trajectories and Efficacy*, BWPI Working Paper 73. Manchester: Brooks World Poverty Institute.