Toward Results-Based Social Policy Design and Implementation

Miguel Székely

Abstract

This paper analyzes some of the elements that cause the apparent perception in the realm of social policy, and in particular in the case of poverty alleviation and education policies in developing countries, that on the one hand, too little evidence is produced on the impact of specific policies and programs on human development, and on the other, that very little use is made of the available knowledge. We label this the "under use of scarce knowledge" paradox. We argue that, in order to move forward, it is necessary to go beyond looking separately at the supply and demand for evidence, which appears to be the prevalent view, and visualize more integrated approaches. One option for greater integration could be the evolution toward Results-Based Social Policy Design and Implementation systems, which, as we propose here, consider incentives for use and production through a set of institutional arrangements that help focus public action on producing better outcomes.

Center for Global Development
www.cgdev.org

Working Paper 249
April 2011

Toward Results-Based Social Policy Design and Implementation

Miguel Székely
Director, Institute for Innovation
in Education Tecnológico de Monterrey

The author thanks Bill Savedoff, Eduardo Amadeo, Michael Clemens, Gonzalo Hernández, for useful feedback and suggestions.

CGD is grateful for contributions from the The William and Flora Hewlett Foundation in support of this work.

Miguel Székely. 2011. Toward Results-Based Social Policy Design and Implementation." CGD Working Paper 249. Washington, D.C.: Center for Global Development. http://www.cgdev.org/content/publications/detail/1425010

Center for Global Development
1800 Massachusetts Ave., NW
Washington, DC 20036

202.416.4000
(f) 202.416.4050

www.cgdev.org

Foreword

In the last decade, efforts to systematically study the effectiveness of programs in developing countries have expanded dramatically. When CGD's Evaluation Gap Working Group published "When Will We Ever Learn?" in 2006, the literature review found very few rigorous impact evaluations on topics like expanding access to health insurance, effectively training teachers, or sustaining water projects. Today, by contrast, organizations ranging from bilateral agencies and multilateral development banks to NGOs, private foundations, and university research centers are engaged in a growing number of rigorous impact evaluations, frequently engaging policymakers in defining questions and research design.

While international groups play an important role in funding, advising and conducting impact evaluations, it is developing countries themselves who are the most critical actors. Mexico has been a pioneer in this regard, demonstrating the value of independent external evaluation to policymaking with regard to its national conditional cash transfer program in the 1990s and later enacting legislation to require that the executive branch evaluate and use the resulting evidence in its policy decisions.

This paper by Miguel Szekely shows how Mexico has improved the evidence base for public policy in a number of ways. Szekely provides a unique perspective, combining the analytical eye of a researcher from his years at the Inter-American Development Bank and the political sense of a policymaker from his time as a high-ranking official in the Ministries of Social Development and Education under Presidents Fox and Calderón. Szekely explains the difficulties of conducting good impact evaluations and then assesses the interests of key stakeholders in promoting or opposing the creation and use of evidence. He draws out lessons from the government's effort to evaluate a major anti-poverty program (PROGRESA-Oportunidades), publish politically-sensitive poverty data, introduce performance measurement in education, and institutionalize learning. He concludes with a proposal, showing how developing countries could systematically incorporate evidence in policymaking.

Szekely is not blindly optimistic about the prospects for institutionalizing the use of evidence in policymaking. The obstacles are real and navigating among the many competing interests is quite complex. Yet, by grounding the analysis in real experiences, this paper shows that pessimism is also unwarranted. Mexico has made tremendous progress in ways that would have seemed

inconceivable only 10 years ago. Other countries can learn from this experience and adapt it to their own needs.

William D. Savedoff
Senior Fellow
Center for Global Development

Toward Results-Based Social Policy Design and Implementation

Introduction

Development is a moving target. One the one hand, the benchmarks by which its success or failure is assessed, are modified continually. Country objectives and goals are updated and adapted constantly because of the process of development itself, so that what may have been considered a success in the past can be deemed totally unacceptable today. Expectations and benchmarks can also be affected by the speed at which different countries progress, so even improvements in absolute terms with respect to one´s own history may appear to be satisfactory or insufficient depending on performance as compared to that of others.

On the other hand, the way in which development is characterized can also change over time. Modifying the dimensions by which development is defined and measured affects perceptions of whether a country is improving, or not, because inclusion of new variables and indicators may reveal aspects that are hidden by previous alternatives. Take for instance the well-known debate on using the gross domestic product (GDP) of a country as a measure of development. GDP provides information about the capacity of the economy to generate wealth, and while a specific rate of GDP growth can be qualified in different ways depending on historical patterns and relative performance with respect to others, positive growth is normally considered beneficial. However, if the focus is not only on the output level but also on the way in which wealth is distributed, the same GDP growth rate may be considered unacceptable if it is accompanied by greater income concentration or increasing poverty levels.

Because development is such a broad concept, the way in which it has been characterized and evaluated has evolved significantly over time. Broadly speaking, during the 20thcentury there was a gradual shift from a focus on *expenditure* measures such as budgets spent on development efforts, to a concentration on *inputs*—that is, the goods and services that are thought of as determinants of well-being: for instance, the number of roads, schools, health clinics, or available housing. Only relatively recently has interest changed from expenditures and inputs to *outcomes*, understood as evidence on direct indicators of individual or collective well-being such as household welfare, income-earning capacity, human capital accumulation, health status, effective access to basic goods and services, and enhanced opportunities, among others. The evolution toward outcome measures implies using new mechanisms for measuring and evaluating progress.

While the assessment of expenditures and inputs can be monitored through auditing and accounting mechanisms, measuring outcomes is generally much more complicated, since it requires new data collection efforts and statistical tools and methods that enable specific policy actions or programs to be associated with particular impacts.

Broadly speaking, evaluation can include the generation of information at the macro-aggregate level to assess general performance, it can measure and inform on intermediate processes, or it can measure in detail the effect of a specific program or action. At the aggregate level, one influential event for changing the scope of development toward final outcomes is the definition and agreement for using the Millennium Development Goals (MDGs) as measures of achievement in the international community.[1] On the verge of the 21st century, the United Nations (UN) and a group of important multilateral institutions agreed, that development results were not satisfactory in most of the developing world, and that specific efforts should be directed to actions that guaranteed concrete and measurable results in basic areas. Specific indicators were set to follow progress in order to reinforce and refocus policies and investment to reach specified targets by 2015. The main areas included in the MDGs are final aggregate outcomes in health, education, income poverty, gender equality, and environmental sustainability, and in this framework, expenditures (e.g., budgets) and inputs (goods and services) are viewed as intermediate vehicles for achieving the eight higher goals. What might seem to be a subtle shift on emphasis to poverty, health, and education as development targets is having considerable influence in the real world by providing useful guidance for policy action, and redirecting the behavior of individuals and institutions toward quantifiable impacts.

At the micro level, the shift to concentrating on outcomes has been strongly influenced by the perception of diverse actors that the growing amount of resources allocated has not necessarily positively affected development levels or prospects in the poorest countries, in part because of the focus on expenditure and input indicators. Birdsall and Savedoff (2010) show that foreign aid increased by around 70 percent during the 2000s decade—increasing from around US$70 billion to around US$120 billion between the years 2000

[1]The MDGs were agreed on in the late 1990s with the objective of setting homogeneous outcome targets for developing countries for the period 2000-15. They are officially followed up by the United Nations System, and reported by all countries joining the initiative. MDG 1 refers to reducing extreme poverty by half; MDG2 refers to achieving universal primary school enrollment; MDG3 refers to promoting gender equality for women; MDG4 implies reducing child mortality; MDG5 refers to improvements in maternal health; MDG6 has to do with combating HIV/AIDS, malaria, and other diseases; MDG7 refers to ensuring environmental sustainability; MDG 8 is about developing partnerships for development.

and 2007—while concerns about its effectiveness have grown in parallel, in large part because of the scarce evidence of a positive effect on final development results.[2]

This relatively recent shift in focus has been accompanied, paradoxically, by two concerns in the development and social policy literature. On the one hand, as argued by Savedoff, Levine, and Birdsall (2006), among others, the amount of evidence and knowledge on the effects of different types of policy interventions on development results is still scarce. On the other hand, there are documented perceptions that the available evidence is being used in policy design and execution to only a limited extent for improving policy (see, for instance, Ravallion 2008; Jones, et al. 2009; Weiss 1999; and Deaton 2010). So, there appears to be too little evidence and at the same time very little use of the available knowledge. We label this the "under use of scarce knowledge" paradox.

On the supply side, the concern by donors and international agencies has led to the creation of the International Initiative for Impact Evaluation (3ie), anew agency that concentrates on three objectives: 1) providing and summarizing existing evidence on effective development policies; 2) offering policymakers and development practitioners a window where they can submit programs, actions, and interventions that are likely to generate knowledge for improved performance; and 3) offering funding and technical support for evaluators that have identified social programs, actions, and interventions that are prone to impact evaluation for generating knowledge that improves the specific action in question, while generating information that can be useful in other settings. This initiative addresses one of the issues that constrain the production of informative evaluations, which is the "public good" nature of knowledge generation that reduces incentives for local actors for financing the costs of high-quality assessments.

The concern on the demand side, that is, limited use of the available knowledge, is also highly relevant because limited use implies lower-quality policy. Evidence can serve at least three important purposes for improving public action. The first is *measuring* for informing on the evolution of development outcomes. As with the MDGs, simply generating information on the pace at which progress is being made is valuable because it allows for transparency and for adjusting expectations. The second is *assessing* whether progress is or is not satisfactory, which is a key element for accountability. By comparing the evolution of a development dimension with its corresponding target or benchmark, it is possible to judge whether the current evolution is below or above satisfactory levels. The third is *explaining* success or failure, which enables the identification of areas for improvement and generation of feedback for policy design and execution. As judged by

[2]See Banerjee (2007) and Pitman, et al. (2005) for expressions of the same concern.

recent concerns, the limited use of evaluation in the context of development is starting to become a restriction in itself for guaranteeing better results.

This paper analyzes some of the elements that cause the apparent paradox of "under use of scarce knowledge"," in the realm of social policy, in particular, the case of poverty alleviation and education policies in developing countries. We argue that, in order to move forward, it is necessary to go beyond looking separately at the supply and demand for evidence, which appears to be the prevalent view, and visualize more integrated approaches. One option for greater integration could be the evolution toward Results-Based Social Policy Design and Implementation systems, which, as we propose here, consider incentives for use and production through a set of institutional arrangements that help focus public action on producing better outcomes.

The argument is developed in five sections. Section 1 includes a discussion on the complexity of evaluating development effectiveness in the realm of social policy. Section 2 reviews some of the literature on the concept of "evaluation use" and examines the behavior of the main actors involved in using and producing evaluation for development effectiveness. Section 3 presents some evidence from Mexico that is a useful reference point for identifying the elements that a full Results-Based Social Policy system should include. Section 4 sketches the main elements that this kind of system could include. Section 5 concludes.

1. The Complexity of evaluating development effectiveness

Why are knowledge and evidence on what works for human development both scarce and underutilized? One reason has to do with the scope and information that can be provided by evaluations. As is generally the case in all social sciences, dealing with individual behavior as the subject for analysis when evaluating the impact of a particular action involves high degrees of uncertainty, and is much more complex than measuring expenditures or inputs. While most expenditures and inputs can be measured and followed through auditing and accounting mechanisms, outcomes require understanding individual behavior and reactions, and even though substantial methodological advancements have been made over decades, still the literature on this matter can only offer approximate assessments of the full effect of policy actions.[3]

[3]This is not the case in important sciences that analyze inert materials or bodies, where low margins of error can be obtained through a series of evaluations and analysis on how each type will react when exposed to different conditions. The margins of uncertainty in social policy are also generally greater than in scientific analysis in areas such as health, where predictions can be made, for instance, about the mechanical reaction of the human body to an external treatment. In this case there will also always be margins of error resulting

At least four broad features increase the complexity of measuring the impact of policies on human development and make their production and use challenging. The first is heterogeneity. Policy action is usually aimed at groups of individuals who share certain characteristic or problem; it is seldom the case that government action can be tailored for each individual´s circumstances and needs. However, depending on the preferences and values of each individual and his or her exposure to particular conditions, contexts, situations, and backgrounds, totally different reactions and effects can be triggered by the same policy. Since individual preferences and values inherent in each utility function are unknown and unobservable, a margin of error and uncertainty will be present in any attempt to measure the reaction to certain intervention, because of this heterogeneity. Different approaches exist for inferring preferences and values, but in the end the best one can hope for is a good approximation based on averages, where the mean value of an expected intervention may not correspond to the case of any of the individuals in the population.[4]

A second challenge has to do with context. Even in a hypothetical case where enough information on individual preferences was available, preferences interact with particular conditions, contexts, situations, and backgrounds, generating a specific response by the individual in those circumstances. The information generated by the analysis of the response will be useful for measuring the impact of the action, but may not necessarily be relevant for inferring impact in other scenarios. Because it is practically impossible to replicate the exact same setting in all cases, identical effects will not necessarily be observed in a different population. Thus, what works in one case might not necessarily work in another. One example is the effect of scholarships on girls' school attendance. In settings where there is a cultural preference for gender equality and female labor market participation, a cash transfer may have important effects on education, while exactly the same transfer in a setting where early marriage is highly valued may produce a much less significant impact.[5]

Different contexts might lead to a wide range of expected and unexpected effects that can go beyond the measurement and evaluation effort. Evaluations are usually concerned

---

from lack of knowledge of all the possible reactions or response mechanisms, but once a regularity is established variations will be most likely the result of external elements not considered so far, and not by the body "deciding" to behave differently suddenly.

[4]See Heckman (1997) for a detailed discussion on the methodological challenges implied by heterogeneity in program evaluation.

[5]Attanasio, et al. (2004) present an approach to simulate the effect of extending the Mexican cash transfer program Oportunidades to different settings, and discuss the challenges and limitations involved in this kind of analysis.

with measuring the predicted impact of an intervention and may even gather additional information on other related consequences, but unless a full general equilibrium setting where all crossed effects, secondary effects, and spillover effects of policy action can be captured, full information on the total impact will be unavailable. In this case also, there can be good approximations for quantifying foreseeable consequences, but even when this is so, the uncertainty with regard to further impacts can leave high avenues of risk open, and the general effect of the intervention will remain unknown. [6] Take for instance the case of investment in education. By promoting human capital accumulation, one would expect that new generations would face better income opportunities, and this might well be the case. However, if the context is shaped by certain labor market conditions, the incorporation of more educated individuals leads to lower returns to education and reduced wages, the final impact of the intervention might even be negative in the income dimension. This is only one illustration of the fact that no evaluation will be able to capture absolutely all the reactions triggered by a specific policy action, and one can at the most hope to generate some information on the diverse dynamics that can be envisioned.

Another example along the same lines has to do with multidimensionality. Since well-being is expressed through different dimensions, unexpected effects might be observed because they affect well-being from an unpredictable angle. For instance, relatively recently conditional cash transfer programs incorporated women´s empowerment by providing the female household head with control over cash transfers for children´s school attendance and health and nutrition benefits. The argument is that this improves resource allocation within the household in terms of well-being and that it generates greater opportunities for women. There is in fact evidence that this can be the case. However, in some settings, the recipient of new resources can be subject to violence and even greater discrimination to counterbalance the effect of the intervention, resulting in even more negative conditions than those originally observed. Evaluations concentrating only on the direct effect of the transfer may misleadingly conclude that all individuals in the household are better off, when the limited scope of the analysis does not account for the cultural context.

Another feature is that most development interventions have intertemporal effects that cannot be observed immediately. Education is a good example. Enhanced human capital is expected to generate greater opportunities that may translate into increased material wealth, professional development, freedom, culture, knowledge, etc., but all these elements can be fully captured only by following each individual over his/her life-cycle,

[6] See Acemoglou (2010) for a detailed discussion of incorporating general equilibrium effects in policy evaluation.

as well as the effects on the rest of society. Good approximations such as current and past income flows, labor market situation, etc., might provide some information on the relationship between improved education and well-being, but they will always be partial measures. Other positive effects such as reduced crime rates, productivity impacts, and reduced corruption are not normally captured, thereby potentially underestimating the benefits of the intervention.

A third characteristic is that even when global effects can be measured, evaluations rarely are able to identify the exact mechanisms through which an outcome is generated. Every good or service has a production function that requires a mix of inputs that may yield positive or negative effects, and identifying the role of each can be virtually impossible. There may even be situations where an intervention enhances the positive effect of some input at the expense of a negative influence of another, but these dynamics are normally hidden in the process of aggregation of the final impact. The difficulty in identifying each individual contribution can blur the final picture and can lead to mistaken conclusions or misleading impressions about the influence of different elements. When many individuals participate in the delivery of a service or in the operation of a program, the problem can be enhanced. One example is schooling services. Even when going only through primary education, one is exposed to a large number of teachers, materials, equipment, contents, etc., that build up to the human capital level reached upon graduation. Identifying the effect of each of the individual inputs would require following what happens practically every day with each of them, which is normally not feasible. This, added to the intertemporal feature of human capital accumulation, implies that, for instance, the effect of a good or bad teacher will be difficult to identify, and measuring its full effect and influence in the future is unthinkable.

Similarly, isolating the causality between certain policy action and an effect on an individual or group is also a difficult task. Programs are applied in a context where a wide range of forces and elements interact simultaneously, and being able to disentangle an exact response from a specific policy can also be virtually impossible. Even in cases where improvements in living conditions are observed after implementation of a certain program it is not straightforward to attribute the effect to the action, unless the only element modifying the status quo at the time was that specific policy. But even then, if there are differences in background or initial conditions that provoke an outcome observed at the same time that the policy response is expected, but that is unrelated to the program, the effect may erroneously be attributed to the intervention. There are several ways of addressing this issue. For instance, the literature on social program evaluation has borrowed solid methodologies from the medical and other scientific fields to develop experimental designs comparing the evolution of treatment and control groups. This approach has proliferated especially for measuring the impact of conditional cash transfer

programs, which is a considerable evolution from inferring program impact through before-after broad comparisons.[7] It is widely recognized that experimental designs can generate perhaps the most useful and robust approximations to impact, but they are not exempt from their own limitations. Apart from methodological considerations (discussed, for instance, in Deaton 2006), there are important technical challenges involved in the econometric estimation procedures used (some of them summarized in Ravallion 2008 and Jones, et al. 2009), and there are also ethical and political considerations (discussed in Section 2 below) that do not allow for straightforward application and widespread use.

So, evaluating development effectiveness is not easy, and even when it is performed by using the best tools and information available, the evidence it generates is inevitably limited. This of course calls for sustained efforts in improving data and estimation methods, on the one hand, and warning users about the potential as well as the caveats of the evidence, on the other. An additional element that may help explain the apparent status quo of low production and meager use of available knowledge has to do with the nature of the actors participating in the process. The following section analyzes this aspect.

2. Understanding the actors

There is a large literature dating back at least four decades analyzing the process by which evidence from evaluations is used in social policies. Even the concept of what "*use*" means has been subject to debate. For instance, one of the early papers by Patton (1975) analyzes the United States' health system and shows that, from the point of view of those generating evidence, evaluation is useful for constructing a knowledge base to understand social processes better, while for health authorities, the knowledge is of use only if it allows for improving services in a visible way.

Weiss (1979, 1999) throws further light on this debate, by classifying definitions of utilization in seven categories: (i) *knowledge-driven*: where research develops findings and knowledge on its own initiative and, as a consequence, policymakers who find it useful take the new evidence to improve public action; (ii)*problem-solving*: where what triggers research is the quest by policymakers to answer a specific policy-related question, and its findings are used to inform decisions; (iii) *interactive*: where knowledge

[7] Skoufias (2005) documents what was perhaps the first cash transfer program evaluated by using techniques beyond measuring expenditures, and inputs, or simply comparing the situation before and after implementation. The PROGRESA program (for its acronym in Spanish referring to the Food, Education and Health Program) initiated in Mexico in 1998 and was accompanied by an experimental evaluation design from the outset.

is generated by a wide set of actors producing useful data and information in a nonsystematic way, some of which is used for policy; (iv) *political*: where information and knowledge that support a specific point of view, political position, or ideology are promoted by groups with political interests and are used to move an agenda forward by gaining advocates or disqualifying opposing views; (v) *tactical*: when evaluation is performed for the sake of evaluating, and not necessarily for taking concrete action derived from its findings (e.g., to comply with legislation to evaluate); (vi) *enlightenment*: where evidence influences the way in which people think about an issue in a general way by setting new parameters related to policy action—say, defining what an "acceptable" level of achievement means in a particular dimension; and (vii) *intellectual enterprise*: where evaluation is used to build on the knowledge base of society and improve the general understanding of one or more issues, which can help improve a specific policy or a wider set of them.[8]

Along the same lines, Marra (2000) builds on Weiss's definitions and documents another category that has to do with the use of evaluation to justify government action. Under the label of (viii) *instrumental* evaluation, the author explains how policymakers can also use evaluation results to disseminate their accomplishments under a halo of credibility. These eight definitions encompass most of the classifications used to date.

In the context of the paradox of "under use of scarce knowledge"" these categories can be thought of as different dimensions of the use of evaluations, since it is difficult to think about situations that could belong exclusively to one of them. Most commonly, evaluation efforts might fall into a combination depending on the actor and the context under analysis, and it might well be the case that a combination of all eight elements is observed simultaneously, with different emphasis depending on the situation.

Recent literature on this subject has evolved from thinking of evaluation *use* to evaluation *influence* using the argument that, rather than constituting individual episodes that are relevant only for specific cases, evaluation efforts have a longer-lasting life by gradually and cumulatively shaping, affecting, supporting, and changing persons and systems over time. The concept was first proposed by Kirkhart (2000) and developed further by Cummings (2002), Henry and Mark (2003), and Mark and Henry (2004), to explain the mechanisms through which evaluations may have effects on policy.[9] The mechanisms

[8]Weiss (1999) explains the *enlightenment* category in more detail and discusses further some of the issues related to the other categories. Sandison (2005) adapts these and other categories to evaluation within large organizations and offers interesting insights on the practical use of these definitions.

[9]Almeida and Báscolo (2006) present a useful review of some of this literature, while Caro (1971) also offers an insightful review of developments up to the early 1970s.

include the gradual trickledown from the individual to the interpersonal and to the collective transmission of information, and an identification of four processes: general influence, attitudinal, motivational, and behavioral.

There is also extensive literature on the process through which evaluation eventually is used for improved policy. In an early paper, Anderson (1966) described it by analyzing the case of health policy in the United States. According to the author, an evolution was observed from a first phase of using casual intuitive information for decision making that did not help build a knowledge base (policy without research). At this stage, evidence is neither systematically produced nor used, and short-term empirical observation guides most decisions. Eventually, there is an evolution to a second phase (of policy with some evidence), in which isolated disparate facts inform policy and lead to the accumulation of dispersed data that are used intermittently depending on the personal profile of decisionmakers. The main difference with respect to the first phase is that knowledge starts being valued, although used sporadically and nonsystematically. Finally, there is a third phase (of knowledge-based policy) where data and understanding are systematically and deliberately developed continually to improve the design and execution of policy. There is a considerable evolution at this point with respect to the previous two phases, since there is an explicit recognition of the value of knowledge for improved action.

As judged by the case studies in Jones, et al. (2009), this is a useful way of characterizing what today is observed in many countries. Unfortunately, experience indicates that what is most common is to reach the second phase of assuring that some (at least scattered) evidence is available to improve social policy.

Furthermore, many countries are far from the basic cycle outlined in Greenberg (1968) as necessary for providing a framework for the use of evidence in decision making. That framework consists of five simple steps: (i) performing a diagnosis of the problem that public action will address; (ii) design and goal setting; (iii) provision of the service or program; (iv) evaluation; and (v) cost-benefit analysis.

But even when a country or organization manifests the conditions necessary to reach the third phase described above, and can operate under the last five simple steps, working under a full knowledge-based policy setting still is not automatic. After at least three decades of research, there are still no recipes for how to guarantee that evaluation results are used for better policy making. Among the first to research this formally were Alkin and Daillak (1979), who through empirical observation were able to trace down each of the steps in the process of use of evaluations in a set of schools, concluding that both the specific contexts (including the demand for information by the stakeholders in the provision of the service) as well as the specific professional profile of individual

decisionmakers (where some are more receptive to evaluation practices than others) are at the heart of the degree to which evidence is used for improvement. Brown et al. (1984) also identify the context and profile of decisionmakers as relevant, and add the profile of the evaluator itself as an important determinant in the use of evidence. According to their research, one key feature of more effective evaluators is the capacity to interact with the agent that is being evaluated, and identifying issues that are relevant to improving their day-to-day activities (which is an aspect also stressed by Barkdoll (1980), and Lawrence and Cook (1982)).[10]

A clear message emerging from this literature is that a better understanding of the motives, context, and profile of each of the actors involved in the production and use of evaluation is necessary to improve our understanding of the underlying processes involved in its dynamics. In the context of the recent concerns in social policy that lead to the aforementioned paradox, at least five different actors with different priorities, interests, and preferences should be analyzed. The five actors include (i) external donors that allocate resources to countries, programs, or specific actions, and that may demand evidence of the impact of the activities carried out through the resources involved; (ii) high-level policymakers, who on the one hand are responsible for improving policy at the general level or sector, and on the other have the power to authorize and mandate the performance of evaluations; (iii) evaluators in charge of performing the analysis, who may have their own interests and motives; (iv) practitioners or program operators/executors who implement the program´s actions in the field or deliver a public service; and (v) constituencies, public opinion, direct/indirect beneficiaries, and other actors that demand the effective use of public resources. We examine some of their motives for producing or using evaluation results, as well as the limitations and circumstances that may shape their behavior. The analysis suggests that the differences in objectives, incentives, and motives across actors are some of the elements behind the paradox.

External Donors, Investors

Among the most influential actors in the process of evaluation production and use are donors or investors that are external to the operation of the program and to the agency in charge of its execution, but that provide financing for the implementation of policies and actions. The range goes from multilateral institutions and foundations to private actors at

[10]One interesting recent example of following up on the process through which evaluation feeds into policy is CONEVAL (2009), who examine in detail each evaluation performed on Mexican social programs for 2008-09 and their effects on changes in design and implementation.

the local level that are interested in supporting efforts to address certain problem. There are several reasons these actors can consciously not advocate for measuring the impact of their resources on development results, and in some circumstances they may even rationally undermine efforts to evaluate. The first reason is that evaluation requires resources. Normally, for a proper evaluation to take place, tailor-made data must be produced, resources must be invested in developing an adequate design, and funding is necessary for the professional services to analyze the data and report findings and conclusions. When resources are limited (which is most usually the case), a moral dilemma arises on whether funding should be allocated to evaluating impact or to benefiting more people. The obvious argument in favor of allocating it to evaluating impact is that the results will show whether the effectiveness and efficiency of resources spent can be enhanced, with further benefits for larger populations in the future. The argument against devoting resources to this activity is that when needs are severe, diverting funding away from a beneficiary might make the difference between a human being's life and death.

A second reason is that, because knowledge is a public good, the incentives for allocating resources to evaluation hold benefits well beyond the immediate interest of learning more about an action´s impact. This argument is discussed in detail by Avery et al. (1999) and Savedoff et al. (2006). The public good feature of evaluations is exacerbated when there are multiple donors participating in fundraising for large initiatives, which makes each perceive its contribution as only marginal (a drop in the ocean). Under this perception, each donor´s incentives and empowerment for demanding accountability are particularly weak.

A third reason has to do with the fact that the real underlying objective of external donors/investors in some circumstances might not be generating impact, but rather making the statement that they are supporting a particular cause. In such cases, the objective might well be a noble and legitimate one, and the measure of success will be the flow or resources itself, rather than its final impact, but neither the donor/investor nor the executor might have incentives to invest in evaluation.

A fourth reason that is related to the economic costs of performing evaluations is that private donations may be used for self-serving goals, including commercial interests or tax exemptions. Under some circumstances, donations may be attractive not only for their development impact as such, but because they offer an alternative to tax payment (e.g., individuals or organizations that might prefer any option rather than providing funding for the government), or for promoting or advertising certain commercial products. Under this scenario, there are fewer incentives to evaluate impact.

There are evidently also strong reasons why external donors/investors that are genuinely interested in promoting development may advocate and even fund evaluation. Perhaps the strongest reason is the presence of the principal-agent problem, where unless evidence is produced by a credible (usually external) party on the use of the resources, the principal (donor/investor) cannot make sure that its funding is being used for the intended purposes, because the agent (receiving agency/executor) might have its own different priorities and preferences. This is discussed in detail in Savedoff and Birdsall (2010), where the "cash on delivery" mechanisms is proposed for assuring that investment in aid and other kinds of support yields development results.

Another motive is that external agents with longer-term perspectives might find it profitable to invest in evaluation in order to generate knowledge to assure greater effectiveness and efficiency of their own future investments. Still another important cause is that when external donations and investments depend on fundraising, the evidence from evaluations may be critical for convincing donors to continue with the effort. Not having such evidence might be the determinant of the action´s sustainability.

In sum, the attractiveness of performing evaluation is not totally evident from the point of view of external donors/investors that finance certain public social programs and actions. In cases where the motive is the flow of resources in itself, none of the eight different dimensions of use discussed above might appear relevant, and evaluation might even be viewed solely as a financial burden. The lack of incentives to support evaluation in these cases will most probably trickle down to other relevant actors, with low investment, if any, in this activity. On the contrary, when the motive is generating development results, strong incentives for evaluation will normally be in place.

High-Level Decisionmakers

A second strategic agent is high-level policy and decisionmakers that can be users and (indirect) producers of evaluation results simultaneously. They are potential users in the sense that general decisions of policy orientation, funding, and implementation are under their responsibility, and the information generated by evaluations can be strategic in identifying areas for improvement and in providing feedback for better decision making. They can also be considered producers in the sense that authorizing the performance of an evaluation of a public program or activity in their realm is also under their responsibility. In this case also, one can think of reasons for or against devoting resources, and for authorizing an evaluation to be carried out.

On the positive side, the eight dimensions of use described in the previous section might be relevant from this actor´s perspective, and at first sight could appear sufficient to justify full advocacy and use of evaluation.[11]Even from the limited perspective of maximizing the time appointed to an important government position, information from evaluation might be a powerful tool, because it enables the communication of positive results to constituencies and builds their support. In some cases, it even has the advantage of offering insights on how to generate even greater future results (and support). One explanation for the limited use even when these potential advantages are present is that policy decisions are taken within a set of constraints and conditions that may divert interest from evaluation. Some of the most relevant are timing, interest groups, normative factors, institutional decisions, politics, and funding. The timing issue is commonly at odds with the dynamics of producing credible evidence. Time frames for delivering results are usually severely restricted, while a solid assessment of impact requires time for design, the generation of baseline data, allowing for the intervention to produce its effects, generating ex-post information, and, finally, analyzing and reporting results. All of this might even imply years of investment that go beyond a political cycle, and might discourage even strong advocates for evaluation because they will not be able to reap the benefits of the initiative.

Interest groups can also be an important constraint for high-level policymakers, inhibiting them from carrying out impact evaluations, especially in the case of experimental designs. The definition of control and treatment groups might be perfectly justifiable from a methodological point of view, but explaining to program nonparticipants that they have been excluded from a benefit because they were not "randomly selected," while others were, is not an easy task and may create enough opposition to make the evaluation infeasible. Even in cases where an experiment can be launched, the evidence of positive results may understandably generate pressure from the control group for incorporation, creating threats of "contaminating" the experiment. Interestingly, pressures usually arise when maintaining the control group as such is more important, since the full intended effect can be measured only when the intervention has time to yield its medium- or long-term impacts. The capacity of policymakers to deal with political and interest group pressure usually determines the feasibility of the experiment.

Normative factors may also play an important role. If, for instance, rules and regulations impede using program resources to evaluate its impact, even when policymakers are evaluation advocates, there will be underinvestment in these activities. Similarly,

[11]Of course, it is possible that the motive of specific individuals in decision-making positions is different from improving development in the particular area of influence, but we will assume that it is for the sake of our argument.

institutional arrangements may be an obstacle for evaluation. When governments are designed around the concept of measuring expenditures and inputs, the mandate to evaluate outcomes or agencies that can perform them may be nonexistent, making it impossible or even illegal to carry them out.

And there is also politics. Generating information on the efficiency of policy action may be highly risky in some settings. While positive effects can be capitalized politically, unfavorable results may be much more difficult to handle in certain circumstances and may require investing political capital in their management. As argued by Pritchett (2002), the risk of obtaining negative or not-so-positive results from an evaluation might be a strong deterrent for promoting it. Results may provide the opposition with ammunition that may backfire and become lethal politically, or may discourage external donors/investors from allocating additional resources. The risk is usually higher in societies where strong transparency and accountability mechanisms are institutionalized, and in environments of tight budget constraints (where, paradoxically, information on what policies are more effective is more in need) where many interest groups compete for resources. Providing sound evidence on program impact under these circumstances may be equivalent to signing its death certificate.

In the end, the balance among the eight dimensions of the use of evaluation, and the restrictions and circumstances described above, will be critical in defining the feasibility of producing and using evaluation from the point of view of high-level policymakers. An important determinant of which way the balance will go is the technical capacity for understanding evaluations and absorbing their results. Low professional capacities at this level may increase the risks of evaluating government action, and therefore reduce their attractiveness.

Evaluators

Evaluators also play an important role in the process. It is common for evaluators to be external to the operation of the program and to have sufficient independence to guarantee credibility. Self-evaluations can also be performed, but since they can be prone to subjectivity, their credibility is usually hindered. As recognized at least since Alkin and Daillak (1979), the approach chosen by evaluators is critical in promoting or discouraging the use and production of evidence. On the positive side, generating solid credible information on the effect of policy action may be of high value for external donors/investors that need assurance on the development effect of their intervention. For policymakers, timely, credible, and relevant information that helps capitalize positive government action and offers elements to improve policy performance will also be

appreciated. The political risks of not-so-good results can even be ameliorated sometimes by early involvement of the policymaker and program operators in the definition of the questions to be answered, and on the definition of the strategy for adequately communicating results.

On the one hand, however, choosing approaches that restrict the use of evaluations in one or more of the eight dimensions described above may hinder interest by policymakers and external donors and investors. For instance, an exercise that requires waiting for a full generation to obtain information may be totally infeasible for both. Similarly, methodological approaches that generate results that are valid only for a particular setting may be of little use for scaling up or informing decisions in other contexts. Concentrating on academically interesting issues that are operationally irrelevant will also deter the demand for assessments. Conclusions that are methodologically sound but politically unviable will usually find less receptive counterparts, while ideal evaluation designs from a technical point of view that require an unreachable budget, or are politically unmanageable, will most likely not be performed, or if performed will most likely be seldom used.

These situations may arise because the objectives and restrictions of external donors/investors and policymakers (described above) do not necessarily coincide with professionals with the capabilities of performing solid evaluations. Evaluators (mostly from academic circles) may prioritize academic purity, professional prestige, recognition, knowledge generation, academic success (in terms of high-profile publications), etc., that may be incompatible with evaluations that are timely, credible, relevant, pertinent, and communicable from the point of view of users. The perception that the available evidence is underutilized may well be because of its incompatibility with users' needs, along the lines discussed above.

Program Operators and Practitioners

For several reasons, perhaps the actors that are most affected in practical terms by the process of generating evidence from evaluations and of implementing changes to programs and services in line with their results are the program operators and practitioners working directly in the field. From the perspective of generating evidence, one reason is that producing new types of data and information needed for designing and implementing assessments usually requires additional work and efforts that are not necessarily accompanied by short-term identifiable benefits or incentives for them. Additionally, failures in delivering benefits to specific populations under strict schedules may put an evaluation at risk or may contaminate an experiment with important

consequences for the quality of the exercise. Program operators and practitioners may also be pressured by individuals in control groups or other non-beneficiary populations to be incorporated in programs, which can also compromise the quality of the evaluation.

But perhaps the main challenge has to do with the use of evaluation results. Shifting from a policy approach based on expenditures or inputs as measures of success to defining development outcomes as benchmarks is in itself an important cultural change. When new evaluations are presented and a set of recommendations that modify day-to-day practices, procedures, norms, and processes are introduced, those "suffering" the process of putting change into practice are precisely operators in the field. Phasing in and implementing new standards and methods in day-to-day operations and activities can be the most complex and laborious task in using evaluations effectively, since it inherently requires a change in individual behavior and customs. Often, even what can appear to be slight changes in procedures and norms take several years for full implementation, and explicit or implicit opposition at this level may inhibit evaluation use.

It is also common for operators and practitioners in some settings to become "constituencies" of the program to which they have devoted years of effort and experience. When operators become clients of their own program, they can be the first to obstruct or openly oppose change and make evaluation use effectively impossible in practice. Even if at the higher decision-making level there is strong commitment to use evaluation results for improving policy design and implementation, resistance at the bottom may make it impossible. This may be either for the practical reason of requiring additional effort, or because operators have become so closely identified with the program and with the way that it is designed and operated that they take any challenge to the status quo personally.[12]

Intensive training and information dissemination on the nature and purposes of evaluations may help ameliorate negative opposition against production and use, and experience shows that involving these actors in the process of design (for instance, by participating in the definition of the questions addressed by the assessment), including them in the hypothesis-setting process, and even in participating in the identification of potential areas for improvement, may help refocus their efforts toward a more effective use of the results obtained through evaluation.

[12]Ferman (1969) presents an interesting characterization of the general relationship between evaluators and program operators. Even though the paper was published 40 years ago, it illustrates clearly the conflicts and stress that may arise between different actors in the evaluation process, which can determine whether or not the whole evaluation process is feasible.

Constituencies, Public Opinion, and Beneficiaries

Political constituencies, beneficiaries, and public opinion are also relevant in the process of producing and using evaluations in social policy. As mentioned in the introduction, evaluation is a powerful tool for generating information, for making assessments about policy performance, and for offering feedback for improving policy action, and these actors play an important role in each of these functions. Making information on policy impact available enhances transparency and allows the public to know how taxes are being spent, or whether certain goals or benchmarks are achieved. It also allows for value judgments on whether performance is adequate or not, and is therefore a tool for making policymakers and program operators accountable for their actions. Transparency and accountability are highly valued in many political settings and constitute strong incentives for demanding evaluation production and, in fact, there is generally a positive relationship between the demand for evaluation and the level of transparency and accountability in a society.

Evaluation results can also be used by these actors to promote changes for improving policy effectiveness, although direct channels to exert their influence in this aspect are not always available.

When a positive combination is observed, on the one hand, of having an informed society that uses evaluation results to demand policy improvements and, on the other, of the existence of receptive, transparent, and accountable governments that implement improvements and inform and justify the use of results and recommendations, evaluation has an ideal setting for fulfilling its mission of becoming a strategic tool for development.

When either side (constituencies or governments) fails to play this role, however, evaluation can become a threat or even a negative instrument. For instance, when, rather than using results constructively for improving policy, constituencies, beneficiaries, and public opinion use the evidence only to expose failure, to signal and discredit specific participants, or even in a punitive way, the perception of evaluation as a risk for operators and policymakers will increase. It is common to find these types of situations, especially in countries where transparency and accountability are a novelty after long periods of censorship or inhibited citizenship rights, and the longer it lasts, the greater the costs associated with evaluation from the public sector´s perspective. The media usually play a key role in influencing the direction that the discussion and debate will take in this regard. A constructive media focused on improvement may lead the discussion toward follow-up on the use of results, while an extremely critical media focused only on identifying failure, signaling, or penalizing the actors involved in policy making and operation activities, may not necessarily have an impact on better performance.

In sum, each of the five relevant actors involved in the use and generation of evaluations in social policy plays an important role. Additionally, there are interactions among them that determine the final outcome. For instance, the combination of extremely critical constituencies and public opinion may inhibit evaluation practices when policymakers have low technical capacity for using and interpreting evaluation results, or may provoke extreme reactions by program operators who may obstruct further assessments in the future. Similarly, the combination of external donors/investors and policymakers centered on improving policy, combined with professional operators focused on generating outcomes, with feasible and useful evaluation design and implementation, and informed and critical, but still constructive constituencies, may yield a virtuous cycle of knowledge generation and policy improvement. As discussed below, aligning the incentives and objectives of all five actors and providing an adequate institutional setting, are the main challenges for evolving from generating scattered evaluations and intermittent use to a full evaluation system that promotes continuous improvement.

3. Relevant Experiences

In order to sketch some of the main elements of an evaluation system that promotes improved design and implementation of social policy, it is useful to document some recent experiences that highlight the critical elements necessary for triggering the virtuous circle mentioned above. This section draws from the author´s experience in Mexico during the past 10 years to identify some of those elements. Mexico made important strides in going from a system based on measuring expenditures and inputs before the year 2000, to having one of the most developed evaluation systems in Latin America today.

The experiences refer to five practical cases where evaluation instruments, practices, and institutions were introduced in a short time span.[13] They include the introduction of evaluation as a general practice in social development policy, the definition and implementation of poverty measures, the creation of an evaluation system at the school level, the introduction of a national academic assessment test at the high school level, and the creation of the National Council for the Evaluation of Social Policy (CONEVAL). From each of these experiences we can draw important insights into the design of a more

[13]The examples are cases where the author personally played an active role. Since no first-hand example is available for illustrating the role played by external donors or investors, we are not able to include an illustration for this case.

comprehensive approach that aligns incentives and objectives among the five actors discussed in the previous section.

The Evaluation Program at the Ministry of Social Development: Evaluating One Program Is Useful, but Not Enough

Policy making is about choosing from a set of alternative programs, services, or actions to generate a predetermined effect in a specific environment, under specific budgetary, political, and institutional constraints. Having information about the impact of one particular program is certainly a useful first step, but for addressing multidimensional issues such as human welfare it is clearly insufficient.

In 1998, Mexico launched the PROGRESA conditional cash transfer program, which among several innovations featured an impact evaluation designed and implemented from the outset. In 2000, evidence had already been generated on its positive impact on education, health, and nutritional outcomes, which played a key role in its survival as an important poverty alleviation program during a change in administration during the same year (in fact, the only change introduced to the program during the transition was modifying its name to Oportunidades). Historically, one of the first actions taken by a new administration in Mexico had been to eliminate previous social programs and substitute new options for them, even though the same political party ruled the country for more than 60 years. This tradition, added to the fact that the government installed in the year 2000 was from an opposition party (the first transition from one party to another in decades), makes the prevalence of PROGRESA even more surprising. Having an impact evaluation that demonstrated the social value added of the Program made it politically costly to simply cancel it, since none of the potential policy alternatives could account for similar effects. Furthermore, the evaluation exercise inspired many others in Latin America and other regions.[14]

Having a high-quality external evaluation at hand, especially when the immediate precedent is decades of policy without research, can make a huge difference. For the first time, evidence on the impact of each unit of budget becomes available, and this not only reassures that an adequate use of resources is being made, but it becomes the main argument for scaling up and increasing investment. One important feature of experimental designs of the kind used by the PROGRESA-Oportunidades and other

[14]See Skoufias (2005) for a review of this initiative. Several evaluations that followed PROGRESA´s have been summarized by Rawlings (2005). Some recent examples are also included in Volume no. 3 of the *Journal of Development Effectiveness*, 2010, and commented on by Gaarder (2010).

similar evaluations is that they compare the effect of the program with respect to a counterfactual defined as a situation with no intervention. While the treatment group receives the full program, the control group is monitored with no benefit for the household whatsoever. Given that these kinds of poverty alleviation programs involve substantial cash transfers to poor families, it would actually be surprising to observe no effect. For this reason, detailed evaluations are designed to capture not just general straightforward income effects (for instance, improved consumption), but the longer-term impacts on school attendance, health status, and nutritional conditions, which are the final goal of the program and which may not necessarily be observed unless the transfer modifies the behavior of household members.

As already mentioned, policy making is about choosing among different alternatives in different circumstances, and in multidimensional phenomena it is also about modifying different welfare dimensions. Conditional cash transfer programs may well generate a positive impact on the household´s human capital accumulation process, but they may be criticized as possibly creating dependence by not enhancing the current income earning capacity. Other approaches such as micro credit for creating self-employment, investment in social infrastructure, or even subsidies for producing goods or services could be complementary alternatives, but when information on their impact is nonexistent, choices become severely limited.

This is precisely what happened in Mexico, and is a clear example of the role played by high-level decisionmakers in promoting evaluation production and use. After the PROGRESA-Oportunidades results became available and were widely used for identifying areas for the program´s improvement for several years, the natural question that emerged was whether this was really the best possible use of public resources. To answer the question, the government decided to introduce similar evaluation designs for other social programs. Better decisions over allocating resources and choosing which programs to expand and which ones to scale back couldn't be made until information on the impact of many programs becomes available.

One case was the comparison between the nutritional supplement provided by PROGRESA-Oportunidades and another nutritional supplement distributed through a subsidized milk program (Liconsa). In a series of impact evaluations, González de Cossio et al. (2009) found that actually the Liconsa supplement was more effective in producing similar outcomes, which motivated the government to take the decision of upgrading the PROGRESA-Oportunidades supplement to meet its composition. Other comparisons of different instruments evaluated with similar methodologies can be found in CONEVAL (2009) and, as can be verified, having a set of evaluations produced systematically allows

policymakers to choose among a set of alternatives with different documented impacts under different scenarios.

Going from a single evaluation to an evaluation program that generates comparable data from different interventions has its own complexities. It requires shifting from measuring inputs and expenditures to assessing outcomes and to creating a critical mass of knowledge to inform policy, with important consequences such as the elimination of programs that have existed for years or decades, or to drastically changing their operation. Introducing these innovations is a decision taken at the highest levels, and its feasibility depends, on the one hand, on demonstrating that the social benefits are higher than the cost, and on the political context, including the power of interest groups and constituencies that become the losers of the process. On the other hand, it depends on the speed at which program operators and practitioners adapt to new rules of the game, which can be a tortuous process on its own. The experience of Mexico illustrates that after the shock of shifting to a new culture of evidence-based decision making in social policy, eventually evaluation practices start being assimilated, and although resistance, especially at the field level, can be prolonged, the participating actors at some point internalize the new culture and procedures.

The Introduction of Poverty Measurement: Evaluation Requires Investing Political Capital

Apart from generating evidence on the individual impact of specific actions, normally a country´s or government´s performance is also evaluated on achievements at the aggregate level. Variables such as unemployment and inflation rates, consumption levels, average wages, etc., are aggregates of microdynamics that provide a general view of how the economy is working, and in the 21stcentury it would seem unthinkable to be able to manage a country´s macro policy without permanent information on their evolution. However, in the case of social policy, it is relatively only recently that similar data have become widely produced. For instance, in Mexico, until the year 2002, official poverty statistics had never been available. Perhaps the main explanation was that transparency on these grounds was a highly sensitive issue. The country had been governed by a single political party for more than 60 years, and generating evidence of unacceptably high poverty rates would have been politically very costly.

The change of government in 2000 to an administration of an opposition party for the first time in decades opened the possibility of generating official poverty statistics for the first time. One important motive was documenting the social conditions of the country after so many years of a single-party system. This is a case where the two main actors

were high-level decisionmakers promoting the creation of a poverty measurement system on the one hand, and the role of evaluators on the other. The Ministry of Social Development had taken the decision to announce officially poverty statistics in the shortest possible time span—as close as possible to the first years of the administration—and invited a group of respected researchers to propose a methodology that would be solid and rigorous enough to assure credibility and restrict the debate to the use of the information, and not to measurement issues. The seven scholars were asked to deliver their proposal one week after they convened for the first time, but it wasn't until one year later that their deliberations allowed defining a methodology that would be technically solid, easy to communicate, clear, replicable, and useful for the design and evaluation of social programs. In this case, evaluators were able to set the pace and orientation, since their participation was critical for the success of the initiative.[15]From the government´s point of view "investing" in having a plural group validating and supporting the methodology was crucial to dealing with the debates that followed.

The official estimate based on the official methodology was that 53.8 percent of the Mexican population was poor in the year 2000. The perception of unacceptably high poverty rates triggered an unprecedented public debate on the performance of the country, the costs of a nondemocratic system, the economic model followed in the recent years, etc. The media played a key role in fueling the discussion and it became by far the most discussed social policy issue in many years. Even though the data clearly represented the situation before the start of the new administration, substantial political capital had to be invested in supporting the exercise and guaranteeing its continuation. The government´s critics attributed the high poverty rates to the current authorities, while opposition parties (mainly members of the party ruling the country for the previous decades) worked heavily to discredit the figures using the argument that they had been manipulated politically.[16]

[15]Székely (2006) presents a detailed account of this process.

[16]Clarity in the time frame for presenting and interpreting data continues to be a critical issue in Mexico and other countries with other indicators such as the results of the Program for International Student Assessment (PISA) that is coordinated by the Organisation for Economic Co-operation and Development (OECD). The PISA examination is applied every three years to a sample of 15-year-olds in mathematics, science, and language, and is published about one year after it is held. Clearly, the results for 15-year-olds reflect their schooling history since grade 1—for instance, the 2009 data reflect policy decisions and school quality at least between 2000 and 2009. In Mexico, the recent publication of the (not positive) 2009 results was completely attributed to the current administration´s failure to improve quality. Moreover, in dimensions such as education and health, bad quality or insufficiencies in the provision of the service at early ages may limit the effect of further improvements, and even substantial advancements in policy design and implementation may not be reflected at all in current data.

The publication of the 2002 figures was less controversial, and after four rounds of official measurements in 2004, 2005, 2006, and 2008, the publication of poverty rates, similar to inflation, employment, wages, etc., has become a much less politicized activity that is systematically used to assess the performance of the country and its government; most important, it has provided a broad framework where the individual impact of specific programs can be viewed in a wider perspective. Apart from measuring their impact at the household level, it is now possible to trace down the effect of programs such as Oportunidades and others on the dynamics of aggregate poverty over time.[17]

The System for Planning and Evaluation at the School Level: Time, Training, and Capacity Building and Resources to Implement Change Are Fundamental for Making Evaluation Usable

Intensive training, capacity building, and providing the necessary time and conditions for operators/executors of public goods and services can be critical factors in creating the conditions for using evaluation results for improving public action. An illustrative example is the introduction of new procedures for improving school management.

In education, any change to the status quo, even at the classroom level, can require huge efforts and resources to guarantee continuity and assure improved results. It is a sector with a high degree of complexity for generating and using evaluation, because apart from usually being the largest entities in the public sector, decisions tend to be highly politicized and subject to the power of interest groups, among which teachers unions are generally the strongest.

What could initially seem to be a simple process whereby teachers receive feedback for improvement, modify their practice, and assess its effect continually can in practice be highly complex, especially if the central actors do not acquire the necessary capabilities and tools for performing the change, or if they are not adequately informed about the nature and importance of the modifications. In the end, even what could seem to be slight improvements require changes in behavior at the individual level that can take time and resources to become a reality. The complexity arises because "using" evaluation results in this context requires internalizing a recommendation or mandate emerging from an evaluation, and then modifying individual behavior and day-to-day practices in order to change the outcome.

[17]The study by Székely and Rascón (2005) is an illustrative example.

Complexity is exacerbated when change implies threatening the power position of some of the actors involved in the delivery of the services. For instance, in 2008, the Education Ministry in Mexico decided to introduce a new system where each school principal (at the high school level) was provided with a spreadsheet and was required to feed in general administrative data that automatically generated 15 selected indicators of the infrastructure, materials, equipment, alumni, and teachers of their own school. Principals were required to perform two exercises. The first consisted in prioritizing the 15 variables in terms of the importance for improving the quality of education in their context and circumstances. The second was setting a target for improving each indicator during the school year. Starting from a situation of total absence of data generated by schools, the underlying purpose of this activity was to generate a diagnosis (baseline) of school conditions, as well as explicit targets, in order to design a school improvement plan for the year. In the context of our discussion, the example is relevant because it constitutes a case where critical actors for the operation of the system (program operators and practitioners) are required to generate inputs for evaluation and at the same time use them to take specific action for improvement.

This apparently simple exercise of setting priorities and targets was not accompanied by adequate training, information, or capacity building under the assumption that school principals had already developed managerial capabilities to perform what seemed a basic and simple task.

After the first year of implementation of the evaluation and planning system, principals were classified in four different categories. The first included the minority of 3 percent that actually completed a high-quality diagnosis and set useful targets to develop an action plan for improvement for the academic year. About one-half of the school principals in this group actually reached their targets, while the remaining half used the system to fine-tune their planning for the following academic year. The second group included a substantial 17 percent of principals who did not perform the exercise at all. Most of them had participated in a pressure group opposing the introduction of the system on the grounds that they already had a good planning and evaluation system that worked well for them. Clearly, transparency and accountability were a considerable threat for this group, and after repeatedly refusing to participate, they were replaced by new principals selected by a new competition process. There were two key elements to managing opposition by this group and move forward with the initiative. The first is that, before the planning system was introduced, new procedures for appointing and selecting new principals were defined. The procedure consists on an open call to teachers with at least five years of experience and who fulfill a series of requirements that are examined and selected through a rigorous and transparent process that prioritizes academic and managerial skills. The second was the possibility of removing principals. The

administrative procedure for firing and hiring school principals at the high school level in Mexico is highly flexible and with little participation of the (very powerful) teacher´s union. This made it possible to maneuver through the process.

The third group, accounting for 52 percent and characterized by being the most experienced and older principals, engaged in a more professional process in priority setting, but deliberately defined extremely low targets, under the belief that low targets were going to be much easier to achieve. Finally, the fourth group included 30 percent of the total, and consisted of principals with younger and less experienced profiles. The commonality among them was deliberately setting extremely higher targets with the idea that this would make them look ambitious and dynamic before central authorities.

After the use of the system in the first academic year, and the renovation of leadership in the 17 percent of schools that had not participated in the first round, interesting dynamics were observed, with the third and fourth groups slowly converging to setting more realistic and meaningful targets, and most important, using the system as a planning device to identify areas for improvement and for demanding support from central authorities in specific areas. One explanation is that this gradual shift was accompanied by more intensive training, information flow, and capacity building at the level of the service provider (the school); it illustrates the difficulties of generating evidence for better decision making and actually using it for improvement. In developing countries, dealing with severe limitations in capacity for implementation and use is most certainly the rule rather than the exception, and is likely to be one of the strong determinants explaining the paradox of "under use of scarce knowledge"."

An additional feature emerging from this case is that once school principals are engaged in the diagnosis-design-benchmarking-implementation-evaluation cycle, resources for introducing improvements can become an important bottleneck. Even when the cultural shift to a knowledge-based system has been accomplished, if enough resources for introducing improvements do not become available, rather than generate a virtuous circle leading to higher quality, the system can lead to increasing stress and frustration, which might make use infeasible in reality and future production of evidence irrelevant.

Evaluating Education Attainment at the High School Level: In Order to Use Information, It Is Necessary to Understand It and Generate the Capacity to Employ It

There is a large literature analyzing the determinants of a set of factors on school attainment that has been influential in policy design, mainly in developed countries. The range goes from analyzing the effect of class size to modified contents, teacher training,

or the use of technology.[18] The evidence for developing countries is much more limited, but in recent years standardized national tests for measuring educational attainment have started to proliferate. In Latin America, at least seven countries regularly hold these types of examinations, and Mexico was one of the newcomers in 2006 for primary and secondary schools and in 2008 at the high school level. The experience in introducing these assessments is illustrative of another feature of the dynamics of production and use of evaluation evidence.

In 2008, the federal government launched the national examinations at 12th grade (exit from high school), with more than 96 percent of schools participating. As in other cases where the precedent is marked with information scarcity, the first time that results were published a huge reaction in the media and public opinion emerged on the one hand, rightly criticizing the low levels of achievement, and on the other, identifying and aggressively attacking underperforming schools. The natural reaction of schools (mainly low performers) was to discredit the test, and strong opposition also emerged from the teacher´s union, which felt aggravated by the exposure and criticism.

This is an example of how constituencies (e.g., parents) and public opinion can play a critical role in continuing with evaluation efforts despite the opposition generated by operators and practitioners. The high profile of discussions around education results can in part be deemed as the immediate natural consequence of introducing transparency and accountability in the schooling system for the first time.

The main feature of this process is that even though low-performing high schools have been under intensive public pressure to improve their results, after three rounds of application in 2008, 2009, and 2010, their capacity to absorb and internalize the information to take action for improvement has seemed extremely slow. Some schools have chosen to go beyond criticizing the test to openly opposing its application, and after a field exploration by the author (of 10 nonrepresentative schools in Mexico City), the common complaint across the different cases is that, while schools were provided with an initial diagnosis and recurrent evaluations thereafter, they were not provided with

[18] See, for instance, the analysis of the relationship between teacher and school quality by Sanders and Rivers (1996), Jordan etal. (1997), Rivkin etal. (2005), Jung (2005), OECD (2005), and Hanushek etal. (2005), and Hanushek and Woessman (2007), among many others, and Barber and Moushed (2008) for a wider review of the determinants of school quality.Haddad and Draxler (2002), Sweet and Meates (2004), Pelgrum (2004), and OECD (2010) present some recent analysis of the relationship between access to technology and schooling outcomes.

guidance or orientation, or resources in order to introduce improvements and perform better in the next round.

This is a case where evidence is substantial, but low use is not due to lack of interest, but due to the low local capabilities to transform information into better practices. If we were to judge the "under use of scarce knowledge"" paradox by this case, the conclusion would be that additional assessments would be welcome, but that without the development of local capabilities at the level of the provider of the services, this valuable information will be permanently underutilized. It is not just a matter of generating and making information available, but of developing the capabilities for its adequate use.

The Creation of the National Council for the Evaluation of Social Programs: The Need to Institutionalize Cultural Change and Creating the Right Incentives

As already mentioned, rather than simply encouraging individual evaluations, the current challenge for countries seems to be creating evaluation systems that promote the generation and use of evidence and knowledge for improving policy. In Mexico, after taking a first step in creating an evaluation program in the Social Development Ministry, Congress introduced the mandate by law of evaluating all social programs that were funded by public resources. The broad concept of "social" adopted in this initiative included poverty alleviation, health, education, agriculture, environmental, micro enterprise, and other related sectors. Simultaneously, a Social Development Law was approved in 2004, which included the formal creation of the CONEVAL.

This was an unusual development for Mexico given its recent history—of about only five years—in producing evaluations. Its main motive, however, had a high political component. Debates had been growing between the ruling party and the opposition, around the possible use of social programs for electoral purposes. Opposition parties promoted the legislation of evaluation practices for the first time mainly as a mechanism to generate credible evidence on the selection of beneficiaries and geographical distribution of resources, although along the way it was realized that this information could be valuable for making better budgetary decisions.

The underlying design was that CONEVAL, which depends on the Social Development Ministry but is external to its bureaucracy, would be allocated the funds for evaluating all social programs, would define a basic structure for their design—for instance, privileging experimental impact evaluation where possible—and would launch a call for proposals among academic public and private institutions for performing the evaluations under certain guidelines. This, added to the mandate by Congress, guarantees that all programs

will be evaluated regardless of the interest or profile of the decisionmakers in charge, or of the operators delivering the goods and services. One important feature is the requirement for the area under evaluation to engage in the design, including suggesting relevant questions and even providing feedback on design aspects.

The main impact of CONEVAL and its evaluations so far has been on transparency and accountability. All evaluations are made public, and their presentation since 2006, which was the first year of formal operation beyond the Social Development Ministry, has caused intensive debate and, most of the time, criticism and discrediting of government action in the media. The process of publication has commonly generated tension and confrontation with other government offices responsible for different programs, especially since the media still tend to highlight whatever negative element arises from the analysis, while ignoring any positive impact or achievement. Tensions reached the highest levels when CONEVAL—also in charge of publishing the official poverty statistics since 2005—released poverty figures for 2008 revealing soaring poverty levels. The institution had to steer through complex situations and questioning from government authorities themselves, since the news generated high political costs. Having been created by congressional mandate and being supported by legislation (through the Social Development Law) have been the main assets of the institution in maintaining its integrity.

After five years of operation, the main challenge remaining is to go beyond the transparency and accountability benefits to guaranteeing that results are used for improving policy. Congress mandates all programs to be evaluated, but has not made a stand on using the information. Still, until 2010, the definition of the public budget for 2011, which by law is defined by Congress, has been guided by inertial elements and political arrangements rather than by policy effectiveness and efficiency. Until budget decisions are tightly linked to evaluation results, it is unlikely that the social sector in Mexico will fully reach the stage of knowledge-based policy.

There are at least four important lessons from the CONEVAL experience. The first is that in many settings, evaluation can be more of a political than a technical issue. In fact, its motivation can be to generate enough transparent information to restrict the discretional use of resources, rather than maximize their impact. The second is that a solid legal foundation for evaluation can make a huge difference and can literally be the determinant in sustaining efforts in the long run, especially when evaluations do not bring good news to decisionmakers. The third has do to with the fact that the production and use of evaluations take time, among other things because they trigger a change in culture that requires adapting day-to-day practices and entrenched procedures. The fact that the Mexican Congress still does not use the information generated from evaluations beyond

marginal budgetary decisions—which is an improvement with respect to no use at all—illustrates this point. The fourth lesson is that independence is critical. Had CONEVAL been independent from the central government, it could have devoted valuable resources to fulfilling its mandate rather than dealing with reactions generated by the results it offers. A critical next step is granting autonomy for the institution.

4. Toward Results-Based Social Policy Design and Implementation

The five examples presented above illustrate some of the real world practical complexities of systematically generating and using evidence from evaluations for improving social policy. The case of Mexico, as many other countries, is still far from an ideal setting where enough information is available on the impact of a variety of interventions in different contexts; enough technical capacity has been developed by decisionmakers and program operators to use information productively and identify the best set of interventions given the characteristics of the populations to serve and the context of reference; and good judgment and common sense are prevalent for correctly identifying the political, human, and financial restrictions for implementation.

Unfortunately, most have only reached the second stage described by Anderson (1966) of "policy with some evidence," and cases where decision-making goes through the diagnosis-design-implementation-evaluation-analysis-finetuning-implementation, and so on, described several decades ago by Greenberg (1968),are actually more an exception than a rule.

The current concerns outlined above and the apparent paradox of "under use of scarce knowledge""may provide an opportunity to move toward the third stage of knowledge-based social policy. This paper has discussed some of the limitations of program evaluations (in Section 1); the different motives, incentives, and dynamics of the different actors involved in the generation and use of evaluations (in Section 2);and some of the real-world complexities in generating and using evidence for improving social programs (Section 3). Following this analysis, we propose that a desirable next step in developing countries is to move toward Results-Based Social Policy Implementation and Design systems (RBSP) at the national level, which could be thought of as complete systems—as opposed to focusing only on individual evaluation efforts—that provide incentives for generation and use, with the support of an adequate institutional setting.

Based on the experience from Mexico, the four central elements of such a system would be:

1) Decisionmakers and personnel in charge of program operation

2) External agency in charge of evaluation
3) External agency in charge of training, information flow, and capacity building
4) Technical bodies in Congress analyzing evaluation results and determining budgets accordingly.

The first actor in the RBSP system is decisionmakers and program operators. Their role is to set targets and get involved in evaluation design, implementing programs and actions, and using evaluation results for improving the performance of their activities. A second actor (institution) would be similar in spirit to CONEVAL, and would concentrate on four activities: defining methodology and evaluation approaches in coordination with the first actor; launching calls for evaluators and selecting those that will perform the evaluation; monitoring the quality of each evaluation; and analyzing evaluations to produce reports that identify areas and specific actions for improvement to be distributed to the first actor. The third actor in this scheme would be an additional (new) public institution totally focused on capacity building, training, and coaching potential users to assure that results can be translated into action. It seems desirable to de-link this activity from the design responsibilities (carried out by the second actor) to avoid conflict of interest. Countries a with strong civil service tradition may already operate along these lines by providing training to civil servants before or while in service, but for others where this is not the case, an institution playing this role specifically might be a more realistic target than waiting for the development of a full civil service career. The fourth actor is the entity that determines public budgets. In the case of political systems where Congress plays this role, implementation of an RBSP system would require strict discipline in linking evaluation results to future funding. There will evidently be other important factors that must be taken into consideration for guiding budgeting decisions, so a minimum share of the budget to be distributed under this scheme could be established as an initial step.

A simple RBSP system of this type could develop the capacity to address most of the challenges identified in the previous sections. For instance, the external agency in charge of implementing evaluations could make sure that the methodological and data issues discussed in Section 1 are properly addressed in each assessment. It would also concentrate evaluation resources to assure that relevant programs are subject to a proper process, which would substitute for the decisions by donors and policymakers (discussed in Section 2) who might even deter evaluation efforts in some circumstances, and would assure that rather than individual evaluations, an evaluation system emerges. This would also promote the creation of a "market for evaluations," where enough funding is provided to carry out rigorous assessments, and evaluators are monitored and kept in line with the central objective of the evaluation activity. The RBSP would also provide a solid

setting for resisting interest group pressures and opposition (for instance, in the case of program operators and practitioners). An important effort would have to be made to "educate the public" in terms of guaranteeing that the publication of evaluation results has an adequate balance between challenges and achievements, to avoid the exclusive focus on bad results or negative outcomes.

The role of the third actor, concentrated on training, informing, and capacity building, would be critical for success. As is evident from the examples in Section 3, even well-intentioned operators and practitioners are not necessarily able to implement the changes or suggestions derived from evaluations. Sustained efforts to assure that policymakers and operators are sufficiently informed and have developed the necessary tools for this endeavor are essential.

Finally, attaching results to budgets is obviously a central requirement. Congress or the entity defining public budgets also needs training and capacity building to play this role adequately. Having technical bodies that can perform this task and support them is critical. The third actor could also be mandated to assure that these capacities are in place. If budgets are aligned to evaluation results, at least to some significant extent, the move toward an RBSP would surely be faster and smoother.

Evidently, building a full RBSP like the one suggested here is not an easy or quick task. It requires, among other elements, strong institutional settings, resources, and technical capacity. However, having this as an image toward which it would be desirable to evolve can be a useful benchmark in assessing how far social policy is from reaching a stage where incentives are more aligned to generate better outcomes.

This kind of system would be complementary to recent international initiatives such as the 3ie, which as already mentioned is devoted to concentrating and summarizing evidence on what works, finances evaluations of programs that promise relevant evidence, and sponsors public programs with similar characteristics, in order to contribute to the construction of a knowledge base for improved social policy. In the same way that national RBSP systems could benefit from the 3ie initiative, the 3ie could pull local information to make it available at low cost for any interested party.

5. Conclusions

This paper describes what could be called a Results-Based Social Policy Design and Implementation system, where programs and actions are continually upgraded and fine-tuned to assure better development outcomes. The system consists of four elements that could provide an adequate institutional setting while providing more incentives to use and

produce evidence for improving policy action: policymakers and program operators; an external agency in charge of assuring that all programs are evaluated with common standards and methodologies; an agency responsible for training, providing information, and building capacity for using evaluations in the public sector; and an external entity (technical bodies in Congress) that defines public budgets based on evidence.

The RBSP is thought of as a possible response to recent concerns, on the one hand, that too little evidence is available on what works in development policy, and on the other that too little of the available knowledge is used in social policy making. The literature has pointed out at least eight different attractive dimensions that make it evident that evaluation can be a powerful instrument for achieving multiple goals. We argue that the apparent paradox is not the result of lack of interest but may be the consequence of two issues. The first are the limitations of the evidence produced on the impact of specific policies on development outcomes, including population heterogeneity, differences in context, general equilibrium effects, the multidimensionality of development, intertemporal impacts, identifying the underlying mechanisms causing effects, and isolating the impact of policy action. There have been important methodological developments to address these issues, but even so, they may deter interest in generating evidence or in using it for real-world decisions. The second has to do with the motivations, constraints, and technical capacity to generate and use evaluation in a profitable way. External donors/investors, policymakers, evaluators, program operators, and the general public all have different incentives, motives, and goals, that might be in contradiction and generate an underutilization and undergeneration of evidence.

The RBSP system, although perhaps out of reach for many countries at the moment, could be a useful reference point for evolving from what seems to be generally the status quo of "policy with some evidence" toward much more efficient settings where knowledge is continually generated to assure that development resources improve the standard of living of large sectors of the population and provide them with better prospects for the future.

Because development is a moving target, the generation and use of evidence on what works, and under what circumstances, should also be in continual evolution. The concept of Results-Based Social Policy Design and Implementation may be a useful target to keep pace with it.

Bibliography

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

Alkin, Marvin C.,& Daillak, Richard H. A Study of Evaluation Utilization. Educational Evaluation and Policy Analysis, Vol. 1, No. 4 Jul.-Aug, pp. 41-49. 1979.

Alkin, M. C., Daillak, R., & White, P. Using evaluations: Does evaluation make a difference? Library of Social Research #76. Beverly Hills: Sage Publications. 1979.

Anderson, O.W. "Influence of Social and Economic Research on Public Policy in the Health Field: A Review." Milbank Memorial Fund Quarterly, Vol 44, No. 3, pp. 11-51.1966.

Attanasio, O., C. Meghir, & M. Székely. "Using Randomized Experiments and Structural Models for "Scaling Up": Evidence from the Progresa Evaluation."I In Annual World Bank Conference in Development Economics, F. Bourguignon, and B. Pleskovic, eds., Oxford University Press. 2004.

Avery, C., Resnick, P., &Zeckhauser, R. The Market for Evaluations. The American Economic Review, Vol. 89, No. 3, pp. 564-584. American Economic Association. 1999.

Banerjee, A. "Making Aid Work." MIT Press. 2007.

Barber, M., & M. Mourshed. "Cómo Hicieron los Sistemas Educativos con Mejor Desempeño del Mundo para Alcanzar sus Objetivos." Programa de Promoción de la Reforma Educativa en América Latina y el Caribe (PREAL), No. 41, Santiago de Chile. Julio 2008.

Barkdoll Gerald L. Type III Evaluations: Consultation and Consensus. Public Administration Review, Vol. 40, No. 2 Mar. - Apr, pp. 174-179. 1980.

Birdsall, N.,& W. Savedoff, "Cash on Delivery: A New Approach to Foreign Aid." Center for Global Development. 2010.

Braskamp, L. A., & Brown, R. D. Utilization of evaluative information. San Francisco: Jossey-Bass. 1980.

Brown Robert D., Newman Dianna L.,& Rivers, Linda S. A Decision making Context Model for Enhancing Evaluation Utilization. Educational Evaluation and Policy Analysis, Vol. 6, No. 4 Winter, pp. 393-400. 1984.

Caro, Francis G. Issues in the Evaluation of Social Programs., Review of Educational Research, Vol. 41, No. 2, Science and Mathematics Education, Apr, pp. 87-114.1971.

CONEVAL, "Informe de seguimiento a los Aspectos Susceptibles de Mejora de Programas Federales 2009, Proceso de Evaluación externa 2009 del Gobierno Federal", México, 2009.

Cummings, R."Rethinking Evaluation Use." Australasian Evaluation Society, November 2002.

Deaton, A. "Evidence-based aid must not become the latest in a long string of development fads." *Boston Review*. July 2006.

Deaton, A. "Understanding the Mechanisms of Economic Development." *Journal of Economic Perspectives*, 24(3): 3–16, 2010.

Drummond, M. Making Economic Evaluations More Accessible to Health Care Decision-Makers. The European Journal of Health Economics, Vol. 4, No. 4, pp. 246-247. 2003.

Ferman, L. A. Some perspectives on evaluating social welfare programs. Annals of the American Academy of Political and Social Science. 1969.

Gaarder, M. "Introduction," Journal of Development Effectiveness, Vol. 3, No. 3.2010.

Gaarder, M., Glassman, A., &Todd, J. "Conditional cash transfers and health: unpacking the causal chain." Journal of Development Effectiveness, Vo. 3, No. 3.2010.

Gonzalez de Cossio, T., Rivera, J., &López Acevedo, G. "Nutrición y Pobreza." World Bank, Washington DC. 2008.

Greenberg, B. G. Evaluation of social programs. Review of the International Statistical Institute. 1968.

Haddad, W., & A. Draxler. "*Technologies for Education: potentials, parameters, and prospects.*" UNESCO, Paris. 2002.

Hanushek, E., J. Kain, D.M. O'Brien, & S. Rivkin. "The Market for Teacher Quality."Working Paper 11154, National Bureau of Economic Research, Cambridge, Massachusetts. 2005.

Hanushek, E., & L. Woessmann. "The Role of School Improvement i9n Economic Development." *CESifo Working Paper* No. 1911. February 2007.

Henry, G.,& Mark, M. “Beyond Use: Understanding Evaluation´s Influence on Attitudes and Actions.” American Journal of Evaluation, Vol. 24, No.3, pp. 293-314. 2003.

Heckman, J., Smith, J., & Clemens,N. “Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts.” Review of Economic Studies, 64, No. 4, pp. 487-535. 1997.

Howard R., Davis, [[initial?]],& Susan E. Salasin. The Utilization of Evaluation.I In Handbook of Evaluation Research, Vol. 1., ed., E. L. Streuning and M. Gutentag, Beverly Hills, CA: Sage. 1975.

John, E., S. Lawrence, & Thomas J. Cook. Designing Useful Evaluations. Evaluation and Program Planning, 5(4):327-36. 1982.

Jones, N., Jones, H., Steer, L., &Data, A. Improving impact evaluation production and use. Overseas Development Institute. 2009.

Jordan, H., R. Mendro, & D. Weerasinghe, “Teacher Effects on Longitudinal Student Achievement.” *Mimeo,* Indianapolis, IN. 1997.

Jung, I. “ICT-Pedagogy Integration in Teacher Training: Application Cases Worldwide.” *Educational Technology &Society*, 8 (2), pp. 94-101. 2005.

Koretz, Daniel. Developing Useful Evaluations: A Case History and Some Practical Guidelines.In New Directions for Program Evaluation: Making Evaluation Research Useful to Congress, ed. L. Saxe & D. Koretz, San Francisco: Jossey-Bass, pp. 25-50.1982.

Kirkhart, K.E. “Reconceptualizing Evaluation Use: An Integrated Theory of Influence.” Chapter in VJ Caracelli and H. Preskill (eds), “The Expanding Scope of Evaluation Use.” New Directions for Evaluation, No. 88. 2000.

Leonard, Rutman. “Barriers to the Utilization of Evaluation Research.”Paper delivered at the annual meeting of the Society for the Study of Social Problems, Chicago, IL, Sept. 1977.

Levinton L. A., & Hughes, E.F. Research on utilization of evaluations: A review and synthesis. Evaluation Review, pp. 525-548. 1981.

Mark, M., & Henry, G. “The Mechanisms and Outcomes of Evaluation Influence.” Evaluation, No. 10. 2004.

Marra, M. “How Much Does Evaluation Matter: some examples of the utilization of the evaluation of the World Bank´s anti corruption activities.” Evaluation Vol. 6, No. 1, pp-22-36. 2000.

OECD. “*Teachers Matter; Attracting, Developing and Retaining Effective Teachers.*” Organisation for Economic Co-Operation and Development, Paris. 2005.

OECD. “*Are the New Millennium Learners Making the Grade?*” Organisation for Economic Co-Operation and Development, Centre for Educational Research and Innovation, Paris. 2010.

Patton, M. “In Search of Impact: An Analysis of the utilization of Federal Health Evaluation Research.” Center for Social Research, Minnesota. 1975.

Patton, M. Q. Utilization-focused evaluation. Beverly Hills: Sage Publications. 1978.

Pelgrum, H. “Promoting Equity through ICT: What Can International Assesments Contribute to Help Fight Low Achievement?” “*Promoting Equity Through ICT in Education: Projects, Problems, Prospects,*” Capítulo en Karpati, A., ed., Budapest, Hungarian Ministry of Education, OECD. 2004.

Pittman, G., Feinstein, O., &Ingram, G. “Evaluation Development Effectiveness.” New Brunwsick, Transaction Publishers. 2005.

Pritchett, L. “It Pays to be Ignorant.” Journal of Economic Policy Reform, Vol. 5, No. 4, pp. 251-269. 2002.

Ravallion, M. Evaluation in the Practice of Development. World Bank. 2008.

Rawlings, L and Rubio, G. “Evaluating the Impact of Conditional Cash Transfer Programs.” World Bank Research Observer, Vol. 20 (1), pp. 29-55**.** 2005.

Rivkin, S.G., E. Hanushek, & J. Kain. “Teachers, Schools and Academic Achievement.” *Econometrica*, Vol. 73, No. 2, pp. 417-458. March 2005.

Sanders, W., & C. Rivers. “Cumulative and Residual Effects of Teachers on Future Student Academic Achievement.” *University of Tennesse.*, 1996.

Savedoff, W., R. Levine, & N. Birdsall, “When Will we Ever Learn? Improving Lives through Impact Evaluation.” The Evaluation Gap Working Group, Center for Global Development, Washington DC. 2006.

Sadofsky, S. Utilization of evaluation results: Feedback into the action program. In J. Shmelzer (ed.), Learning in action. Washington: U.S. Government Printing Office. 1966.

Schulberg, H., & Baker, F. Program evaluation models and the implementation of research findings. American Journal of Public Health. 1968.

Skoufias, E. "Progresa and Its Impact on the Welfare of Rural Households in Mexico." International Food Policy Research, Washington DC. 2005.

Sweet, R., & A. Meates. "ICT and Low Achievers: What Does PISA Tell Us?" Capítulo en Karpati, A., (ed), "*Promoting Equity Through ICT in Education: Projects, Problems, Prospects.*" Budapest, Hungarian Ministry of Education, OECD. 2004.

Székely, M. "*Números que Mueven al Mundo: la medición de la pobreza en México.*" Editorial Porrúa, México DF. 2006.

Székely, M., &E. Rascón. "México 2000-2002: Reducción de la Pobreza con Estabilidad y Expansión de Programas Sociales." *Economía Mexicana,* vol. XIV, núm. 2. Second Semester de 2005.

Walter, J. Can Evaluations Influence Programs? The Case of Compensatory Education.*Journal of Policy Analysis and Management,* Vol. 2, No. 2 Winter, pp. 174-184.1983.

Weiss, C.H. "The Many Meanings of Research Utilization." Public Administration Review. September/October, 1979.

Weiss, C.H. "Measuring the use of evaluations." In E. House (ed.), Evaluation studies review annual Vol. 7, pp. 129-146, Beverly Hills, CA: Sage.1982.

Weiss, C.H. "Utilization of Evaluation: Toward Comparative Study, in Readings in Evaluation Research," F. G. Caro (ed.). New York: Russell Sage Foundation, p. 141.1971.

Weiss, C.H. "The Interface between Evaluation and Public Policy." Evaluation, Vol. 5, No. 4, pp. 468-486. 1999.

Weiss, C.H. "Theory-Based Evaluation: Theories of Change for Poverty Reduction Programs." Chapter in O. Feinstein and R. Piccioto (eds.), "Evaluation and Poverty Reduction." Transaction Publications. 2001.