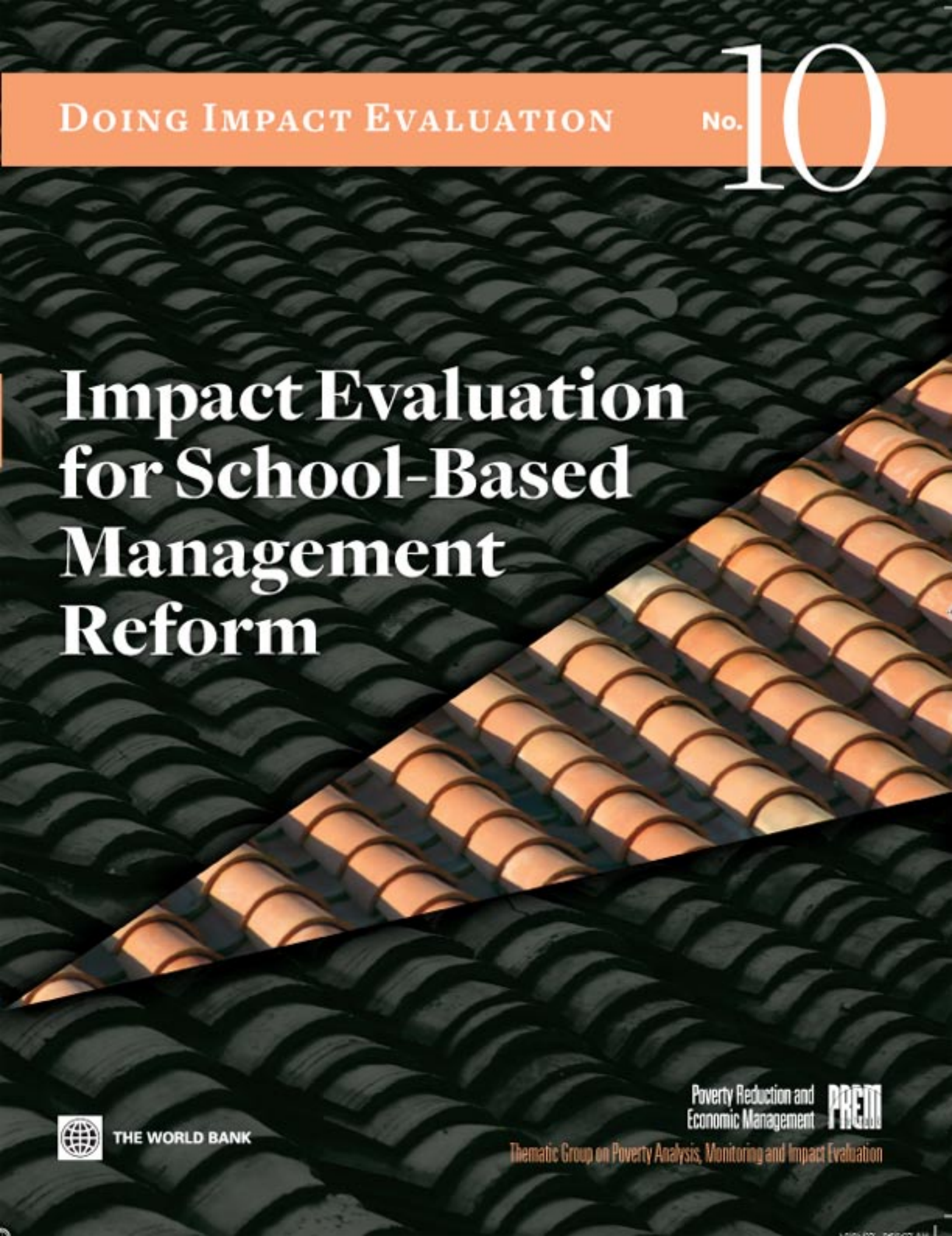
DOING IMPACT EVALUATION No. 10

Impact Evaluation for School-Based Management Reform



THE WORLD BANK

Poverty Reduction and Economic Management PREM

Thematic Group on Poverty Analysis, Monitoring and Impact Evaluation

Impact Evaluation for School-Based Management Reform

December 2007

Acknowledgement

This paper[1] was written by Paul Gertler,[2] Harry Anthony Patrinos,[3] and Marta Rubio-Codina.[4] The authors thank Felipe Barrera, Thomas Cook, Tazeen Fasih, Vicente Garcia Moreno, Markus Goldstein, and the participants of a World Bank workshop on school-based management for useful comments and suggestions. The work was task managed by Markus Goldstein and financed through grants from the Trust Fund for Environmentally and Socially Sustainable Development supported by Finland and Norway and by the Bank-Netherlands Partnership Program.

[1] First draft: February 2007

[2] Haas School of Business, University of California at Berkeley, gertler@haas.berkeley.edu

[3] The World Bank, hpatrinos@worldbank.org

[4] University College of London, m.rubio-codina@ucl.ac.uk

TABLE OF CONTENTS

Introduction

This report is designed to provide guidance on the design of impact evaluations of school-based management (SBM) initiatives in developing countries. SBM is a reform movement that consists in allowing schools more autonomy in decisions about their management; that is, in the use of their human, material, and financial resources. Also referred to as school based governance, school self management, or school site management, this trend has become very popular over the past decade (Caldwell 2005). Today, countries as diverse as New Zealand, United States, the United Kingdom, El Salvador, Nicaragua, Guatemala, Mexico, Spain, the Netherlands, Hong Kong (SAR), Thailand, and Israel have instituted SBM programs.

Many governments and international agencies are increasingly interested in finding ways to boost learning outcomes and get maximum benefit from their education investments, especially in developing countries. Indeed, education quality continues to be very low in middle- and low-income countries despite the success in expanding schooling access and enrollment in the last decades. Education systems in developing countries are usually highly centralized and have very strong teacher unions. Teachers often lack strong incentives and accountability mechanisms, which results in high teacher absenteeism rates (Banerjee and Duflo 2006; Chaudhury and others 2006). Moreover, many schools lack the basic equipment and school supplies, and many children learn much less than the learning objectives set in the official curriculum.

Not surprisingly, policymakers and researchers in developing countries have shifted their focus to policy reforms that attempt to reduce distortions and inefficiencies in the education system and its institutions. Nowadays, these reform initiatives range from pay per performance schemes that link teacher wages to student performance, to introducing vouchers and other methods to expand school choice, to decentralizing school functions and processes so that local communities have more power to allocate and manage their resources. The World Development Report 2004 claims that placing educational resources, decision-making, and responsibilities closer to the beneficiaries is one approach for the improvement of schools (World Bank 2003). Local communities arguably have the best knowledge about the needs of their children, stronger incentives to monitor the performance of teachers and principals, and a comparative advantage in conducting this monitoring. However, while decentralization reforms appear promising and are increasingly being adopted, rigorous empirical evidence on their impact is scarce (Glewwe and Kremer 2006).

This is partly due to the context in which many SBM reforms have been implemented, at least in the developing world. Sometimes, they have been adopted as a response to crises in the educational system, for instance in Chicago; or to empower teachers; or even – albeit not very often – to ensure educational quality as in Hong Kong. However, many other times, SBM initiatives have been introduced as a political reform to increase school access and transfer power to devastated communities after a disaster, when centralized coverage is unfeasible. For example, the EDUCO (*Educación con Participación de la Comunidad*, Education with Community Participation) program was introduced after

the civil war in El Salvador and the PROHECO (*Proyecto Hondureño de Educación Comunitaria*, Honduran Project of Community Education) program in Honduras after Hurricane Mitch. In both situations, the intervention had to be set up quickly and delaying it to plan a well-thought evaluation design was not practical. As a consequence, many evaluation studies have had to rely on limited data and have struggled to find a valid comparison group of schools that allowed for causal interpretation of the effects.

Hence, reliable and well-conducted evaluations of SBM programs that can lend empirical support to the various claims on the advantages of SBM are needed; and more so, given the increasing number of countries that are adopting these reforms. Rigorous program evaluations can certainly serve several purposes. First, they offer a direct assessment of the impact of the program on the welfare of its targeted population and verify whether funds have been spent as intended. Second, evaluations provide insights on how the intervention is affecting outcomes, thus shedding light on the relative efficiency of implementing one particular intervention versus another. Third, evaluations inform policy decisions on how to improve existing programs and whether to continue and/or expand them to environments different from those for which they were first designed.

Despite the wider evidence in developed countries – the United States, notably (see Borman and others 2003) – this note will focus on developing countries. A focus on developing countries fits better the objectives of this series. Also, and more importantly, SBM reforms in developed countries are applied to schools with very different initial conditions from those in developing countries. While there are surely lessons to be drawn/learnt from the experience in developed countries, comparisons between developed and developing countries could be – at times – misleading.

Thoroughly planning and conducting a good impact evaluation is a long and challenging task that requires three key elements, all of which need one another to exist:

1. An *appropriate model of behavior* that provides a theoretical framework to guide the formulation of hypotheses on the expected effects of the intervention and the mechanisms underlying these effects. Therefore, it is crucial to start by defining the intervention and clearly stating its objectives, targeted population and implementation details.
2. *Detailed micro-level data* over an *appropriate time frame* that measures the response of individual agents (students, teachers, schools) to the proposed program.
3. A *good identification strategy* that allows the measurement of a *counterfactual* – namely, how would the lives of program participants been had they not received the program – in order to attribute changes in outcomes to the program and only the program.

The purpose of this note is to address each of these points for the specific evaluation of impact of SBM interventions. We begin in section I by defining SBM programs and their types, and reviewing the arguments why SBM reforms should or should not be introduced. In section II, we describe common indicators of treatment and outcomes, and discuss potential data sources. Section III reviews the different designs available to target beneficiaries and define a counterfactual, and describes the array of estimation methods that

are operationally feasible under each design. When possible we illustrate the methods with examples from past SBM evaluations, mainly from the developing world. Section IV summarizes the main results from past evaluations. In section V, we address ethical considerations. Section VI concludes with a discussion of outstanding issues for the evaluation of SBM reforms.

I. School-Based Management (SBM)

A. Definition and Goals of SBM Interventions

SBM is the decentralization of authority to the school level. It involves the transfer of responsibility and decision-making over school operations and school management to principals, teachers, parents, sometimes students, and other school community members. The school-level actors, however, have to conform to, or operate within, a set of centrally determined policies (Caldwell 1998). The basic principle around SBM is that giving school-level actors more autonomy over school affairs will result in school improvement as they are in a better position to make decisions to meet school needs in a more efficient manner (Malen, Ogawa and Kranz 1990).

SBM reforms are far from uniform. SBM encompasses a wide variety of strategies, ranging from fully autonomous schools with authority over every educational, financial, and personnel matter to more restrictive versions that allow autonomy over certain areas of school operations. Another dimension of variability revolves around to whom greater decision power and accountability are transferred. Similarly, the goals of SBM reforms vary substantially, although they typically involve: (i) increasing the participation of parents and communities in schools; (ii) empowering principals and teachers; (iii) building local level capacity; (iv) creating accountability mechanisms for site-based actors and improving the transparency of processes by devolution of authority; and (v) improving quality and efficiency of schooling, thus raising student achievement levels. Only recently has SBM been adopted as a mean to an end, which is providing good quality education to students and improving school management, transparency, and accountability. In the early years of SBM, the mere transferring of autonomy and authority to the school local agents was considered a goal on its own.

B. Arguments For and Against the Introduction of SBM Reforms

There are a number of arguments put forth in favor of the introduction of SBM. First, allowing school agents (principals, teachers, and parents) to make decisions about relevant educational issues is believed to be a more democratic process than keeping these decisions in the hands of a selected group of central level officials (Malen, Ogawa, and Kranz 1990). Second, locating the decision-making power closer to the final users will arguably lead to more relevant policies, as local actors generally have better information about local needs, and thus are able to make the best decisions. Third, additional gains in efficiency could come from making the decision-making process less bureaucratic. Fourth, empowering the school personnel and the community might lead to higher commitment, involvement, and effort. This will result in a greater resource mobilization and possibly a

more enjoyable school climate if all different agents involved in the decision-making process cooperate and coordinate efforts. The closer parent-school partnership might also improve the home environment with respect to learning. Fifth, involving parents in school management or in monitoring and evaluation activities is likely to increase the levels of transparency and accountability within the school. This might in turn improve school effectiveness and school quality.

The empirical evidence thus far – although limited in both quantity and quality – seems to support some of these arguments. It has been demonstrated that the quality of education depends primarily on the way schools are managed, more than on the availability of resources (Hanushek 2003). It has also been shown that the capacity of schools to improve teaching and learning is strongly mediated by the quality of the leadership provided by the principal (Caldwell 2005). Both factors would argue for stronger control over management within the school.

However, governments are faced with many challenges in delegating responsibility and power to the school that can threaten the success of the reform. Ex-ante the government has to decide whom to devolve decision-making authority to and to which degree – namely, which functions to decentralize. Moreover, the government has to be able to provide appropriate incentives that will minimize conflicting interests amongst school agents. For example, policies that put school budgets in the hands of the communities might not be very popular amongst school staff, whereas policies that strengthen the role of the principal might gain little sympathy amongst teachers (Wohlstetter and Briggs 1994). Conflicts amongst school agents about the use of funds and the evaluation of performance can have an adverse impact on school quality. Ex-post, the government has to offer an accountability framework that provides support to decentralized schools and ensure enough local capacity to manage the powers and resources transferred.

Two groups are expected to be the main guarantors of the successful implementation of SBM reforms: senior teachers, especially the school's principal, and the parents – and, at times, the wider community (De Grauwe 2004). However, it is wrong to presume that school staff is always ready and willing to undertake the reform. SBM has in several cases made life harder for school principals by increasing their administrative and managerial workload, to the detriment of their role as a pedagogical leader (Caldwell 1993; Odden and Odden 1994; Wylie 1996). In addition, many of the management-related decisions SBM reforms involved – especially financing and staffing issues – are intricate and complex. With regard to the community, its involvement in school life might also impose considerable coordination and time demands. These can represent a significant cost for low-income parents who might have to forego some wage-earning work time to participate in the school committees. Moreover, in communities with many social and political tensions, the school committee can become an instrument in the hands of an elite group, and no increased transparency and accountability will be achieved. Given these potential problems, additional rigorous evidence is needed to examine the impacts of different ways of implementing SBM.

C. Types of SBM Interventions

SBM is a very broad concept. It includes a variety of interventions and experiences that admit many different classifications. A first classification is according to whom in the school is authority transferred. Caldwell (1998) draws a distinction between *school-based management* and *school-based governance* initiatives. The former applies to initiatives that transfer responsibilities to professionals within the school, generally the principal and senior teachers, whereas the later implies giving authority to an elected school board, which represents parents and the community. Similarly, Leithwood and Menzies (1998) identify four types of SBM reforms:

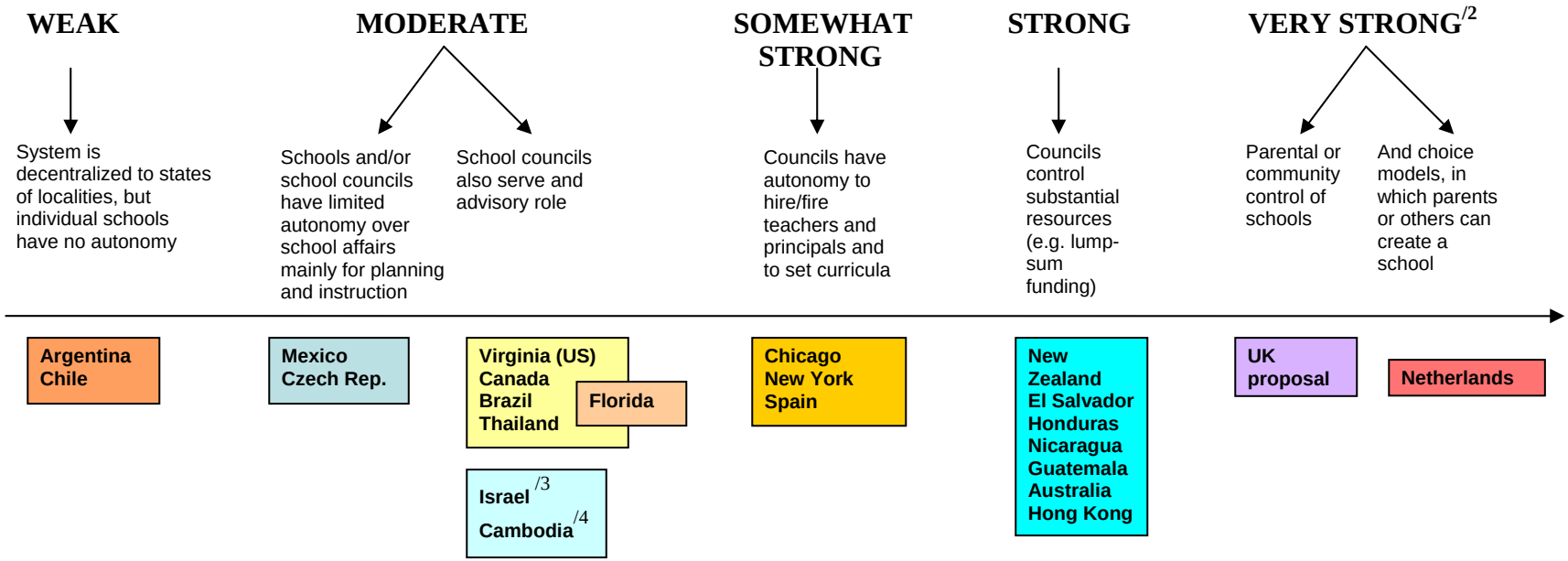
1. *Administrative control reforms*: the principal is the key-decision maker. The reform is intended to provide more accountability and improve the efficient use of resources.
2. *Professional control reforms*: the body of teachers receives the authority. Teacher empowerment is usually the primary objective.
3. *Community control reforms*: the parents or the community are in charge through a parent association. The reform tends to focus on accountability to parents and choice.
4. *Balanced control reforms*: parents, teachers, and principals share responsibilities. Empowering all actors is the main reform objective.

An alternative way of classifying SBM reforms is according to the processes they decentralize and the level of autonomy they transfer. In this case, the diversity of SBM reforms might be better represented as a continuum of reforms that are differentiated by the degree of autonomy granted to schools and to each school agent (Fasih and Patrinos 2006). In this continuum, the range of SBM reforms goes from "weak" reforms that decentralize very little autonomy, over a few areas only, to "strong" reforms in which schools are basically stand-alone units, responsible for almost all decisions concerning what goes on inside their buildings. Any type of reform in the continuum can be evaluated provided the degree of autonomy granted to the school is clear to the researcher.

Figure 1 depicts such a continuum and classifies the countries that have implemented SBM reforms in the various stages of this continuum.[5] For instance, weak to moderate intensity SBM reforms are those in which schools and/or school councils have limited autonomy, usually over areas having to do with instructional methods or planning for school improvement. Such would be the case of schools in the PEC (*Programa Escuelas de Calidad*, School Quality Program) in Mexico. Or of schools in Prince William County (Virginia, US) or in Edmonton (Canada), where councils merely serve an advisory role. As councils become more autonomous, receive funds directly from the central or other relevant level of government (for example lump-sum funding or grants), can hire and fire teachers and principals, or set curricula, SBM becomes a much stronger type of reform. Schools like these can be found in El Salvador and New Zealand. At the end of the continuum are systems in which schools councils or school administrators have full autonomy over the school educational, operational, and financial decisions. Some schools even engage in their own fundraising activities. In these cases, parents or others can even establish fully autonomous public (charter) schools, such as in the Netherlands and the United Kingdom.

[5] Note that the terms "weak" and "strong" are not used to classify any SBM system as better or worse than any other but simply to define the degree of autonomy awarded to the school-based agents.

Figure 1: Classification of SBM reforms implemented in various countries /1



/1 Source: adapted by the authors from Fasih and Patrinos (2006).
/2 These represent ratings in the continuum of autonomy and authority vested to schools by the various types of SBM reforms.
/3 Israeli schools have autonomy to control their budget. School locally-controlled budgets represent a small fraction of total public expenditures because most expenditures are controlled and made centrally. There are no school councils or parent associations with decision-making authority.
/4 Cambodia schools in the EQIP program receive cash grants and have participatory decision making, but schools councils are not formally established.

II. Key Elements in the Evaluation of SBM Interventions

The design of the impact analysis of any social policy should carefully address the following issues:

1. Define the intervention and clearly state its objectives and targeted population.
2. Define what it means for an individual unit (a school, a student) to participate in the program and whether this has changed over time.
3. Establish the relevant outcome measures and a sensible time frame over which we would expect these measures to show program impacts.
4. Establish a sensible strategy to define a counterfactual; this is to say, establish the evaluation design which will inform the empirical analysis.

In the next sections, we provide guidance on how to address each of these points in the impact evaluation of SBM interventions. We start by discussing issues related to the definition of treatment. Next, we comment on the different units of analysis and outcome measures one can consider, and we put forward some timing considerations. Finally, we summarize the data sources one can exploit for the impact evaluation of SBM and comment on sample size determination issues. Section III reviews the array of designs available to target beneficiaries and define a counterfactual and briefly describes the estimation methods that are technically feasible under each design.

A. Definition of Treatment

As discussed in section I, SBM reforms are extremely varied in form. Understanding the intervention and explicitly defining what it means for a school to participate in the program are crucial in the definition of the treatment variable. Other considerations involve ascertaining how the intervention is operationalized and the quality of this implementation. Making sure that only beneficiary schools receive benefits, there is no leakage of resources, responsibilities are effectively transferred to the school, and so on and so forth is key before attempting to identify treatment effects on beneficiary schools.

Depending on what the objective of the evaluation is and assuming that the evaluation will develop as rigorous a counterfactual as possible, one can undertake program evaluations, process evaluations, or combine both types.[6] SBM *program evaluations* measure the overall impact of the intervention on the school and the school community: parents, principals, teachers and, students. They take the SBM reform as a black-box and measure the impact of receiving the SBM package versus not receiving any package at all. It is standard to categorize treatment using a dichotomous variable that equals one if the school (or the relevant unit of analysis) receives the intervention – namely, operates under an autonomous or decentralized mode – at a certain point in time and zero otherwise. While program evaluations are useful and necessary, they are not sufficient to determine which

[6] While in some contexts process evaluations are posed as an alternative to impact evaluations, what we mean here are impact evaluations that measure effects on ultimate impacts (e.g. learning outcomes) while unpacking the effects of different procedural changes.

particular components of the intervention are affecting outcomes, let alone the mechanisms. On the other hand, SBM *process evaluations* put a larger effort into trying to identify the mechanisms by which the SBM reform is affecting outcomes. They decompose the SBM intervention into its different components (autonomy to hire and fire teachers, control over resources, autonomy over school planning and instruction, etc.) and attempt to identify the effects of each sub-component separately. Hence, treatment is characterized by a set of dummies, each of them equal to one if the school receives a particular sub-component and zero otherwise. Process evaluations are clearly more informative on what practices to adopt and mimic in future interventions than program evaluations. Unfortunately, they are also more demanding in terms of data and more challenging in terms of identification. Because several treatment variables are defined, at least one per intervention sub-component, a valid counterfactual for each of them has to be identified.

Some authors have suggested using "de facto" autonomy – as opposed to "de jure" autonomy – as the relevant measure of autonomy (King and Ozler 1998). While "de jure" autonomy refers to whether the school has been appointed as autonomous or not, "de facto" autonomy is related to the level of autonomy the school is actually enjoying or exercising as measured by the number (or the percentage) of decisions the school makes. Alternatively, it is possible to construct an "index of autonomy" using information on the different functions the school reports having a say on: selection of didactic material and textbooks, curricular innovations, criteria for evaluation of teachers and students, infrastructure works, etc. A natural concern in defining the index is how to assign weights to each function. One possibility is to apply principal components or factor analysis techniques. In any event, the problem with "de facto" or effective autonomy is that it is very likely to be correlated with unobserved school characteristics that are simultaneously correlated with outcomes, even if "de jure" autonomy (the SBM reform) has been assigned randomly to some schools and not to some others. We will return to this and other endogeneity issues in section III.

If schools are gradually decentralized, there will be variation in the length of time a school has been under treatment at each point in time. As a consequence, it is possible to define treatment in several ways. A first criterion is to characterize treatment with a dichotomous variable equal to one if the school has received treatment in *all* of the years under evaluation and zero if the school has not received benefits in any year during the evaluation period. This approach will exclude from the analysis those schools that received treatment only in some of the years in the evaluation period. A less strict criterion would be to set the treatment variable equal to one if the school has received benefits in *any* year during the evaluation period and zero otherwise. Both these definitions of treatment, however, ignore the variation in the length of time under treatment and any potential differences between schools introduced earlier versus those phased-in at later dates. Assuming there is no reversion in a school treatment status – namely, no attrition amongst participant schools so that decentralized schools stay decentralized[7] – an alternative is to set

[7] Indeed, most evaluations work only with schools that have received the program continuously since their starting date. However, if an evaluation is to be used to inform a policy decision about whether to continue a program, it should take into account the fact that the program did not continue to be attractive to some participating schools (attriters). One should then include in the treatment group any school that ever participated irrespective of how long for. We return to the issue of attrition in section III.C.

the treatment variable equal to one from the first year the school is given autonomy onwards. Further interacting this dummy with year dummies will pick out differential effects of treatment for schools that were phased-in in successive periods. Two last possibilities are to either define treatment as a continuous variable equal to the number of periods (months, years) the school has been autonomous for at each point in time or to use a set of dummies D_x, each of them equaling one if the school has been autonomous for $x = \{1, 2, \ldots, n\}$ number of periods and zero otherwise.

It is equally important to check whether the intervention or its implementation have changed over time, as these changes might introduce an additional time dimension in the effects. Moreover, if schools are phased-in gradually, there is also scope for the existence of heterogeneous effects between schools intervened at different dates. As noted above, interacting time dummies with treatment status will capture this source of heterogeneity.

B. Unit of Analysis and Outcome Indicators

The natural unit of analysis in the evaluation of SBM interventions is the school. Nonetheless, all members of the school community – students, teachers, principals, teacher-aides, parents – are likely to benefit from the reform more or less directly. Therefore, provided there exist sufficiently disaggregated data, the analysis can be performed at any of these lower levels.[8] The unit of analysis will in turn determine the outcome indicators we should measure impacts on. These indicators must be observable before and after and/or with and without the intervention. Ideally, the evaluation of the SBM reform should not only focus on final educational outcomes (student learning) but should also examine whether the reform has transformed the relationships amongst school principals, teachers, parents, and government officials and the school operations and decision-making processes. One possible classification of outcome indicators is:

Process Outcomes

Process outcomes will be useful to examine whether autonomous schools effectively exercise greater autonomy over their own management than non-autonomous schools and whether this increased influence over school decisions is positively viewed by local stakeholders. They can also be informative about whether the reform has encouraged changes in teaching effort and in pedagogic and operational practices, which might be conducive of a more favorable learning environment.

Purposive surveys can be designed to measure several indicator variables that can fall into the process outcomes category, such as: whether there have been improvements in the school security, infrastructure, or equipment; whether there have been curricular and teaching innovations; and whether training courses on pedagogic and/or managerial matters have been introduced along with the reform. These questions can be either asked to the principal, the teachers, or a relevant member of the parents association – the president, for example. Other relevant questions are those related to parental (or the parents association) and community involvement in school matters and activities. An easy way to measure these

[8] Note that the school is likely to be the primary sampling unit. Hence, any analysis performed at a lower level should cluster standard errors at the school level.

is by inquiring about the number of meetings between parents, teachers, and principals in the school over a certain period of time.

A more direct measurement of the school level of decentralization would be to ask who has the major influence – the central government, the local government, or the school – in decisions related to each school function. In order to find out about the distribution of responsibilities within the school, one can ask how much responsibility each school agent has over each of those functions. This will additionally provide some insight on the level of influence felt by principals, teachers, and parents on school matters. The degree of autonomy they feel and the satisfaction they derive from it will be important determinants of the subsequent effects on student performance, if any. Related to the feeling of influence is the question of excess burden and responsibility felt by parents, teachers, and principals. These variables should be very carefully measured – for example, in terms of the extra time devoted to meetings, participation in school activities, and managerial tasks related to administering the intervention and in terms of principal and teacher turnover – as they will constitute a measure of the indirect costs of the SBM intervention.

Lastly, evaluations might also attempt to measure the impact of SBM on teacher effort and performance. However, teacher behavior is an abstract concept that is extremely difficult to quantify. One approach is to design a teacher survey with a series of questions on each of the following areas:

- First, the *teacher level of interaction with other school members* – namely students, parents, the director, and other teachers. The number of interactions can be measured in terms of number of meetings or number of hours in meetings with each of these agents over the last month, semester, etc.
- Second, *teacher motivation*, which can be measured in terms of: the number of hours preparing for class, grading homework, teaching regular classes, teaching support classes for students that lag behind, and attending training sessions on pedagogic, teaching and even managerial practices – were these available in the school or in a nearby education centre. The relevant time span for the formulation of these questions is probably the week before the interview from Monday through Friday.
- Third, questions on *what teachers do when students are absent for extended period of time* will provide an idea of how concerned teachers are about their students. An alternative to applying a teacher survey to all teachers in the school, which can be too time-demanding, is to randomly select a sample of teachers. Nonetheless, if student level data is collected, then collecting data on teacher characteristics and performance for those teachers teaching students in the evaluation sample is a must.

The main drawback of the above measures of teacher behavior is that they are often reported by the teachers themselves and over a set period of time. As it is unlikely that teachers keep precise records, there is a risk that the answers are not accurate besides being unreliable. One alternative is to ask some of the questions related to teacher behavior to the school principal, the students, or the parents; and to contrast answers from the different information sources. In some contexts, it might simply be unfeasible or extremely inappropriate to ask the teacher about certain aspects, such as absenteeism rates. A proxy

variable could be obtained by asking the principal or rather asking students or their parents how many days they missed class over the previous week or month because the teacher was absent (Jimenez and Sawada 1999). It might also be informative to ask students about teaching practices using questions such as whether the teacher repeats what is not understood, whether she encourages students to study, whether she encourages teamwork, and whether she comments on the homework and allows discussion in the class. Carrying out a pilot questionnaire will be very useful in order to give some insight on the most appropriate way to formulate these types of questions in different contexts.

School Access Outcomes

Under certain circumstances it might be of interest to determine whether SBM reforms are successful in expanding school access – for example, when decision-making power is transferred to the school in devastated and isolated communities where a more centralized provision of education is unfeasible, as has been the case in Central America (Di Gropello 2006). Increased school access can be measured using different variables such as total school enrollment or the number of days a student attends school over the number of days the school is open, which is a measure of individual (student) participation rate.

Intermediate Quality of Education Outcomes

SBM interventions can have a positive impact in improving student flows, namely, dropout, repetition, and failure. If measured at the school level, a dropout rate is an indicator of the success of a school in retaining those students who enroll. It is believed that this outcome measure is a useful indicator of school quality as perceived by parents (Murnane and others 2006). A related measure is the proportion of students that transit to the next grade or to the next school level. Another possibility is to look at the proportion of students that are over age for the grade in which they are enrolled in. "Overage" students are more likely to drop out – as they have a higher opportunity cost of being in school – and their presence stretches resources across a larger number of students.

Intermediate quality education outcomes can be measured at the school level across all grades – or preferably by grade – in the form of percentages or rates. Also, one could look at heterogeneous responses by sex and other student characteristics such as ethnicity or parental background. If detailed student information exists, however, it might be worth exploiting the individual variability in the data and using the individual probability of outcome measure y happening as the relevant impact indicator.

Student Achievement Outcomes

Improving learning outcomes has rarely been an explicit goal that has motivated the introduction of SBM programs. Although there are good reasons to believe it can have positive impacts in learning, establishing the direction of the causal relationship between SBM and student test scores remains an open empirical question. So far, robust evidence on the topic is scarce. This is partly due to the fact that successful SBM programs may take a few years to affect learning outcomes. Nonetheless, collecting math and language test score data using standardized evaluations on the evaluation sample of students for long enough periods of time might be a challenging and costly task. Moreover, representative and

comparable test score data is rarely readily available for the econometrician to use, which hinders the study of student achievement in retrospective evaluations. Many of the test score data available – usually collected by the relevant governmental education agency – is not usable for evaluation purposes for two reasons: first, because it is often collected on a non-representative sample of the sample under study; and second, because it is often not comparable over time due to changes in the structure of the examinations. However, test score data are available, and efforts should be made to obtain them and to request adjustments to make it usable for future evaluations of specific programs as has been done in the past with Mexico's national assessments, for example.

C. Data Sources

In retrospective evaluations, the existing data available will be a key factor in the determination of the evaluation method to apply. In prospective evaluations, two elements will be crucial in the success of the evaluation: the design of the survey instruments and the selection and size of the evaluation sample that should include both a treatment and a comparison group. In practice, particularly in a budget-constrained environment, prospective evaluations are also feasible using administrative data. Albeit not ideal – especially if the evaluation seeks to identify the mechanisms whereby the reform affects outcomes – this may be the only feasible option.

Data used in the evaluation of SBM interventions typically come from a variety of sources. Some studies use data purposely collected for the evaluation of the intervention, as is the case of the evaluations of the SBM programs implemented in El Salvador and Honduras (Jimenez and Sawada 1999; Sawada and Ragatz 2005; Di Gropello 2006). Ideally, purposive surveys should be collected before and after the intervention and on nationally representative samples. However, in some evaluation designs (experimental designs), it might be too costly to sample a nationally representative set of schools.

Purposive surveys are usually composed of different questionnaires, each of them applied to a relevant school agent. The school principal questionnaire should collect school-level questions about the school type, facilities (infrastructure, equipment), student enrollment and other student census data, teacher quality and quantity, and the school finances, operations, and management. The teacher questionnaire should contain teacher-specific information such as her educational background, years of experience, wages, teaching practices and methods, and meetings with other agents, as well as classroom-specific information. The questionnaire applied to members of the parents' association should include questions on the organization and practices of the association and its influence in the school administration and management. The student questionnaire should collect student level data on her individual characteristics such as age, gender, achievement test results, educational background, time use, habits and studying practices, and health status and other key family background data such as household demographic composition and living standards (asset ownership, consumption, etc.), and parental education and labor force participation. Data on household characteristics could be also collected in a separate questionnaire applied directly to the students' parents.

Many countries collect school and population census data routinely. School censuses can provide most of the information on school characteristics needed as covariates in the analysis as well as school-aggregated or grade-aggregated intermediate education outcome variables (failure, drop out, overage rates, etc.). Population and housing censuses can provide useful information related to the targeting rule used by the government to identify beneficiary schools and other community or regional time-invariant or time-varying characteristics worth controlling for. Sometimes even test score data might be available for a nationally representative sample of students. Administrative data on the implementation of the program can include relevant information on the targeting criteria, take up and participation rates, money disbursed or other benefits provided, the type of responsibilities transferred, and the timing. All these pieces of information can be very useful in the construction of the treatment variable. Administrative data on other educational or social programs can also be useful to control for other interventions simultaneously intervening in the school or the region that are likely to affect the supply and demand for schooling in the area. All of these data sources can be combined with additional purposive surveys thus lowering the data collection cost and effort substantially. Unique geographical, school, and student identifiers will then be essential to efficiently combine all different data sets. Hence, efforts should be made to request that individual administrative data systems are designed with the goal of linking them to other datasets in mind – which is not often the case.

Quantitative data can be complemented with qualitative interviews with school agents. These can be collected either before or after the intervention. If collected before, they will help form hypotheses and define the type of data that needs to be collected and the main dimensions of heterogeneity of impacts. They might also inform the intervention design through the ex-ante identification of the administrative problems that departments of educations and schools might experience in supporting the intervention. If carried out after the intervention, they might help assess the plausibility of the results and interpretation. Moreover, they can provide high quality information on the indirect costs of the intervention, the level of decision-making devolved to the school, processes, school management, and the school agents' feelings about having more influence in the decision-making process. We will return to this topic in section III.D when discussing qualitative evaluation methods.

D. Sample Sizes

Planning prospective evaluations should also take into account sample size considerations. This is usually a complex issue. Sometimes, however, sample sizes will be determined by budget constraints. If there are no limitations, then the study needs to have the adequate size relative to its goals. Calculating the correct sample size for a survey is an extension of calculating the sample size for each relevant outcome question. Two general formulae exist to compute sample sizes: one calculates sample sizes when estimating averages or means (continuous variables), and the other one calculates sample sizes for proportions (dichotomous and polychotomous variables and rates). In either case, applying the formula implies knowledge of the following elements: (i) the hypothesis test on the parameter of interest and the underlying probability model for the data; (ii) the significance level of the test (90 or 95 percent significance level are usual values); (iii) the desired effect

size (a *x* percent decrease in the failure rate, for example); (iv) historical values or estimates of parameters (usually the variance) of the outcome variable of interest; (v) the tolerance for error or power of the test (0.80 to 0.90 are common power values).

Outcomes in SBM evaluations can be continuous (test scores, number of meetings, hours devoted to teaching), binary (whether or not the school has autonomy over a certain process, whether or not a student repeats a grade), polychotomous (how much autonomy a school agent has on a process over a scale) or proportions (percentage of students failing a grade or dropping out). If scales refer to qualifiable attributes (satisfaction, perceptions) they should be treated as a proportion. Working with dummies and proportions is easier, as the variance is entirely determined mathematically from the mean. When working with continuous variables, historical data can sometimes be used to estimate the variance of the outcome variables. Alternatively, a pilot study can be very useful in this respect.

Logistical, financial, and ethical considerations make sample size issues specially pressing in the case of control randomized experiments, which will be discussed at length in section III.C. It seems to be a rule of thumb amongst educational researchers that 40 to 50 schools (clustered unit of treatment) with 40 to 60 students (unit on which impact is measured) are needed for a cluster randomized trial contrasting two equally-sized treatment groups at conventional power and significance levels in order to detect intercept differences in student achievement test scores between 0.10 and 0.25 standard deviations (Bloom and others 1999; Raundenbush and others 2004). Notably, the number of clusters or sampling units (schools) needed will be larger if the analysis is performed at any other level with fewer observations per cluster (school, group or teacher level, for example), which will in turn increase the cost of the study considerably. Nonetheless, recent developments demonstrate that introducing cluster-level covariates that are highly correlated with the outcome variable can reduce sample size considerably with no power detriment (see Gargani and Cook 2007). Commonly used statistical packages such as STATA include statistical software that performs power and sample size calculations for non-cluster and cluster sample studies.[9] Free downloadable software is also available on line.[10]

E. Timing of Outcomes and Length of Evaluation

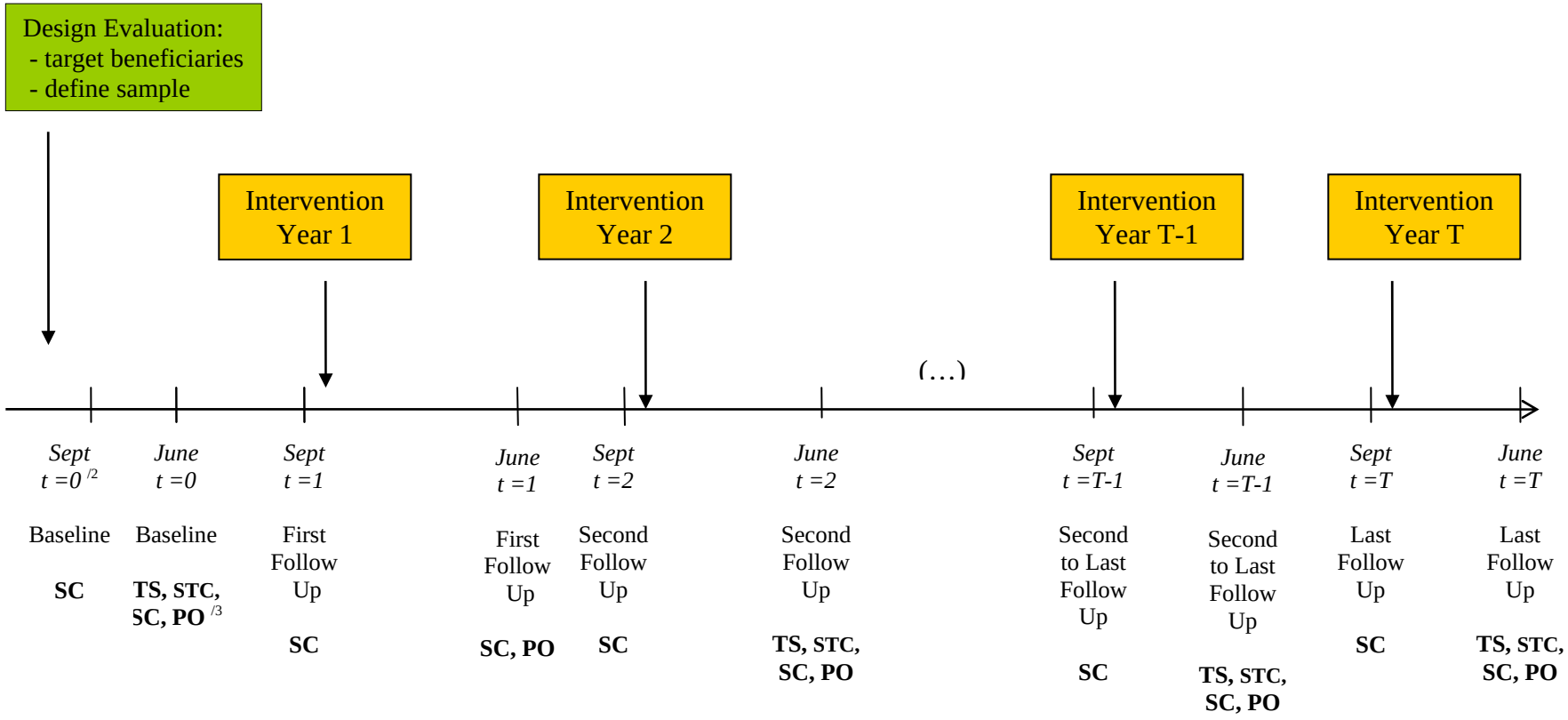
A reasonable time frame for impacts to become evident will not only depend on the outcome measures but also on the nature of the intervention, what it demands from schools, and how developed managerial skills across school members were before the introduction of the program. Test scores, for instance, will likely take longer to react to increased autonomy in the school than parental involvement. Similarly, if school managers have never engaged in strategic planning activities, the intervention might take a longer time to have effects. Impacts could even be negative during the first years (adjustment period), given high coordination costs between school agents or between the state and the local school environment.

[9] See "sampsi" and "sampclus" commands.
[10] See http://www.ssicentral.com/otherproducts/othersoftware.html and http://www.cs.uiowa.edu/~rlenth/Power/ for software, references, and useful links on power calculations.

One should also think practically about what types of outcomes are likely to be observed at each point in time and when to collect them. Process outcomes should probably be collected for the first time six months or a year after the start of the intervention and every half year or every year from there on. If outcomes on processes are to be collected yearly, it seems natural to collect them at the end of the school year. This strategy will not only give the maximum possible retrospective time frame but will also minimize the recall bias with respect to the alternative procedure of collecting the data at the beginning of the next school year. Intermediate quality of education outcomes such as intra- and inter-year drop out and repetition rates should be computed yearly. Their computation requires knowing how many students were enrolled at the beginning and at the end of the school year. Most school censuses do collect data on the number of students and whether they are first time enrolled or repeating the grade, both at the beginning and at the end of the school year for this purpose. Finally, because test scores may take longer to react to SBM reforms, it might be advisable to allow two complete school years or more before measuring impacts on achievement (Borman and others 2003). It is standard practice to collect achievement data at the end of the school year with the objective of capturing what the student has learnt over the course of the year.

In an attempt to summarize the many points addressed in this section, Figure 2 presents a diagram of the “ideal” timeline and data collection scheme for a hypothetical project, as an example. Note that in this hypothetical scenario the school year is assumed to run from September through June.

Figure 2: <u>Ideal Timeline and Data Collection Scheme of a Hypothetical Intervention</u> /1



/1 Source: created by authors.
/2 The school year is assumed to run from September through June.
/3 **TS**: Test Score Data; **STC**: Student Context Data and/or Household Questionnaire; **SC**: School Census Data; **PO**: Process Outcome Data

III. Evaluation Designs: Targeting of Beneficiaries and Evaluation Methods

The fundamental evaluation question is the measurement of the impact of a certain intervention or program on a set of well-defined variables on the beneficiary population relative to what they would have experienced had they not benefited from the intervention or program. The problem is one of *missing data*: individuals benefiting from the program cannot be simultaneously observed in the alternative state of no treatment. Thus, the central issue evaluation methods address is how to construct a *counterfactual* with no intervention against which to measure the change with intervention. A valid counterfactual should be as similar to the target group as possible except for the fact that its members do not benefit from the program. Only then it will be possible to establish the causality link between the intervention and the observed changes on outcomes.

Solutions to the evaluation problem differ in the method and data used to construct the mean counterfactual term, which is in many cases largely determined by the way beneficiaries are selected. Broadly, the evaluation literature classifies evaluation designs as: non-experimental, quasi-experimental, and experimental. They vary in feasibility, cost, and the degree of clarity and validity of results. The design also determines the set of estimation methods available to obtain an unbiased estimate of the program impact. The methodology employed will further depend on the type of information available, the underlying model, and the parameter of interest. For example, datasets with longitudinal or repeated cross-section information will support less restrictive estimators due to the relative richness of information. The most common estimate of impact is the *Average Treatment Effect on the Treated* (ATT) which is derived by comparing the mean levels of well-being between those in the treatment group that actually received benefits and those in the comparison or control group.[11]

Next, we describe the alternative ways governments can identify and select the beneficiaries of a SBM intervention, classified by the evaluation design employed. We also present the estimation methods available under each design, focusing the discussion on the strengths and weaknesses of each method. When possible, we illustrate each methodology presented with examples from existing SBM impact evaluations.

A. Non-Experimental Designs

Universal Coverage But Non-Universal Participation: Self-Selection Bias

When the SBM intervention has universal coverage but not all schools choose or volunteer to participate, participant schools can be compared to non-participant schools. However, the reasons why a school chooses to participate or not participate might be very diverse and respond to systematically different characteristics between participant and

[11] The term "comparison group" is associated with quasi-experimental designs, while the term "control group" is used when the evaluation employs an experimental design.

non-participant schools. It seems reasonable to think that wealthier schools, better-structured schools, schools more open to change principals, or schools with less discord amongst school agents are more likely to take up the intervention. This is more so, the more demanding the program is in terms of time and effort managing resources, implementing changes, or collecting additional funds (Skoufias and Shapiro 2006). The characteristics that induce participation are also likely to be positively correlated with outcomes: more active and prone-to-change schools and schools that enjoy a better climate are also more likely to have better educational outcomes. Hence, a comparison of participant and non-participant schools will surely suffer from positive *self-selection bias* and overestimate the true program effect.

In the analysis of student-level outcomes – student test scores, for example – *self-selection bias* also arises when students (or their parents) can alter the exogeneity of the treatment variable through school choice: they can choose whether to participate or not in the program by choosing whether to attend or not a SBM beneficiary school.[12] This phenomenon is also known as *sorting bias*. Even in remote rural areas where parents have little (if any) choice over which school to send their kids, they might still send their children to live with relatives (allowing them to attend a non-local school) or to a boarding school. If the selection of which school to go to is influenced by unobserved characteristics (parental or student preferences for education) that are also correlated with the outcome of interest (student progress), then selection bias is in place. Students may decide to exit a treated school if they interpret treatment as a signal of the school's malfunctioning. Then, students with a lower preference for education and lower learning would remain in treatment schools, and the negative correlation between choice and ability will bias downwards the true program effect. Contrarily, students with a higher desire to learn may be encouraged to enter a treated school if they think they can benefit from the additional resources poured into the school. In this case, the positive correlation between choice and ability will likely overestimate the true program impact.

Sorting bias can also affect treatment estimates on school averaged test scores or other education quality outcomes. If autonomous schools do a better job at retaining students who would have otherwise dropped out, then the average school achievement remains lower. This is to say, the achievement effect is washed out by an attainment effect and underestimated. The converse is also possible if autonomous schools attract better performing students. Controlling for this form of bias can be done using data on school rolls on enrollment, passing rates, and desertion rates.

Universal Coverage within a Specific Group According to Certain Criteria: Endogenous Program Placement

Non-experimental designs are also used when governments target interventions to areas with particular needs and characteristics which are thus systematically different to those areas where the program is not allocated. For instance, the state government could assign benefits to more disadvantaged schools first given budget constraints. This would

[12] Note that this decision is conditional on attending school. We will ignore here any considerations on the previous decision of whether to go to school or work or both.

produce a negative correlation between the school unobserved components in the error term and the treatment variable. Hence, estimates of the program impact would be downward biased. On the other hand, governments may be just as likely to place treatment in areas that already have good education outcomes in order to increase the chances of positive outcomes or because they might derive political support from elite groups. Alternatively, better performing schools that have stronger and more concerned parent associations might push the local authority harder to allocate benefits in their school. In either situation, these schools are likely to continue to do better than worse performing and less influential schools even without the program. Hence, program impact estimates will likely be upward biased. Biases coming from this source are known as *endogenous program placement bias*.

A first possibility is to exclude from the analysis those areas where the local authority might have had more discretion in the allocation of treatment. Non-experimental designs, however, rely in the use of econometric techniques to statistically control for differences between participant and non-participant schools or students (self-selection) and targeted and non-targeted schools (endogenous program placement). The simplest strategy is to use multivariate regression analysis and control for all observable characteristics that are thought to determine the school decision to participate in the program or the student (or her parents) decision to attend a treatment school. Nonetheless, if participation is also determined by unobservable characteristics such as the drive of the school principal, his ability to raise funds, and parental or governmental preferences, then OLS estimates will suffer from *omitted variable bias*.

An alternative is to locate one or more instrumental variables (IV) that matter for the treatment status or more generally for participation – *relevance of the instrument* – but that are not correlated with the outcomes of interest given treatment – *exclusion restriction*. Thus, the instruments control for the endogeneity in the choice variable (enter a school, take up a program or allocate a reform) that arises from selection on unobservables. Nonetheless, valid and plausible instrumental variables are usually very difficult to find (Heckman 1979). Impact evaluations of education interventions (SBM and other) often exploit the geographic variation in program availability or program implementation as instruments, especially when endogenous program placement is the main source of bias. In student achievement regressions, school choice is usually instrumented with variables related to the cost of schooling: price of schooling and distance to the school. However, these variables might violate the exclusion restriction if distance is correlated with absences or tardiness – likely to affect learning – or if the price of schooling also depends on the demand for schooling. Another possibility when past (pre-program) data are available is to use lagged (pre-program) values of participation determinants as instruments. However, because past determinants are strongly correlated with current determinants, they are arguably weak instruments.[13]

[13] Note that in a heterogeneous treatment framework, the IV methodology is unfeasible as the instrument is required to be correlated with the participation decision and uncorrelated with the individual specific effect that likely determines participation (Blundell and Costa Dias, 2000). Chapter 7 in Davidson and MacKinnon (2003) offers an excellent overview of the IV methodology.

When the endogenous choice variable takes values between 0 and 1 it can be estimated using a probit model and standard Heckman Selection methods can be applied (Heckman 1979). Theoretically, the selection model can be identified from functional form assumptions on the distribution of the errors in the participation and outcome equations; this is to say, without imposing any restriction on the regressors. Unfortunately, these distributional assumptions are hardly ever defensible, so the empirical analysis will have to rely on the existence of at least one regressor in the participation equation excluded from the outcome equation (instrument) to correct for selection biases.

This is the approach taken by Jimenez and Sawada (1999) in the evaluation of the EDUCO program in El Salvador. EDUCO started as an initiative from the Ministry of Education to expand pre-primary and primary rural education following the civil war. It was based on a community initiative that organized and set up schools in rural areas where education could not be extended during the war. In these communities, associations of households organized, administered, and financially supported the school. Since 1991, EDUCO autonomous schools are responsible for allocating budgets; staffing, equipping, and maintaining the schools; and monitoring teacher performance. They are, however, required to follow a centrally mandated curriculum and maintain a minimum student enrollment level. In non-EDUCO schools, the parent association has no administrative authority over school personnel or budgets.

The authors estimate school production functions at the student level and model selection into an EDUCO school using a Heckman two-stage procedure. They exploit the government prioritizing formula – a *non-linear* function of community and other socioeconomic and geographic variables – as an instrument. More precisely, the authors use district dummy variables as the excluded regressors in the main equation (test scores or days missed) based on two arguments. First, the weights that determine the influence of these variables in the targeting formula are a priori uncorrelated with any individual decision as were exogenously chosen by the government. Second, they are likely to affect the decision to go to an EDUCO school given they grant access. As discussed earlier, the weakness of this approach is related to the fact that variables that affect school access are also likely to affect school absences, tardiness, and ultimately learning.

Similarly, Jimenez and Sawada (2003) study the impact of EDUCO on school drop outs. In this paper, the authors take a slightly different approach and consider the decisions of going to an EDUCO school and staying in the school as simultaneous. They estimate a bivariate probit model and use the proportion of EDUCO schools and traditional schools relative to all primary schools in a municipality as instruments, since these proportions are pre-determined by the municipal authority. Because EDUCO's main purpose was to supply education to underserved areas, the proportion of EDUCO schools in a municipality is likely to be correlated with the density of schools in the municipality – a measure of access likely to affect the decision to stay in school. Moreover, the government allocation rule is not necessarily exogenous or random, and the concern for endogenous program placement bias remains.

Di Gropello and Marshall (2005) also apply Heckman selection correction techniques (amongst others) in their evaluation of PROHECO on student achievement. PROHECO started in Honduras in 1999 with the objective to improve school access and encourage community participation in rural isolated areas that had been affected by Hurricane Mitch. PROHECO schools have a council in charge of selecting and paying teachers, monitoring teacher and student attendance and performance, managing school funds and materials, and building and maintaining the school. Di Gropello and Marshall (2005) use the presence of potable water and the sum of services (post office, water, electricity) in the community as instruments. Because services that affect access might also affect learning, these instruments are not particularly convincing. For example, electricity allows students to study at night and access to potable water is likely to reduce the number of days a student misses school because he is sick.

B. Quasi-Experimental Designs

Universal Coverage: Reflexive Comparisons

In the evaluation of interventions with *nationwide coverage* – where there is no room for a comparison group – it is still possible to compare participating schools to themselves before and after receiving the intervention provided there exists longitudinal data. However, such a strategy – known as reflexive or before and after comparison – presents very serious problems. Indeed, its major drawback is that estimates of the effect of the program also include aggregate effects or trends in the outcome variable. For example, reductions in the aggregate failure rate given a SBM reform can also reflect changes in other aspects of the educational strategy of the country, such as a curricular reform, or a lower student teacher ratio given a decreasing demographic trend in the country. While it is possible to include some of these factors as statistical controls in the regression (the student teacher ratio), others are almost impossible to quantify and control for (the curricular reform). Thus, before and after estimates of impact will inevitably suffer from *omitted variable* and *measurement biases* and *should not* be considered an option when doing impact evaluation unless evaluation is a must *and* coverage is universal. Even in such circumstances, the real need for an impact evaluation should be reconsidered as results could suffer from serious biases and not necessarily reflect the causal effects of the program.

Wylie (1996) uses data from 1989 to 1993 to obtain a *before and after* impact estimate of the New Zealand SBM reform whereby all schools in the country became fully autonomous beginning in 1990. Similarly, Nir (2002) uses three years of data from 28 elementary schools in the municipality of Jerusalem, which was the first one to adopt the Israeli SBM reform. This intervention established a governing body in schools that presented a well-defined work plan and exerted extensive monitoring. The report was commissioned by the Ministry of Education to explore the expansion of the reform at the national level. Given the implementation of the reform, however, results are likely to be non-representative nationally and suffer from selection biases.

Partial Coverage: Selection of a Sub-Population of Non-Beneficiaries as a Comparison Group

When an intervention has *partial coverage,* it is possible to construct the comparison group using the sub-population of non-beneficiaries that is most similar to the treatment group. This can be either done *prospectively* – the treatment and comparison groups are selected before the intervention is in place – or *retrospectively* – the comparison group is identified after the intervention. In either case, however, the comparison between beneficiary and non-beneficiary schools will not be exempt of biases, as there are both observable and unobservable reasons why a school is deemed eligible or ineligible for benefits.

Quasi-experimental designs and methods deal with these biases in different ways. The common factor to all of them is that only non-beneficiary schools with *similar* characteristics to beneficiary schools contribute to the calculation of the expected counterfactual. They differ on whether treatment and comparison groups are selected on the basis of purely observable characteristics (matching and sharp regression discontinuity methods) or also on unobservables (difference-in-difference and fuzzy regression discontinuity methods). These methodologies require – unlike reflexive comparisons – the existence of data on both the treatment and the comparison groups. Difference-in-difference estimation additionally requires the existence of data collected at least in two different periods: before and after the intervention.[14] What follows is a characterization of alternative ways governments can target SBM beneficiary schools and identify a valid comparison group using quasi-experimental designs, along with the estimation techniques applicable to each design.

Exploit Non-Beneficiary Characteristics

Imagine a situation in which a large group of schools has been excluded from a SBM program for reasons unrelated to the reform: for example, because they belong to a different geographical region. Assume also that there exist abundant data on the determinants of (participation in) treatment and on outcomes for the samples of participant and non-participant schools. It is then possible to select for each treated school the (set of) non-treated school(s) that has the same realization (or is most "similar") in terms of some essential observable characteristics. These schools will constitute the comparison group. This technique is called general or exact matching and heavily relies on program participation or allocation being orthogonal to outcomes once one has controlled for these observable variables (*conditional independence*). A widely used matching method is propensity score matching (PSM). The propensity score is the predicted probability of (participation in) treatment given observed characteristics (Dehejia and Wahba 2002). In the current context, the propensity score is the probability that a school is offered a SBM intervention and/or decides to take up the offer to participate. It will generally be a function of observable school characteristics such as

[14] See Blundell and Costa Dias (2000) for an overview of the difference-in-difference and matching methodologies; and Hahn and others (2001) for details on the regression discontinuity design.

school type and size, locality characteristics, etc.[15] Instead of finding the best comparison school(s) for every single characteristic, one defines (the neighborhood of) similar school(s) on the one-dimensional probability to participate conditional on observables: the propensity score. Each treated school is then paired with its selected set of non-participant schools using weights: equal weights to all, unity weight to the nearest observation and zero to others, kernel weights to account for the relative proximity of non-treated schools, etc. Note that this is a nonparametric approach as it does not need to assume any specific relation (linear or other) among treatment, covariates, and outcomes.

One example can be found in Sawada and Ragatz (2005), who inspect how EDUCO affects administrative processes and teacher behavior and how these affect education quality. Building on Jimenez and Sawada (1999), the authors use PSM to address self-selection concerns. They first estimate a probit regression of the probability of being an EDUCO school on school, school agents, and community characteristics. Then, they apply weighted nearest-neighbor matching to match each EDUCO school to the comparison school with the closest propensity score. Some of the results found using OLS are no longer significant when the authors apply PSM, which suggests that selection bias is a concern in the data.

PSM techniques can also be applied to address self-selection biases at the student level coming from the decision of whether to attend an autonomous school. An example is offered in Parker (2005) for the evaluation of the Nicaragua's SBM reform. Begun in 1993, the reform consisted in the transferring of key management tasks (hiring and firing the school principal and maintaining school facilities and academic quality) to school councils. Parker (2005) starts by showing that students who attend autonomous schools are, on average, significantly younger and wealthier than students who attended centralized schools. In consequence, the author uses stratification and nearest-neighbor propensity score matching to compare Spanish and math test scores from students in each type of school. While the paper is not particularly explicit, details on how the PSM procedure was implemented can be obtained from the author upon request.

There are two main challenges with the use of matching methods. The first is related to the heavy requirements they impose on the data. Exhaustive information on the characteristics of participant and non-participant schools is needed to model the participation decision. But the more detailed this information is, the harder it is to find a similar comparison group, as treatment and comparison schools will have to be matched on a larger number of similar characteristics. That is, there is a trade-off between the quantity of information to use and the size of the comparison group. Sawada and Ragatz (2005) report running into this problem when trying to find PSM estimates of the effect of EDUCO on teacher effort. A second challenge is that matching methods hinge on identifying all relevant differences between treatments and comparisons purely on observables. However, if treatment/participation is assigned/decided on the basis of some variable that is not observed by the researcher – the school desire for autonomy, for

[15] A common approach is to construct the propensity score using pre-intervention characteristics. However, this requires the existence of exhaustive pre-intervention (baseline) data for both the group of participant and non-participant schools, which is often not the case.

example – this technique will not correct for selection biases stemming from unobservables, which will – following with the example – overestimate the program impact.

In face of these difficulties and given there are pre- and post-intervention data available (though not necessarily for the same schools), it is possible to use the difference in outcomes between before and after in the comparison group as a counterfactual for the difference in outcomes between before and after in the treatment group. This method is known as difference-in-differences or double differences (DD). Its main advantage over matching methods is that all observed *and* unobserved time *invariant* individual characteristics that determine (participation into) treatment no longer bias impact estimates since they are differenced out in the estimation equation. However, a good argument that the outcome would not have had differential trends in treated schools had they not received treatment has to be made. This is neither (i) likely to occur when there are macro-economic effects in the region that affect treatment and comparison schools differently (other demand or supply education interventions, for example) nor (ii) testable empirically. Instead, one can test whether treatment and comparison schools had differential trends before the introduction of the program and assume that this difference would have been kept constant in the post-intervention period *were* the intervention not in place. This requires access to long time series of pre-intervention data from both types of schools to compare pre-trends over long enough periods of time. A second weakness of this methodology is that it does not control for any time-*varying* school characteristics that affect (participation into) treatment, such as changes amongst the school committee members or changes in the committee's preferences and strategies. In applying DD methods, one must control for as many time varying observable characteristics as are available and include separate time trends for treatment and comparison schools in the estimation, in order to minimize the potential for biases.[16]

Murnane and others (2006) use DD methods to analyze the impacts of the PEC intervention in Mexico on student academic progress. PEC started in 2001 to promote school planning and increase community participation. It is administered by the state Secretariats of Education and guided by national regulations and oversight. To qualify for the program the principal, teachers, and parents in the school prepare a school improvement plan that includes a diagnosis of the school needs, objectives for improvement, and an annual working plan. Schools are chosen to participate in PEC on the basis of their improvement plan, and wining schools receive five years of financial support to bring it about. They also receive financial incentives to engage in their own fund-raising activities. The design of the program is likely to motivate self-selection of schools into it: schools with less discord between staff and more willing to implement changes are more likely to write better structured improvement plans.

Murnane and others (2006) compare PEC schools – defined as schools that joined the program on its second year of operation (2002) – to non-PEC schools after verifying

[16] Bertrand and others (2004) note a third limitation of DD methods coming from strong serial correlation of the error term, which results in an underestimation of the standard deviation of the parameter of interest. Note that serial correlation issues are more stringent when long time series are available – a rare event in the evaluation of SBM interventions.

the equality in their pre-intervention trends over four years of data. Nonetheless, the hypothesis of equal trends is rejected when the authors define as treatment those schools that enrolled in the program on its first year of operation. These schools were improving outcomes more rapidly in the pre-PEC years than comparison schools, which proves that any PEC evaluation is likely to suffer from serious self-selection, at least on its initial stages (best schools apply first). Murnane and others (2006) opted to redefine the sample of treatment schools such that the existing comparison schools were a valid counterfactual. Alternatively, they could have redefined the comparison group of schools.

Time-invariant selection bias – equal *linear* pre-intervention trends – might sometimes be too strong an assumption. There might be situations where the SBM intervention is allocated to the worst performing schools with a higher potential to improve outcomes *and* at a faster pace. DD estimates will then be biased given that post-intervention changes in the outcome variable are a function of the same initial conditions that influenced (participation into) treatment. In such situation, controlling for initial heterogeneity is crucial to obtain credible DD estimates. Using PSM to select the comparison group is an obvious corrective that will produce more accurate estimates under less restrictive assumptions. This combination of methods boils down to applying matching techniques to changes rather than to levels. The approach will reduce the bias in both the DD estimates – by better accounting for initial heterogeneity – and the PSM estimates. Because the main matching hypothesis is now stated in terms of the before-after evolution, there is room for unobserved determinants of participation as long as they do not vary across observations and over time. In other words, selection can be on individual- and time-specific components of the error term.

Skoufias and Shapiro (2006) adopt this strategy to evaluate PEC on school averaged repetition, failure, and drop out rates. Their approach builds on the standard DD methodology in Murnane and others (2006) in that it estimates the differences between PEC and non-PEC schools in a more flexible way, using semi-parametric methods. Moreover, it matches each PEC school to a set of comparison schools with similar observed characteristics and located in similar communities, rather than to all comparison schools. However, because they only have data on two periods, the authors are forced to assume that pre-intervention trends between treatment and comparison schools were the same. Skoufias and Shapiro (2006) also report simple DD and PSM estimates but propose DD PSM as their preferred method for the reasons mentioned above.

The underlying identifying assumption of linear equality of pre-intervention trends can be relaxed in two additional ways. A first approach comes from the labor literature (Bell and others 1999) and consists in allowing macro effects across treatment and comparison schools and including in the estimation another time *interval* over which a similar macro trend has occurred (as opposed to including a single pre-intervention time period). A second possibility that accommodates differences in trends between treatments and comparisons is the non-linear DD model – the non-parametric changes-in-changes estimated procedure proposed by Athey and Imbens (2006).

Exploit a Discontinuity in the Targeting Rule

Program targeting rules often create discontinuities that can be used to identify the effects of the program by comparing schools that are just above the discontinuity and schools that are just below. Imagine, for example, that the SBM intervention is awarded to schools on the basis of the quality of a school improvement plan presented to the state government. Suppose that the quality of the proposal is graded according to some specific criteria and that only schools with a score above a certain threshold receive the program. If this quantifiable score is well-defined, known by the researcher, continuous, and strictly enforced, it is possible to apply regression discontinuity (RD) methods. RD techniques consist in comparing schools just above the threshold (SBM beneficiaries) to schools just below the threshold (non-beneficiaries). Its implementation requires that schools are ordered along the score and are large in number around the discontinuity or cutoff score. While it is always possible to open a larger window around the threshold to increase sample sizes, this will come in detriment of the comparability between treatment and comparison schools and hence increase the scope for biases.

If panel data are available, one can combine RD with before and after comparisons and compare schools above and below the threshold before and after the intervention. In this case, the "before" data can be used to test how well balanced (how similar) schools on either side of the cutoff were prior to the intervention. Ideally, there should be no difference (no "jump") in the outcome values of schools at the discontinuity before the intervention. Provided there are data for more than two points in time, RD and DD methods can be combined to additionally control for all observed and unobserved time invariant factors correlated both with a school's treatment status and its outcome levels. The advantage is that one can strengthen the case for the comparison group by testing the equality of treatment and comparison school pre-trends.

RD designs can be a convenient method to solve *reversion to the mean bias*. This is a common source of bias in the evaluation of educational interventions targeted using rankings of schools that contain past outcome measures. Imagine, for instance, that the targeting criteria used to identify eligible schools include grade-averaged student test scores. Suppose that there is noise in the measurement of test scores such that the low performance of the school is correlated to a bad shock. A bad cohort of students would be an example of a negative shock affecting test scores on a particular year. Then the ranking of schools can be misleading. Unless shocks are correlated over time, we would expect mean grade test scores in the bad performing school to increase over time even in the absence of any intervention. Thus, standard DD estimates of impact will be upward biased as they will reflect a combination of a true program effect and a spurious mean reversion. Chay and others (2005) show that – whenever they are feasible – RD designs control for all omitted factors correlated with being selected for treatment, including the intensity of the mean reversion. An alternative is to include pre-intervention values of the mean and variance of the variables that cause the bias (test scores in the example) in the regression. Another advantage of RD methods is that they allow for heterogeneity in treatment under some additional assumptions (see Hahn and others 2001).

If the discontinuity is not enforced very strictly, the possibilities to apply a RD design are reduced. However, the chances to find a large enough number of treatment and comparison schools that overlap over a common support of targeting scores increase. This will make it possible to apply matching techniques. A second disadvantage of RD methods is that because they estimate effects at the discontinuity, treatment effects are rather local (only for a selected sample of participants) and not always generalizable to the full population. Thirdly, RD methods require larger sample sizes to identify effects – as suggested earlier. A fourth concern arises when there is a mismatch between the targeting unit (school) and the unit of analysis (students). Schools just above and just below the threshold might have very similar characteristics, at least in terms of the factors that contribute to the targeting rule. However, students in comparable schools do not necessarily need to be comparable, and the RD will not adjust for differences in students within comparable schools. Although it seems reasonable to expect these students to have common unobserved characteristics – given they chose similar schools – it will be crucial to include as many student covariates as possible in the regression. Lastly, the RD method can be very sensitive to misspecifications of the functional form chosen to model the relationship between the assignment and the outcome variables. One should test for the significance of higher order terms and interactions – and even apply semi- or non-parametric estimation techniques – to prove the robustness of results to misspecification.

To our knowledge, there is no SBM evaluation that applies RD methods. Nevertheless, the PEC program in Mexico uses a RD design to target schools: participant schools are chosen on the basis of the quality of a school working plan. The reason why this discontinuity has not been exploited for identification might lay in the fact that the targeting rule was not strictly enforced and did not generate a *sharp discontinuity*. Indeed, Murnane and others (2006) report that "(…) after being reviewed by a technical committee (…), scores are submitted to a "social involvement committee" (…). This committee selects the schools to be enrolled in PEC, based on the comments by the technical committee *and other criteria like poverty levels*." This additional criterion to select schools makes the *discontinuity fuzzy*. In such situations where the cutoff is not perfectly known, it is still possible to obtain consistent estimates using a two-step procedure, whereby the propensity score of receiving treatment as a function of the targeting rule is estimated in the first step (see Van der Klaauw 2002 for an application).

Exploit the Program Phase-in Over Time, Space, or Both

"Random" differences in the timing of program implementation in different schools or in different geographical areas (school districts, localities, states, etc.) can also facilitate forming comparison groups. Examples of exogenous variations of this sort are administrative delays in program implementation or the application of a time-varying geographic targeting rule uncorrelated with outcomes. In these situations, the implementation of the intervention automatically generates a valid counterfactual net of potential self-selection biases as both participant and not-yet participant schools are potential beneficiaries and are thus likely to have similar observed and unobserved characteristics. It is then possible to compare schools that are already being treated with schools that will be treated in the future using matching methods or simple differences. As usual, if there are longitudinal data available, DD or a combination of methods apply.

Note that it is crucial for identification that the variation is exogenous. If contrarily, the allocation of treatment responds to certain political criteria, then comparisons of this sort will result in estimates that suffer from endogenous program placement bias.

This is the approach taken by Gertler and others (2006) in their study of the impacts of the SBM component of the Mexican Compensatory Education Program on school-averaged grade failure, grade repetition, and intra-year drop out rates. This component, the AGES (*Apoyo a la Gestión Escolar*, Support to School Management), began in 1996 and provides cash grants to parent associations who can then spend the AGEs money on the educational purpose of their choosing. Spending is in many cases limited to small civil works and infrastructure improvements. The authors compare schools that received the AGEs at earlier dates with schools that received or will receive the intervention at later dates using DD methods. All schools had values of the targeting index that the Ministry of Education used to determine eligibility over a common support, which backs its comparability. Gertler and co-authors also check for balance in pre-intervention trends to dismiss endogenous program placement bias. The authors also test and reject the existence of sorting of students by looking at changes in enrollment and control for other educational interventions that are simultaneously operating in the schools and that could possibly confound effects.

Paes de Barros and Mendonca (1998) use a similar strategy to evaluate a variety of SBM reforms that several Brazilian states progressively undertook between the 1980s and 1990s. The reforms had three main pillars: give financial autonomy to the school; establish school councils with participatory decision-making power; and either democratic election of school principals by school officials, parents, and teachers or competitive appointment via examinations. The authors use data from 1981 to 1993 and estimate a state fixed effects model to compare intermediate educational outcomes between states and points in time where the reforms had been instituted and states and points in time where they had not. As noted, the unit of analysis is the state which is likely to cloud important within-state variation. Moreover, the reasons why different states adopt different reforms at different times are unclear, which raises the standard concern about the allocation of treatment being endogenous.

Encouragement Designs

Randomized encouragement designs consist in randomly marketing or advertising the intervention to some schools and not to some others. In other words, in randomized encouragement designs, only schools in the treatment group – which have been selected randomly – are encouraged to participate or to apply to the SBM intervention although they are not required to participate. Schools in the control group are not marketed the SBM intervention but are still able to participate or apply for it if they choose to. This design is thus attractive when – because of ethical reasons or political pressure – it is infeasible to deny an intervention to schools that would like to participate, but full coverage is unaffordable given budget or capacity constraints.

One can use IV methods to estimate treatment effects under this setting. The randomly assigned advertisement campaign to apply for the intervention – the assignment

to treatment – appears as a natural and robust instrument for actual participation – treatment itself. If the campaign has been successful, the assignment to treatment will be highly correlated with program take up. Moreover, because the assignment to treatment is random, it is by definition orthogonal to any observables or unobservables correlated with outcomes. In the estimation, it will be crucial to use the assignment to treatment as opposed to treatment itself to differentiate treatment and comparison groups. Note also that the larger the number of participating schools in the control group and/or the number of non-participating schools in the treatment group, the larger the sample sizes needed; furthermore, the more likely is self-selection bias to affect estimates. Hence, a well designed and accurately implemented advertising campaign will be key elements for the validity of this approach.

To date, there have been no published results of a SBM intervention that has been rolled out using encouragement design, though results from on-going experiments in Nepal and possibly Mexico may soon prove useful. Duflo and Saez (2004) and Hirano and others (2000) are useful references for applications of this approach in other areas.

C. Experimental Designs

The most effective way of minimizing the potential for biases and obtaining credible impact estimates is the use of cluster based randomized designs. In the particular evaluation of SBM programs the cluster unit at which the randomization is performed is the school. Randomized designs involve two-stages: first, the identification of a group of potential beneficiary schools (or willing participant schools) with similar characteristics; and second, the random allocation of treatment to a subset of schools – the *treatment group* – and not to others – the *control group*. By definition, being a SBM beneficiary school is exogenous, or in other words, uncorrelated with any school observed or unobserved characteristics. This makes treatment and control schools virtually identical and causal inference feasible. Hence, a well-executed randomized experiment will greatly simplify the statistical analysis: it will suffice to compute the mean difference in outcomes for the treatment and control groups to obtain consistent estimates of impact. For this matter, when using experimental data, one should always assess the balance of the treatment and control groups by, for example, testing the equality of the means and/or the distribution of student, school, and locality characteristics between treatment and control schools at baseline. Similarly, one should compare pre-intervention means or distributions of the outcome variables (exogeneity test).

The key advantage of randomized designs is that they overcome many of the problems encountered when using other evaluation practices without the need to have to resort to difficult to test and hard to satisfy behavioral assumptions. Nonetheless, they are subject to several potential limitations. First, they are very costly to run. They involve establishing a rigorous evaluation design ex-ante, which might entail long and tedious negotiations with the different parts involved in the design, implementation, and evaluation of the intervention. Moreover, if the treatment variable has different dimensions or if extrapolation is an important objective, then experimental designs will require large samples that should preferably be followed over several periods.

Second, randomized designs might not always be feasible: policymakers can be uncomfortable with the idea of denying the monetary and training benefits from a SBM intervention to schools that deserve it on the basis of their characteristics, their past performance, or their past effort in succeeding. Some program designs, however, might favor the introduction of a randomized component better than others (*program induced randomization*). Imagine a SBM intervention that assigns benefits to applicant schools via competition on the basis of the quality of their proposals. Imagine also that there are more quality applicant schools than available resources to allocate. Then, it would seem fair to allocate benefits to particular winning schools and not to others using a lottery (see Angrist and others (2002) on the evaluation of a school voucher program in Colombia that benefited from a lottery design to facilitate the excess demand for school places). It could also be the case that the relevant authority temporarily lacks the capacity or resources to decentralize all applicant schools or the entire population of schools in a certain area. A natural solution would be to randomize the order in which the intervention expands to all potential beneficiaries. This design – known as *randomized phase-in* – would not only allow all schools to eventually benefit from the reform but would also provide the researcher with a clean source of identification. It is, however, important to allow enough time for the intervention to have effects amongst the initially treated before allocating benefits to the control group.

Third, impact estimates from randomized designs might not be generalizable to the population at large. The analysis of experimental designs is of a partial equilibrium nature and its results are unlikely to hold in a general equilibrium framework (Zellner and Rossi 1986). In other words, a broadly applied SBM reform might change the economic environment (the demand for schooling or the returns to education, for example) sufficiently to invalidate the predictions from the experimental setup.

Lastly, randomized evaluations are not exempt from a series of potential biases. Although most of these biases – except the so-called randomization bias – also apply to non-experimental and quasi-experimental designs, the difference in the case of random treatment assignment designs is that biases are well known and can often be corrected. Contrarily, biases caused by non-random treatment assignment are much harder to either sign or estimate. Next, we summarize the main biases that randomized trials (as well as other methods such as propensity score matching and regression discontinuity) can suffer from:

Sample Selection Bias

A first possibility is that the initial randomization is not respected. For example, schools in the control group might end up receiving the program as a result of pressure exerted by the local authorities. Or program administrators might deny treatment to certain eligible schools given logistical and accessibility constraints or managerial problems (endogenous program placement). Likewise, schools in the treatment group may not receive treatment simply because they decide to not take up the program given their expectation on how much they can benefit from treatment (self-selection). Even though the intended allocation of the program was random, the actual allocation is not. Consequently, the *intention to treat estimate* – the average treatment effect on the

randomly assigned treatments – differs from the *treatment on the treated estimate*, the average treatment effect on those who effectively participated in the program.

However, in most cases it is at least more likely that a school receives the program if it was initially allocated to it. The researcher can thus compare outcomes in the initially assigned group of schools and scale up the difference by dividing it by the difference in the probability of receiving the treatment in those two groups. This estimate is known as the LATE (Local Average Treatment Effect) estimate and was first introduced in Imbens and Angrist (1994). It is equivalent to estimating the treatment effect on the treated sample using the random treatment assignment as an instrument for take up. A major inconvenience is that it requires large sample sizes to perform well.

Sorting of Students or School Staff

As discussed in section III.A, students (or their parents) might respond to treatment in the school by withdrawing from or enrolling in treated schools. Changes in the enrollment and drop out behavior given the intervention can alter the distribution of observed and unobserved student attributes across treatment and control schools. A way to test how serious the bias is might be to explore changes in post-intervention enrollment patterns in treatment and control schools. If differences exist, one way to proceed is to control for changes in enrollment in the regression, as well as for as many school and student characteristics as are available. It is standard practice in the literature to control for parental education as a proxy for parental motivation to keep or withdraw their children in the school. Another possibility is to bound the treatment effects. When the outcome variable is itself bounded (binary or proportion outcomes), the lower (upper) bound of the treatment estimate is found by subtracting (adding) the estimated probability of being treatment (control) given a set of observable characteristics to the estimate of the treatment effect (Manski 1989, 1990). Altonji and others (2005a, b) suggest an alternative approach to setting bounds when the outcome variable is continuous.

Similarly, teachers and principals might react to treatment in the school. For example, they might choose to move to non-treatment schools with lower work loads if the intervention demands high levels of responsibility and accountability. On the other hand, more motivated and more adventurous teachers and principals might choose to work in a treatment school where they can enjoy a larger degree of autonomy. In either case, both observed and unobserved characteristics of the school are likely to change with changes in the staff composition and will be no longer balanced between treatment and control schools. Thus, it might be worth investigating and, if relevant, controlling for changes in the school staff composition throughout the evaluation period.

Attrition Bias

If the intervention has no set length of participation, schools in the treatment group might decide to not take up treatment in successive periods and drop out of the sample. Student and teacher attrition are more likely to occur in control schools if we assume that those who benefit from the program are less likely to migrate or exit the school than those who do not. Nonetheless, as noted above, students and teachers in a

treatment school might decide to exit the school if they interpret treatment as a sign of the school malfunctioning or if the work load given treatment is too heavy. Even if the initial sample was random, if the process whereby individuals (schools, students, or teachers) drop out of the sample is non-random, then the composition of the sample will be shifted, and estimates will suffer from attrition bias.

There exist statistical techniques, namely selection models extended to panel data, to mitigate the biases coming from attrition. These methods model selection into the sample – the probability of not dropping out – as a function of past observable characteristics and use the predicted probabilities to weight observations in the outcome (main) equation (Heckman 1976; Moffitt and others 1999). Tracking down attriters to find out their reasons to drop out is likely to improve the predictive power of the selection (on observables) equation. King and Ozler (1998) tackle large sample attrition in their evaluation of the Nicaraguan SBM reform using non-experimental data.

Spillover Effects

In the presence of spillovers, schools in the comparison group are indirectly affected by the treatment and consequently will not serve as a pure comparison group. For example, a higher degree of autonomy and community involvement in one school might trigger similar behaviors in neighboring control schools. This becomes more of a concern the higher the mobility of teachers, principals, and students and the closer schools are to each other. The natural and practical way to proceed in this case is to randomize the allocation of treatment across communities rather than across schools. In this case, it is important to take into account the appropriate sampling unit (community as opposed to school) and the grouped nature of the data when computing confidence intervals for the estimates of the impact of the program (clustering of standard errors).

Hawthorne Effect

It might also be that staff and students in treatment schools are initially more motivated and enthusiastic to undertake reforms because they know they are being treated. Their treatment status improves their "morale," which could result in improvements in performance likely to vanish as the enthusiasm wanes. In this case, the variation in the response observed under experimental conditions cannot be attributed solely to treatment. The term originated in a social psychology research project at the Hawthorne Plant of the Western Electric Company in Cicero, Illinois, from 1926 to 1932

John Henry Effect

The John Henry effect refers to situations where school agents (teachers, principals, students, parents) in either treatment or comparison group change their behavior, voluntarily or involuntarily, because they know they are part of an experiment (Duflo 2004) – or more generally, of an evaluation.

Randomization Bias

The administration and operation of a social experiment often involve a bureaucracy that might assign treatment or operate the program systematically differently in the experimental situation, with respect to its normal operation situation. This might generate systemic differences between schools that would normally be attracted to a SBM intervention from schools randomly assigned to the intervention. Heckman and Smith (1995) document the possibilities of such bias in actual experiments.

Substitution Bias

Substitution bias arises if the control group can receive some form of treatment that substitutes for the experimental treatment (Heckman and Smith 1995). For example, schools in the control group might receive benefits from other sources or of different nature than the SBM intervention – a compensatory program for instance – that might also affect outcomes. The researcher can minimize the potential for this bias by controlling for other programs the schools or the students are receiving in the analysis. Substitution bias can be really serious and even undermine an evaluation if the alternative program is put in place to compensate the control group for not receiving the SBM benefits.

To our knowledge, there is thus far no SBM evaluation in developing countries that uses experimental data. Although, several initiatives are under way in Indonesia, Kenya, Mexico, and Pakistan, results will not be ready for some time.

D. Qualitative Designs

So far we have been discussing quantitative methods. But often there are qualitative methods for data collection that can play an important role in evaluation. Qualitative techniques are used for carrying out evaluation with the intent to determine impact by the reliance on something other than the counterfactual to make a causal inference (Mohr 1995). The focus instead is on understanding processes, behaviors, and conditions – extremely important in the case of SBM since it is an innovation that requires significant changes in processes, behaviors, and conditions before one sees an impact in terms of schooling outcomes – as they are perceived by the individuals or groups being studied (Valadez and Bamberger 1994). Because measuring the counterfactual is at the core of impact analysis techniques, qualitative designs have generally been used in conjunction with other – quantitative – evaluation techniques.

Among the different qualitative methods are in-depth interviews. These entail asking questions, listening to and recording answers, and then posing additional questions to clarify or expand on a particular issue. Questions are open-ended, and respondents are encouraged to express their own perceptions in their own words. In-depth interviewing aims at understanding the beneficiaries' view of a program, their terminology, and judgments. Such interviews could be with individual respondents or group interviews (Marshall and Rossman 1995).

Focus group interviews are with small groups of relatively homogeneous people with similar background and experience. Participants are asked to reflect on the questions asked by the interviewers, provide their own comments, listen to what the rest of the group have to say, and react to their observations. The main purpose is to elicit ideas, insights, and experiences in a social context where people stimulate each other and consider their own views along with the views of others. Typically, these interviews are conducted several times with different groups so that the evaluator can identify trends in the perceptions and opinions expressed. One of the main advantages of this technique is that participant interaction helps weed out false or extreme views, thus providing a quality control mechanism (Chung 2000).

Other methods are observational, or firsthand observation of a program. The main purpose of observational evaluation is to obtain a thorough description of the program, including program activities, participants, and the meaning they attach to the program. It involves careful identification and accurate description of relevant human interactions and processes (Patton 1987).

Participant observation is at one end of the participation spectrum and consists of the evaluation observer becoming a member of the community or population being studied. The researcher participates in activities of the community, observing how people behave and interact with each other and outside organizations. The evaluator tries to become accepted as a neighbor or participant rather than as an outsider. The purpose of such participation is not only to see what is happening but to feel what it is like to be part of the group. The extent to which this is possible depends on the characteristics of program participants, the type of questions being studied, and the socio-political context of the setting. The strength of this approach is that the researcher is able to experience and presumably better understand any project impacts. The main weakness is that it is likely to alter the behavior that is being observed. In addition, ethical issues may arise if the participant observer misrepresents herself in order to be accepted by the community (Patton 1990).

Direct observation tends to be at the other end of the spectrum. It involves the systematic noting and recording of activities, behaviors, and physical objects in the evaluation setting as an unobtrusive observer. It can often be a rapid an economical way of obtaining basic socio-economic information on households or communities. The main advantage of this method is that if participants are not aware that they are being observed, then they are less likely to change their behavior and compromise the validity of the evaluation.

There is a growing acceptance of the need for integrating the different approaches to evaluation (Baker 2000). While quantitative methods are better suited to assess causality and reach generalizable conclusions, qualitative methods allow the in-depth study of selected issues, cases, or events and can provide critical insights into beneficiaries' perspectives, the dynamics of a particular reform, or the reasons behind certain results observed in a quantitative analysis. Combining quantitative and qualitative methods can often be the best vehicle for meeting the program's information needs. For example, qualitative methods can be used to inform the evaluation questions and the

questionnaire design, as well as to analyze the social, economic, and political context within which a program or policy takes place. Similarly, quantitative methods can be used to inform qualitative data collection strategies, including sample design, and to apply statistical analysis to control for household characteristics and the socio-economic conditions of different study areas, thereby eliminating alternative explanations of the observed outcomes.

SBM offers significant opportunities for mixed-method evaluation. The Nicaragua School Autonomy Reform provides a good example (Fuller and Rivarola 1998; Rawlings 2000). Quantitative methods following a quasi-experimental design were used to determine the relationship between SBM and learning. Qualitative techniques, including a series of key informant interviews and focus group discussions with different school based staff and parents, were utilized to analyze the context in which the reform was introduced, examine the decision-making dynamics in each school, and assess the perspectives of different school community actors on the autonomy process. The qualitative study pointed out that policy changes at the central level do not always result in tidy causal flows to the local level. In general, reforms are associated with increased parental participation as well as management and leadership improvements. But the degree of success with which reforms are implemented varies with school context. Of particular importance are the degree of impoverishment of the surrounding community (in poor communities, increasing local school financing is difficult) and the degree of cohesion among school staff (when key actors such as teachers do not feel integrated into the reform process, success at decentralization has been limited). Policymakers often ignore the highly variable local contexts into which new programs are introduced. The qualitative results point out that in the Nicaraguan context the goal of increased local financing for schools is likely to be derailed in practice, particularly in poor communities, and therefore merits rethinking. Gertler and others (2006) study of SBM in rural Mexico also used qualitative analysis to corroborate the quantitative results. The qualitative work consisted of discussions with parents, teachers, and school directors in beneficiary and non-beneficiary schools. The qualitative work suggests that more active participation by parents is behind the measured improvement in outcomes associated with the program. The program helps improve relations between parents and teachers, as well as the overall school climate.

IV. Summary of Evidence on the Effects of SBM Interventions

After more than a quarter century of SBM reforms around the world, there is still little conclusive evidence on the effects of these interventions. Many of the works assessing SBM have weak methodological designs with questionable identification strategies. Hence, little evidence allows for causal interpretations of the effects of the reform on outcomes. Because the focus of this note is on methods rather than on results, we will limit the discussion here to a brief summary of the findings in the (relatively) more sound studies – methodologically speaking. Santibáñez (2006) provides a

comprehensive review of the evidence available for both developed and developing countries. We refer the interested reader to this paper and to the references therein.

Evidence from Honduras and El Salvador suggests correlations between SBM reforms and improved school access and coverage in rural areas and poor communities (see for instance, Di Gropello 2006).

Although there seems to be consensus in that SBM reduces dropout and repetition rates, the magnitude of the effect varies across countries. Jimenez and Sawada (2003) find that third graders in EDUCO schools were more likely to continue studying than third graders in traditional schools. When the authors add a community participation variable in the estimation, the EDUCO coefficient loses magnitude and significance and community participation emerges as positive and statistically significant. The authors concluded that a significant portion of the EDUCO effect can be explained by community participation. Recall however, that the authors' efforts to correct for selection had weaknesses of their own. Skoufias and Shapiro (2006) use simple and difference-in-difference propensity score matching techniques and find that participation in the Mexican SBM program (PEC) decreases dropout and failure rates by 0.24 percentage points and repetition rates by 0.31 percentage points. Similarly, Murnane and others (2006) estimate 0.27 percentage points reductions in dropout rates in PEC schools using difference-in-difference methods on a slightly different sample of PEC and non-PEC schools. Gertler and others (2006) employ a similar methodology and find reductions in school-averaged failure and repetition rates of 0.4 percentage points as a result of the SBM component of Mexico's compensatory education program, the AGEs. As noted earlier, the authors corroborate the quantitative findings with a series of qualitative interviews with principals, parents, and teachers.

The evidence on student achievement is mixed and weakly identified. Hess (1999) suggests that after initial slippage, student achievement increased in Chicago public schools. In their meta-analysis of 29 SBM programs in the US, Borman and others (2003) conclude that schools that implemented the models for 5 years showed strong effects on achievement. In Honduras, the PROHECO intervention is correlated with higher test scores in science, although there is no evidence of significant effects on math or language test scores (Di Gropello and Marshall 2005). Student level propensity score matching estimates show positive effects of the Nicaraguan SBM reform on math scores for students in third grade (Parker 2005). Effects are negative for students in sixth grade and non-significant on Spanish scores for any grade.

The evidence from Central America also suggests positive correlations between autonomy and increased parental involvement as measured in terms of parents meetings with school personnel and parental visits to classrooms. On teacher effort, findings are unclear. Drury and Levin (1994), for instance, find changes in curriculum innovation and mixed results in terms of increased teacher professionalism. Sawada and Ragatz (2005) find that teachers in EDUCO schools spend more time meeting with parents and teaching, and that they are absent fewer days. Di Gropello and Marshall (2005) did not find significant differences in teacher effort nor pedagogical methods between PROHECO

schools and their traditional counterparts. As noted earlier, their results suffer from selection bias and are thus non-conclusive.

V. Political Economy and Ethics of SBM Evaluations

From a data requirement perspective, SBM evaluations might raise some ethical concerns as they are likely to involve the collection of sensitive personal data on students, teachers, and principals, in addition to administrative data collected for programmatic purposes (general census and school census data). In consequence, the data collection procedure might usually require clearance from a Protection of Human Subjects board that guarantees the protection of the interviewee's identity and her consent to participate in the survey and in the evaluation.

A more sensitive issue is that of denying benefits to students or schools on methodological grounds, once their need for benefits (eligibility status) has been identified. The issue is one of major concern in the application of experimental designs, although as noted in the previous section, randomized experiments do not necessarily have to deny benefits to anybody. Moreover, until a reform has been properly evaluated, it is wrong to assume that we are denying eligible schools a beneficial intervention. It can be argued that a more sensible approach is to first ascertain whether the reform does have a positive impact relative to the next-best alternative and for what type of schools. SBM reforms might represent a large administrative and logistical burden for some schools. The extra work load imposed on teachers and principals might consume time they previously devoted to teaching and managerial activities and hence damage teaching quality in the school. Moreover, some schools and/or local departments of education might lack the technical ability, experience, and capacity to assume autonomy on some areas. Arguably, these schools or education departments would be better off with a more limited transfer of responsibilities or with no transfer of responsibilities at all.

A few studies explore equity issues related to SBM reforms. They look at heterogeneous responses to increased autonomy and decentralization across schools and geographical areas with varying wealth and capacity levels. Findings are generally consistent with the existence of an efficiency-equity trade-off. Galiani and others (2005) show how the decentralization of education decisions to provincial governments and increased budgetary autonomy to secondary schools in Argentina resulted in larger inequalities in education outcomes. Schools in poor municipalities – defined as those where more than 30 percent of the households do not meet a list of basic needs – in weakly managed provinces experienced a reduction in test scores as a result of the transferring of responsibilities. In contrast, Eskeland and Filmer (2002) find a stronger correlation between increased autonomy and student test scores amongst the sub-sample of "poorest" primary urban Argentinean schools. Moreover, they do not find any indication that autonomy and parental participation are less correlated with learning for students from poorer households. Note that, unlike Galiani and others (2005), the Eskeland and Filmer (2002) study is at the student level and defines poor schools in terms of average family wealth in the school and poor households in terms of family wealth and maternal years of education. More importantly, Eskeland and Filmer (2002) suffers from

self-selection biases: identification is solely based on the inclusion of a large number of inputs in the test score production function and on a series of specification checks.

The evaluations on PEC also show that the decentralization of public management can improve outcomes in wealthy areas but has no impact in more disadvantaged areas. Skoufias and Shapiro (2006) find no significant effects of PEC on failure, dropout, and repetition in indigenous schools, as opposed to substantial reductions in outcome rates in non-indigenous schools. Similarly, Murnane and others (2006) report no significant reductions on average school drop out given PEC in schools in "low outcome" states, defined as those states with a lower value of the Human Development Index. Effects are however significant in "high" and "medium outcome" states, where effects are the largest. The authors propose the existence of differential capacities to support schools across states as a potential explanation. Overall, the crucial political economy issue these pieces of evidence raise is whether SBM reforms contribute to increasing inequalities by benefiting more disadvantaged schools less. The implication may be that more attention should be given to raising capacity of weaker schools.

Ethical concerns might also be present when SBM interventions are assigned to schools via competition. When schools are selected into treatment on the basis of the school improvement plan or proposal they present to the evaluation committee, better performing schools are likely to have a higher probability of receiving treatment as they are more likely to present better proposals. Then, benefits might be denied to the most needy (worse performing schools) who, as seen, might also be the less able to implement and benefit from reform measures.

VI. Final Considerations: Outstanding Issues for Evaluation of SBM Reforms

In this last section, we pinpoint some of the main issues we think future evaluations of SBM reforms should address in order to obtain more conclusive results that can better inform educational policy and orient the scale up of current interventions.

First, evaluations of SBM interventions should allow *longer time horizons* as effects might differ in the short- and long-run. For example, although we might see an initial response in teacher behavior, they might adjust to the intervention later on. Moreover, because these reforms imply changes in the school environment and in the relationships across the school community members, certain disruption and even negative impacts might be observed during the initial years of adjustment. Indeed, evidence from the US reveals that it might take at least five years before a successful SBM program can achieve results in learning outcomes (Borman and others 2003). New evaluation efforts should probably be oriented towards following cohorts of students over several academic years in both autonomous and comparable non-autonomous schools. They should also collect information on changes in the school organization and climate and on the involvement of local agents in school matters over the same period of time. Although very demanding in terms of time, money, and effort; collecting these data will help

address and correct for many of the biases that threaten existing evaluations. It will also contribute to the currently unresolved question on achievement.

Second, even if the final evaluation goal is to identify the effects of SBM on learning, evaluation efforts should also be able to *explain the channels and mechanisms by which increased autonomy results in better education*. Further work that models how the different school agents react to the intervention incentives is needed. A better understanding of both the mechanisms (parental involvement, increased monitoring, teacher effort, satisfaction) and the type of incentives (financial autonomy, personnel hiring and firing autonomy, autonomy over planning and instruction or over the curriculum) will nurture and refine the design of new policies. Furthermore, it will shed some light on the possibility of extending existing policies to environments different from those for which they were first designed. More and more precise data on school agents' responses to SBM incentives will be very useful in this respect.

Third, there is also a need for *more and better measures on the direct and indirect cost of implementing and managing SBM interventions*. Because SBM involves local agents more directly in school affairs, improved achievement – through greater monitoring of school personnel, better student evaluation, closer match between school needs and policies, and a more efficient use of resources – it is clearly an expected benefit. On the cost side, devolving more authority to parents, teachers, and principals should, at least directly, cost very little. There are, however, indirect costs borne not by the government but by the stakeholders themselves in terms of time, effort, and satisfaction. Some of the existing studies reveal how teachers and principals felt overworked or overstressed by the higher demands of responsibility and accountability, which can damage their teaching and managing practices. Data on time devoted to managing the intervention, extra responsibilities, teaching practices, etc. will inform the cost-effectiveness analysis.

Fourth, *more attention should be paid to the potential for increasing schooling inequalities* as a result of the implementation of SBM reforms. More generally, SBM evaluations should address heterogeneity issues in a more rigorous and systematic way in order to inform policymaking on best practices. However, measurement demands are considerable – both in terms of data required and feasible methods – in the treatment of heterogeneity.

Fifth, the lack of rigorous studies of SBM interventions – namely randomized or regression discontinuity designs – highlights the *need for more research that can lend empirical credibility to the many claims on SBM.* As noted earlier, the implementation of randomized evaluations raises some fairness concerns. Conducting partial randomizations within a school that involve less of a feeling of exclusion by non recipient students is not a feasible alternative in the design of SBM interventions. Nonetheless, in certain situations, capacity and logistic constraints might favor the random allocation of benefits. Whenever possible, we advocate the use of experimental evaluation designs. Not only do they overcome many of the problems often encountered when using other evaluation practices by identifying ex-ante a valid counterfactual but they also simplify the statistical analysis and provide reliable estimates than can inform policymaking.

Lastly, *quantitative evaluations should more often be complemented with qualitative evaluations*. Qualitative interviews with school agents might be very informative both if performed before and after the intervention is in place. If before, they will help form hypothesis, define the type of data that needs to be collected, and identify the main dimensions of heterogeneity of impact. They might also inform the intervention design through the ex-ante identification of the administrative problems departments of educations and schools might experience in supporting the intervention. In turn, these might help identify the reasons to participate in the program or to drop out of it. If carried out after the intervention, they might help assess the plausibility of the results and its interpretation and provide high quality information on some of the crucial aspects aforementioned: indirect costs, and processes and mechanisms by which increased autonomy improves educational outcomes. Case studies are of dubious causal value but are good for describing implementation dynamics.

Bibliography

Altonji, J., E. Todd and C. Taber. 2005a. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy* 113(1): 151-183.

Altonji, J., E. Todd and C. Taber. 2005b. "An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schools." *Journal of Human Resources* 40(4): 791-821.

Angrist, J., E. Bettinger, E. Bloom, E. King and M. Kremer. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." *American Economic Review* 92(5): 1535-1558.

Athey, S., and G. Imbens. 2006. "Identification and Inference in Nonlinear Difference-In-Difference Models." *Econometrica* 74(2): 431-497.

Baker, J. 2000. *Evaluating the impact of development projects on poverty : a handbook for practitioners*. Directions in development. Washington DC: World Bank.

Banerjee, A. and E. Duflo. 2006. "Addressing Absence." *Journal of Economic Perspectives* 20 (1): 117–32.

Bell, B., R. Blundell and J. Van Reenen. 1999. "Getting the Unemployed Back to Work: The Role of Targeted Wage Subsidies." *International Tax and Public Finance* 6(3): 339-360.

Bertrand, M., E. Duflo and S. Mullainathan. 2004. "How Much Should We Trust Differences-In-Differences Estimates." *The Quarterly Journal of Economics* 119(1): 249-276.

Bloom, H., J. Bos and S. Lee. 1999. "Using cluster random assignment to measure program impacts." *Evaluation Review* 23(4): 445-469.

Blundell, R. and M. Costa Dias. 2000. "Evaluation Methods for Non-Experimental Data", *Fiscal Studies*, 21(4): 427-468.

Borman, G., G. Hewes, L. Overman and S. Brown. 2003. "Comprehensive school reform and achievement: A meta-analysis." *Review of Educational Research* 73(2): 125-230.

Caldwell, B. 1993. "Leading the Transformation of Australia's Schools." *Network News* 5(4): 2-6.Caldwell, B. 1998. "Strategic Leadership, Resource Management, and Effective School Reform." *Journal of Educational Administration* 36(5): 445-461.

Caldwell, B. 2005. *School-Based Management*. Paris: IIEP-UNESCO.

Chaudhury, N., J. Hammer, M. Kremer, K. Muralidharan, and F. H. Rogers. 2006. "Missing in Action: Teacher and Health Worker Absence in Developing Countries." *Journal of Economic Perspectives* 20 (1): 91–116.

Chay, K., P. McEwan and M. Urquiola. 2005. "The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools." *American Economic Review* 95(4): 1237-1258.

Chung, K. 2000. "Qualitative Data Collection Techniques," in M. Grosh and P. Glewwe, eds., *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study*. World Bank: Washington, D.C.

Davidson R. and J. MacKinnon. 2003. *Econometric Theory and Methods*. Oxford University Press.

Dehejia, R. and S. Wahba. 2002. "Propensity Score-Matching Methods for Nonexperimental Casual Studies." *Review of Economics and Statistics* 84(1): 151-161.

De Grauwe, A. 2004. "School Based Management (SBM): does it matter?" Paper commissioned for the *EFA Global Monitoring Report 2005, The Quality Imperative*. UNESCO, Paris.

Di Gropello, E. and J. Marshall. 2005. "Teacher Effort and Schooling Outcomes in Rural Honduras," in E. Vegas (ed.), *Incentives to Improve Teaching*. Washington, D.C.: World Bank.

Di Gropello, E. 2006. "A Comparative Analysis of School-Based Management in Central America." World Bank Working Paper No. 72, The World Bank, Washington, D.C.

Drury, D. and D. Levin. 1994. "School-Based Management: The Changing Locus of Control in American Public Education." Report prepared for the U.S. Department of Education, Office of Educational Research and Improvement, by Pelavin Associates.

Duflo, E. 2004. "Scaling Up and Evaluation," in F. Bourguignon and B. Pleskovic, eds., *Annual World Bank Conference on Development Economics 2004: Accelerating Development*. Washington DC and Oxford: World Bank and Oxford University Press.

Duflo, E. and E. Saez. 2003. "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment." *Quarterly Journal of Economics* 118(3): 815-842.

Eskeland, G. and D. Filmer. 2002. "Autonomy, Participation, and Learning in Argentine Schools. Findings and Their Implications for Decentralization." World Bank Policy Research Working Paper No. 2766, Washington, D.C.

Fasih, T. and H. Patrinos. 2006. "Impact of Organization on School Performance School-Based Management." Human Development Network, World Bank, Washington, D.C. (processed)

Fuller, B. and M. Rivarola. 1998. "Nicaragua's Experiment to Decentralize Schools: Views of Parents, Teachers and Directors." Working Paper Series on Impact Evaluation of Education Reforms No. 5. World Bank, Washington, D.C.

Galiani, S., P. Gertler and E. Schargrodsky. 2005. "School Decentralization: Helping the Good Get Better, but Leaving the Poor Behind." (processed)

Gargani, J. and T. Cook. 2007. "How Many Schools? Limits of the Conventional Wisdom about Sample Size Requirements for Cluster Randomized Trials." Institute for Policy Research, Northwestern University. (processed)

Gertler, P., H. Patrinos and M. Rubio-Codina. 2006. "Empowering Parents to Improve Education. Evidence from Rural Mexico." World Bank Policy Research Working Paper Series No. 3935, Washington, D.C.

Glewwe, P. and M. Kremer. 2006. "Schools, Teachers, and Education Outcomes in Developing Countries," in E.A. Hanushek and F. Welch, eds,. *Handbook of the Economics of Education*. New York: Elsevier.

Hahn, J., P. Todd and W. Van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression Discontinuity Design." *Econometrica* 69(1): 201-209.

Hanushek, E. 2003. "The Failure of Input-Based Schooling Policies." *Economic Journal* 113(485): 64-98.

Heckman, J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models." *Annals of Economic and Social Measurement* 5(4): 475-492.

Heckman, J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47(1): 153-161.

Heckman, J. and J. Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives* 9(2): 85-110.

Hess, G.A. 1999. "Expectations, Opportunity, Capacity, and Will: The Four Essential Components of Chicago School Reform." *Educational Policy* 13(4): 494-517.

Hirano K., G. Imbens, D. Rubin and X. Zhou. 2000. “Assessing the effects of influenza vaccine in an encouragement design.” *Biostatistics* 1(1): 69-88.

Imbens G. and J. Angrist. 1994. “Identification and Estimation of Local Average Treatment Effects.” *Econometrica* 62(2): 467-476.

Jimenez, E. and Y. Sawada. 1999. “Do Community-managed Schools Work? An Evaluation of El Salvador's EDUCO Program.” *World Bank Economic Review* 13(3): 415-41.

Jimenez, E. and Y. Sawada. 2003. “Does community management help keep kids in schools? Evidence using panel data from El Salvador's EDUCO program.” CIRJE Discussion Paper N. F-236.

King, E. and B. Özler. 1998. “What's Decentralization Got To Do With Learning? The Case of Nicaragua's School Autonomy Reform.” Development Research Group, The World Bank, Washington DC. (processed)

Leithwood K. and T. Menzies. 1998. “Forms and effects of school-based management: a review.” *Educational policy* 12(3): 325-347.

Malen, B., R. Ogawa and J. Kranz. 1990. “What do we know about site based management: a case study of the literature - a call for research,” in W. Clune and J. Witte, eds., *Choice and Control in American Education*. London: Falmer Press.

Manski, C. 1989. “Anatomy of the Selection Problem.” *Journal of Human Resources* 24(3): 343-360.

Manski, C. 1990. “Non-parametric Bounds on Treatment Effects.” *American Economic Review Papers and Proceedings* 80(2): 319-323.

Marshall, C. and G. Rossman. 1995. *Designing Qualitative Research*. Thousand Oaks, California: Sage Publications, Inc.

Moffitt, R., J. Fitzgerald and P. Gottschalk. 1999. “Sample Attrition in Panel Data: The Role of Selection on Observables.” *Annale d’Economie et de Statistique* 55/56: 129-152.

Mohr, L. 1995. *Impact Analysis for Program Evaluation* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Murnane, R., J. Willet and S. Cardenas. 2006. “Did Participation of Schools in *Programa Escuelas de Calidad* (PEC) Influence Student Outcomes?” Harvard University Graduate School of Education. (processed)

Nir, A. 2002. “School-Based Management and Its Effect on Teacher Commitment.” *International Journal of Leadership in Education* 5(4): 323-341.

Odden, A. and Odden, E. 1994. "Applying the high involvement framework to local management of schools in Victoria, Australia." Working Paper The School-Based Management Project, University of Southern California.

Paes de Barros, R. and R. Mendoca. 1998. "The Impact of Three Institutional Innovations in Brazilian Education," in W.D. Savedoff, ed., *Organization Matters. Agency Problems in Health and Education in Latin America*. Washington DC: Inter-American Development Bank.

Parker, C. 2005. "Teacher Incentives and Student Achievement in Nicaraguan Autonomous Schools," in E. Vegas, ed., *Incentives to Improve Teaching*. Washington, D.C.: World Bank.

Patton, M. 1987. *How to Use Qualitative Methods in Evaluation*. Newbury Park: Sage Publications.

Patton, M. 1990. *Qualitative Evaluation and Research Methods*. Newbury Park: Sage Publications.

Raudenbush, S., X. Liu and R. Congdon. 2004. "Optimal design for longitudinal and multilevel research: Documentation for the Optimal Design software"

Rawlings, L. 2000. "Assessing Educational Management and Quality in Nicaragua," in M. Bamberger, *Integrating Quantitative and Qualitative Methods in Development Research*. Washington DC: World Bank.

Santibañez, L. 2006. "School-Based Management Effects on Educational Outcomes: A Literature Review and Assessment of the Evidence Base." The World Bank, Washington, D.C. (processed)

Sawada, Y. and A. Ragatz. 2005. "Decentralization of Education, Teacher Behavior and Outcomes," in E. Vegas, ed., *Incentives to Improve Teaching*. Washington, D.C.: World Bank.

Skoufias, E. and J. Shapiro. 2006. "The Pitfalls of Evaluating a School Grants Program Using Non-experimental Data." World Bank Policy Research Working Paper No. 4036, Washington, D.C.

Valadez, J. and M. Bamberger (eds.) 1994. *Monitoring and Evaluating Social Programs in Developing Countries*. Economic Development Institute of the World Bank Series. Washington DC: World Bank.

Van der Klaauw, W. 2002. "Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach." *International Economic Review* 43 (4): 1249-1287.

Wohlstetter, P. and K. Briggs. 1994. “The Principal's Role in School-Based Management.” *Principal* 74(2): 14-17.

Wylie, C. 1996. “Finessing Site-Based Management with Balancing Acts.” *Educational Leadership* 53(4): 54-59.

Zellner, A. and P. Rossi. 1986. “Evaluating the Methodology of Social Experiments,” in A. Munnell, ed., *Lessons from the Income Maintenance Experiments*. Federal Reserve Bank of Boston.

World Bank. 2003. *World Development Report: Making Services Work for Poor People*. Washington DC: World Bank.

wb171275
M:\! M&E\Impact Evaluation\Sector Focus\Final (Printed) Versions - Word and PDF\Doing IE Series - Final Word Versions\Doing_ie_series_10 SBM - Final.doc
01/02/2008 11:44:00 AM