Inter-American
Development Bank

Office of Strategic
Planning and
Development
Effectiveness (SPD)

TECHNICAL NOTE

No. IDB-TN-519

Guidelines for Impact Evaluation in Education Using Experimental Design

Rosangela Bando

September 2013

Guidelines for Impact Evaluation in Education Using Experimental Design

Rosangela Bando

Inter-American Development Bank

2013

Cataloging-in-Publication data provided by the Inter-American Development Bank Felipe Herrera Library

Bando, Rosangela.
Guidelines for impact evaluation in education using experimental design / Rosangela Bando.
p. cm. — (IDB Technical Note ; 519)
Includes bibliographic references.
1. Education. 2. Education—Evaluation. I. Inter-American Development Bank. Office of Strategic Planning and Development Effectiveness. II. Title. III. Series.
IDB-TN-519

http://www.iadb.org

Rosangela Bando is an Evaluation Economist Specialist with the IDB's Office of Strategic Planning and Development Effectiveness.

E-mail: rosangelab@iadb.org

Guidelines for Impact Evaluation in Education Using Experimental Design

Rosangela Bando[1]

Abstract

There has been extensive research and investment in strategies that aim to improve the quality of education around the world. But despite rigorous evidence, the question of what to do in a specific context usually remains only partially answered at best. As a result, numerous impact evaluations are being conducted with the goal of learning what works. The goal of the guidelines presented here is to summarize the evidence and provide references in a single document in order to save time for those implementing education impact evaluations through randomized control trials. The guidelines focus on supporting evaluations that help policymakers determine where to assign limited resources to improve the quality of education. The document has five sections. The first reviews empirical findings to provide a general idea of why education is important, which inputs matter, and how to conduct an impact evaluation. The second section provides guidance on how to define the impact evaluation hypothesis. The third section presents a methodology for selecting the sample and setting up a randomized impact evaluation. The fourth section provides information on what data to collect for the evaluation, and the final section looks at how to analyze data and strategies to adjust the analysis for changes in the original design. The guidelines are not exhaustive: their main contribution is to present a methodology to design an impact evaluation, outline the necessary inputs to start the design of such an evaluation, and provide a structured source of rigorous references.

JEL Classification: C21, C93, H43, I25, I38, J24

Keywords: Education, Impact Evaluation, Policy Evaluation, Randomized Trials, Experimental Design, Average Treatment Effect, Development Effectiveness

[1]Rosangela Bando is an Evaluation Economist Specialist with the IDB's Office of Strategic Planning and Development Effectiveness. E-mail: rosangelab@iadb.org. This document has benefited from comments made by Sebastian Galiani, Emma Näslund-Hadley, and an anonymous referee. Paloma Acevedo, David Einhorn and Xia Li contributed with editing.

Contents

List of Figures

List of Tables

Acronyms

ATE	Average treatment effect
ATOT	Average treatment on the treated
DID	Differences in differences
ENLACE	Evaluación del Logro Académico en Centros Escolares (Mexico)
ICC	Intra-cluster correlation
ITT	Intention to treat
IV	Instrumental variables
LATE	Local average treatment effect
MDE	Minimum detectable effect
OLS	Ordinary least squares
PIRLS	Progress in International Reading Literacy Study
PISA	Program for International Student Assessment
PSM	Propensity score matching
RCT	Randomized control trial
RD	Regression discontinuity
SMART	Strategic, measurable, attributable, realistic, and targeted
SUTVA	Stable unit treatment value assumption
TIMSS	Trends in International Mathematics and Science Study
TOT	Treatment on treated

Introduction

There has been extensive research and investment in strategies that aim to improve the quality of education around the world. But despite rigorous evidence, the question of what to do in a specific country, state, or city usually remains only partially answered at best. As a result, numerous impact evaluations are being implemented with the goal of learning what works.

The goal of the guidelines presented here is to the summarize the evidence and provide references in a single document in order to save time for those conducting education impact evaluations through randomized control trials. The aim is to provide a starting point to pose questions and present evaluation strategies based on the work done to date, as well as to present a road map of relevant and critical references. The document is written for individuals who have taken university degree econometrics. It focuses on supporting evaluations in order to help policymakers determine where to assign limited resources to improve the quality of education. Evaluation of interventions that aim to enroll students in school are out of the scope of this work. Enrollment and dropout decisions are discussed, but only in the context of interventions aimed at improving the quality of education for those already enrolled in school.

The document has five sections. The first reviews empirical findings to provide a general idea of why education is important, which inputs matter, and how to conduct an impact evaluation. It includes a discussion on basic theory in order to understand tradeoffs and restrictions faced by different individuals and institutions involved in the education process. The second section provides guidance on how to define the impact evaluation hypothesis. The third section presents a framework for selecting the sample and setting up a randomized impact evaluation. The fourth section provides information on what data to collect for the evaluation, and the final section looks at how to analyze data and adjust the analysis for changes in the original design.

The guidelines are designed to incorporate key ideas and advances on an ongoing basis so as to support future work.

1 STEP 1: Save Time by Learning What Is Already Known

This section reviews the theoretical and empirical work that has been done on the quality of education. To clarify the scope of this section, education literature can be classified in two stages. The stages refer to the process through which inputs to education lead to welfare outcomes. As diagrammed below:

		First stage		**Second stage**
Inputs and organization (measured in dollars and time)	→	Human capital production function (measured in test scores)	→	Wages, growth, etc. (measured in dollars and other units)

This section, as well as the guidelines in general, concentrates on the first stage. The first subsection discusses the second stage of the process in order to provide evidence on the importance of education, but the rest of the section focuses on the first stage of the literature. The second subsection includes a description of four models, with an emphasis on household investment. The goal is to provide a framework to understand the tradeoffs that decisionmakers may face when they determine the amount of inputs to devote to education. The third subsection reviews evidence on which inputs and organizational aspects may play a role in the quality of education. The aim is to give an idea of the factors for which we have evidence that play a role into determining the quality of education. The final subsection introduces the theory behind how to evaluate. It is not limited to education per se and it rounds out the elements needed to understand the steps outlined in the rest of the document. The study of education is diverse. The aim of this section is to provide a starting point. Providing a full description and discussion of the various paths that play a role in education is out of the scope of this work. The goal is to give a general idea of what matters and why, and to provide a road map with multiple references to studies that discuss in detail a specific subject within education.

1.1 Justify the Evaluation: Why Education Is Important

The goal of this section is to provide references with evidence on the importance of education. The evidence shows that education leads to economic growth, higher wages, adoption of new technologies, improved health outcomes, decreased fertility, higher participation in civic affairs, and decreased risky behavior. Additionally, education plays a role in reducing inequality. Table 1 shows a list of studies in which causality is attributed to the impact of education on the quality of life and

Education leads to economic growth.

human development. The table provides an initial list of studies that can support an evaluation. The context in which education provides benefits may be relevant: for example, improvements in education in the absence of property rights or in an open economy have not been found to foster economic growth (Hanushek and Woessmann, 2007).

Table 1: **Summary of evidence on the importance of education**

• Education promotes economic growth (Krueger and Lindahl, 2001).
• Quality is important because what students know is more important than school attainment alone in explaining economic growth (Hanushek and Woessmann, 2007)
• An increase in education increases wages; for example, an increase of 0.12 to 0.19 years of education increased wages by 6.8% to 10.6% in Indonesia (Duflo, 2001) and an increase in test scores in 1 standard deviation in literacy increased wages by 15% to 20% in Chile (Patrinos, Ridao-Cano, and Sakellariou, 2006)
• Human capital can account for between one-half and two-thirds of the income differences between Latin America and the rest of the world (Hanushek and Woessmann, 2012).
• Education increases the adoption of new agricultural technologies (Foster and Rosenzweig, 1995).
• Education decreases fertility (Thomas, 1996).
• Mother's education improves their children's health (Chen and Li, 2009).
• Education increases participation in civic activities (Dee, 2004).
• Education decreases risky behavior such as criminality, drug use, and teen pregnancy (Heckman, Stixrud, and Urzua, 2006).

Source: Prepared by the author.

1.2 A Framework to Think About Education

Economists see education as a service that is traded. Under this framework, there are agents on the demand side (usually households) that buy education services. On the supply side there are agents that sell education services (usually private institutions or the government). The market is the place where these services are traded. Governments are usually organized in such a way that the social protection ministry covers demand-side factors while the education ministry takes care of the supply-side and market factors. Therefore, a way to classify studies is to talk about whether they address behavior of individuals on the demand side, the supply side, or in the market (exchange):

Economists see education as a service that is sold by the government and private institutions and purchased by households.

Market (Exchange)
Demand (Households) $\iff$ Supply (Schools, tutors, etc.).

This section provides a brief introduction to four theories on human capital. It starts with the demand side by providing the basic intuition of the classical model of household decisions regarding investment in human capital. The section includes a simplification of the model by Glewwe and Jacoby (2004). which is introduced because, on average, household characteristics explain more variation in test scores than factors on the supply side (Hanushek and Luque, 2003; Mizala, Romaguera, and Urquiola, 2007; Romeo and Raffinetti, 2012). The section briefly discusses the classical model of firms' choices of training by Becker (1964) in order to explain the role of firms in education investments. Companies are likely to supply firm-specific training but not general-knowledge training. The section continues by looking at the supply side, including the main ideas behind political economy models that examine how governments decide how to allocate resources. Finally, the section briefly describes a model regarding the market with the signaling model by Spence (1973). This model is interesting because while education can tell us who is more productive, it does not increase productivity. The section concludes by providing references to other models that look at specific market features.

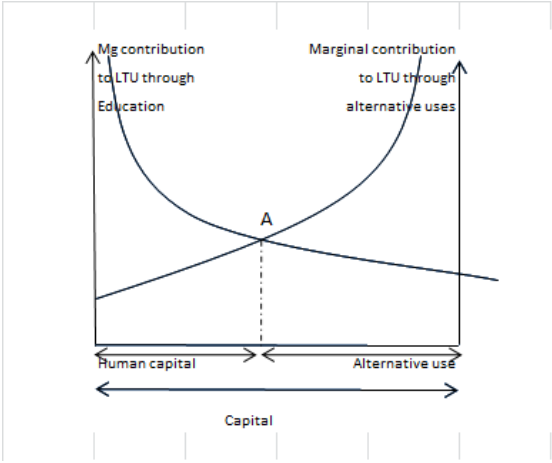
This review clearly outlines the range of incentives, trade-offs, and constraints faced by the different players in the education process. It is important to note that this discussion is limited in the context and range of modeling in education. These models are included because it is important to have some idea of the literature necessary to support the rationale of a given evaluation. Those conducting evaluations should look for models specific to the program they are evaluating. See Mas-Colell, Whinston, and Green (1995) for a formal derivation of the market framework.

In the classical model, parents maximize lifetime utility, which is a function of consumption and accumulated human capital. The optimal allocation of resources is such that the relative marginal return to human capital production equals the relative marginal return in consumption.

Households will invest to the point where the marginal contribution to utility equals the marginal utility of those resources for alternative uses.

Consider a household that experiences two periods of time with school-age children. Parents in the household are endowed with resources $R \geq 0$ at time zero. In this period, parents allocate resources either to current consumption C or to investment in human capital accumulation M. Children go to school in the first time period. The household only benefits from returns to education after the schooling period is over at time 1. Suppose there are also inputs to human capital production that are exogenous to the household, such as state-supplied inputs B. Therefore, the human capital accumulation function is given as $Y = Y(M, B)$, where the function Y is a neoclassical human-capital production function, and therefore $Y' > 0$ and $Y'' < 0$. Y may include other parameters that capture factors such as an individual child's ability, school characteristics, etc. The utility of the household U is a concave function of consumption C so that $U' > 0$ and $U'' < 0$. The household allocates

Figure 1: **Household Decisions on Investments in Human Capital**



Note: LTU = Lifetime utility.
Source: Prepared by the author.

resources to human capital and consumption so as to maximize utility. Note that with a binding budget restriction, consumption equals remaining resources after investment in education. Therefore $C = R - M$. Let δ denote the subjective discount factor. Then the household problem is

$$max_{\{M\}} U(R - M) + \delta Y(M, B), \tag{1}$$

subject to the borrowing constraint $R \geq 0$ and non-negativity conditions, $M \geq 0$, $C \geq 0$. The first-order condition for the solution to the problem described in equation 1 implies that the marginal returns to lifetime utility through consumption must equal the marginal returns to lifetime utility through investment in human capital after the schooling period is over:

$$U'(R - M^*) = \delta Y'(M^*, B), \tag{2}$$

Note that the optimal level of household investment M^* is a function of the parameter B. Therefore, one would expect household investment to adjust to other inputs. Figure 1 is a graphic representation of equation 2. The horizontal axis represents the amount of resources available to the household. The optimal amount of investment in human capital will be such that the marginal contribution to utility through education is equal to the marginal utility of those resources for alternative uses. Intuitively, this means that the last dollar invested in education should provide as much additional expected utility as the last dollar invested in alternative uses. This is point A in the figure.

Note that households and government subsidies finance education, but firms play a limited role. Becker (1964) pointed out that education leads to skills that can be classified as general or firm-specific. General skills are those that are useful in a wide variety of settings. As a result, an individual can benefit from them in the labor market. Firm-specific skills are those that are useful only at a specific firm. Becker maintained that in competitive markets, firms would only provide general training if there was no cost to them. Firms will invest in a worker up to the point where the last dollar used to pay for the services equals the additional benefit provided. The model also predicts that people with more education are more productive, receive higher wages, and are trained when they are young. Another important prediction of the model is that an additional year of education must be approximately equal to the interest rate. This has been observed empirically in the United States.

Firms generally invest in firm-specific skills and not general skills.

Decisions on investment on the supply side depend on whether provision is public or private. Private institutions will invest up to the point where marginal benefits, as defined by fees and the number of students enrolled, equal marginal costs. These models are usually relevant in higher education contexts. Decisions on investment in government-provided education are modeled in the political economy literature. These models predict that poverty, electoral dates, political alignment, and level of competence among parties will influence resource allocation.

There are two dimensions across which resources are allocated: groups and time. Government officials allocate so as to maximize votes. If voters can attribute transfers to the incumbent, then voting areas with a higher population density, swing areas, and poor areas where vote-buying has a lower cost are more likely to receive investments (Dixit and Londregan, 1996). If voters do not specifically identify who is providing the transfers, then the incumbent will allocate those transfers to politically aligned groups in order to avoid "leakage" (Arulampalam et al., 2009).

Political economy models predict that governments will invest more in areas with high poverty and during time periods close to electoral dates.

Regarding the time dimension, government officials will allocate more resources before elections because voters like low taxes and high government expenditures. Other models predict more government spending during elections for candidates to signal candidate ability to provide public goods (Rogoff, 1990). Clientelism is the exchange of goods and services for political support, and it may lead to inefficient and nonequitable allocation of resources (Díaz-Cayeros, 2008; Bullock, 2004) and to the restriction of public services (Ferraz and Finan, 2008). Pritchett and Filmer (1997) state that political decisions on education are made based on the power of stakeholders in the system, and that those decisions do not necessarily focus on maximizing outcomes. See Banerjee and Somanathan (2007) for more information on government investment in public goods.

The final set of models discussed in this section pertains to what economists call market failures. Market failure models describe cases in which the allocation of education services is not efficient in a free market or exchange. These models include information asymmetries, as in the signaling model (Spence, 1973), non-competitive markets (as in discussions on vouchers, see Hanushek and Woessmann (2007), and Barrera-Osorio (2007)), principal agent-problems (as in teacher incentives, see Bruns, Filmer, and Patrinos (2011)) , and externalities (as in peer effects, see Gould, Lavy, and Paserman (2003)). In Spence (1973) signaling model, people have either high H or low L ability and employers cannot observe what level of ability a given individual has. People with higher ability have a lower cost of education. Therefore, if employers offer high wages to educated employers, then high-ability people will be educated so as to become eligible. In a separating equilibrium, the cost of education is too high for low-ability individuals. As a result, high-ability individuals choose to get educated and are employed. Note that the signaling model does not require education to increase productivity. A school diploma works only as a signaling device.

The signaling model does not require education to increase productivity. A school diploma works only as a signaling device.

This section has reviewed models that aim to determine the extent of investment and education, and therefore education outcomes. Other theories are proposed for specific inputs but are beyond of the scope of this document. The next section reviews the literature on specific investments in education. This branch of the literature aims to evaluate which kinds of inputs (or investments) are effective in a given context.

1.3 Evidence of Inputs that Matter for the Quality of Education

Identifying the extent to which (and how) different variables contribute to student learning is difficult (Vegas and Umansky, 2005). Debates continue on which inputs should be included and how they should be measured (Hedges, Laine, and Greenwald, 1994). Even with a specific input, the literature usually shows mixed evidence on the effects of inputs on education outcomes. For example, Harris and Sass (2011) review studies on teacher training and Bruns, Filmer, and Patrinos (2011) discuss school-based management. Differences across findings may be attributed to the context and characteristics of the population being studied. Another challenge is measurement. A given input is usually provided by more than one source and in more than one way, making it challenging to measure. Some inputs, such as the quality of human interaction, are difficult to quantify. In an attempt to identify what inputs matter in the human capital production function, some authors have looked at sources of variation. Mizala, Romaguera, and Urquiola (2007) found in Chile that 75% to 86% of test score variation across schools is explained by parental schooling and household income. The authors also found that one-third of test score variation is between schools. This means that the worst students at a "good"

In Italy, the variation of test scores is 83% within the classroom, 11% across groups within schools, and 6% across schools.

school will score worse than the best students at a "bad" one. Romeo and Raffinetti (2012) find that in Italy, the variation of test scores is 83% within the classroom, 11% across groups within schools, and 6% across schools. Appendix 6 shows how to decompose test score variation in groups. Using census data from the standardized mathematics test in public schools in the state of Puebla in Mexico (*Evaluación del Logro Académico en Centros Escolares* - ENLACE), this methodology allows for decomposing the variation as 49% within student variations, 11% across schools, 11% across localities, 2% across municipalities, 1% across years, 9% across cohorts, and 17% not explained.

Table 2 presents a brief summary of inputs that have been found in the listed studies to cause changes in test scores. The objective is to provide a list of references that review the evidence of the factors that determine the quality of education. The studies included are peer reviewed and have rigorous evaluation designs. The main idea of each study is stated, but only briefly so as to keep the list practical. The table includes the order of magnitude of the effect in standard deviations to provide a sense of the potential relevance that a factor may have. Note that the magnitude of the effect depends on the context of the evaluation and not just the intervention. The list of studies is not likely to be exhaustive and should not replace a careful review. A wider and more detailed review of the role of inputs on education outcomes can be found in Vegas and Petrow (2008) Inputs are categorized in five groups: Y_0 (investments made in the past), I(inputs directly related to instruction), F(inputs not directly related to instruction), and according to whether the provision comes from the supply side S or demand side D.

Table 2: **Summary of Studies on Inputs that Cause Changes in Test Scores**

Category	Input	Effect size in test score standard deviations	Reference	Methodology
Investments made in the past (Y_0)	Access to pre-primary schooling	Increase of 0.33	Berlinski, Galiani, and Gertler (2009)	DID
	Increase 1 year of age upon entry	Decrease of 0.4	McEwan and Shapiro (2008)	RD
	Early literacy	Increase of 0.2	Woessmann and Fuchs (2005)	OLS
Instruction provided on the supply side (I^S)	Increase of one standard deviation in teacher subject knowledge (test scores)	Increase of 0.1	Metzler and Woessmann (2010)	OLS
	Improve teacher ability to teach rational numbers (training + coaching)	No effect	Garet et al. (2011)	RCT
	Pedagogy and content knowledge training (48 hours of instruction + 20 hours of coaching) + resource materials	No effect	Garet et al. (2011)	RTC
	Alternative versus traditional teacher certification	No effect	Constantine et al. (2009)	RCT
	Teacher training before service as induction	No effect	Glazerman et al. (2008)	RCT
	Less teacher experience and below-average credentials	Decrease of 0.21	Clotfelter, Ladd, and Vigdor (2010)	OLS

Continued on next page...

... table 2 continued

	Input	Effect size	Reference	Methodology
	Performance-based promotion	No effect	McEwan and Santibanez (2005)	RD
	Performance pay	Increase of 0.27	Muralidharan and Sundararaman (2011)	RCT
	Teacher incentives	No effect	Glewwe, Ilias, and Kremer (2010)	RCT
	Monitoring such as teacher attendance tracked by photos	Increase of 0.17	Duflo, Dupas, and Kremer (2011)	RCT
	Diagnostic tests and feedback with low-stakes monitoring	No effect	Muralidharan and Sundararaman (2010)	RCT
	Indian Balsakhi tutoring	Increase of 0.27	Banerjee and Somanathan (2007)	RCT
Instruction provided on the demand side (I^D)	Incentives such as merit scholarship for girls.	Increase of 0.28	Glewwe, Kremer, and Moulin (2009)	RD
	Parent with university degree	Increase 0.41 in Argentina and 0.19 in Colombia	Woessmann and Fuchs (2005)	OLS
	Parents' occupational status	Increase of two-thirds of a school year		

Continued on next page...

... table 2 continued

	Input	Effect size	Reference	Methodology
	Report card to parents	Increase of 0.1	Andrabi, Das, and Khwaja (2009)	RCT
	Parents' choice of school through vouchers	Increase of 0.2	Angrist et al. (2002)	RCT
	for accountability accountability	Increase in learning outcomes	Bruns, Filmer, and Patrinos (2011)	RCT and others
Inputs not related to instruction provided on the supply side (F^S)	Decrease Class size from 20 to 15	Increase of 0.21	Krueger and Whitmore (2000)	RCT
	Textbooks	Increase of 0.36, mostly benefiting top students	Lockheed and Hanushek (1988)	
	Computer-assisted instruction	Increase of 0.37	Banerjee and Somanathan (2007)	RCT
	School autonomy over personnel management and process decisions	Increase of 0.2	Fuchs and Woessmann (2007)	OLS
	School-based management	Increase of 1.50	Sawada and Ragatz (2005)	IV and PSM

Continued on next page...

... table 2 continued

	Input	Effect size	Reference	Methodology
	Autonomy and parental participation	Increase of 0.05, with the largest effect on the poorest	Eskeland and Filmer (2002)	OLS
	External exit exams	Increase of 0.04	Fuchs and Woessmann (2007)	OLS
	Increase per student spending	No effect	Woessmann (2003)	OLS
	School funding has little effect on student performance or, equivalently, schools do not matter and only families and peers affect student performance	No effect	Coleman (1966) (not causal but commonly cited)	
Inputs not related to instruction provided on the demand side (F^D)	Improve Nutrition at 12 to 24 months	Increase of 0.8	Glewwe, Jacoby, and King (2001)	IV
	Iron supplements	Increase of 0.4	Luo et al. (2010)	RCT

Continued on next page...

... table 2 continued

Input	Effect size	Reference	Methodology
Books at home	Increase of 0.56 in Argentina and 0.18 in Colombia for more than 200 books at home	Woessmann and Fuchs (2005)	OLS
Increase home educational resources (PISA index) one standard deviation	Increase of one-third of a school year	Woessmann and Fuchs (2005)	OLS

Note: DID = differences in differences; IV = instrumental variables; ; OLS = ordinary least squares with fixed effects; PISA = Program for International Student Assessment: PSM = propensity score matching; RCT = randomized control trial; RD = regression discontinuity.
Source: Prepared by the author.

An input of special interest is the initial test score Y_0. Assume the education production function is one such that

$$Y' = f(Y), \tag{3}$$

The solution to equation 3 is not linear if $f(Y)$ is not constant. Therefore, it is unlikely that a linear specification would correctly describe test score changes. In the United States, increases in test scores have been assessed at decreasing rates. These estimates, taken from Hill et al. (2007), are shown in Table 3.

Table 3: **Annual Gains for Students in Standard Deviations**

Grade	Reading	Math
Grade K to 1	1.52 (+/- 0.21)	1.14 (+/- 0.22)
Grade 1 - 2	0.97 (+/- 0.10)	1.03 (+/- 0.11)
Grade 2 - 3	0.60 (+/- 0.10)	0.89 (+/- 0.12)
Grade 3 - 4	0.36 (+/- 0.12)	0.52 (+/- 0.11)
Grade 4 - 5	0.40 (+/- 0.06)	0.56 (+/- 0.08)
Grade 5 - 6	0.32 (+/- 0.11)	0.41 (+/- 0.06)
Grade 6 - 7	0.23 (+/- 0.11)	0.30 (+/- 0.05)
Grade 7 - 8	0.26 (+/- 0.03)	0.32 (+/- 0.03)
Grade 8 - 9	0.24 (+/- 0.10)	0.22 (+/- 0.08)
Grade 9 - 10	0.19 (+/- 0.08)	0.25 (+/- 0.05)
Grade 10 - 11	0.19 (+/- 0.17)	0.14 (+/- 0.12)
Grade 11 - 12	0.06 (+/- 0.11)	0.01 (+/- 0.11)

Source: Hill et al. (2007)

A broader view of what matters in education without conditioning for causal impacts provides some hints on key lessons that tend to hold in spite of a change of setting and time. First, it is likely that there is no single set of inputs that matters more in every context. In other words, there is no "magic bullet" in education (Kielstra, 2012). Second, how and when resources are provided matters more than the amount of resources. For example, early childhood interventions seem to be more effective than later remedial interventions (Berlinski, Galiani, and Gertler, 2009; McEwan and Shapiro, 2008), and culture seems to play a role in how learning processes develop (Kielstra, 2012; Bruns, Filmer, and Patrinos, 2011; Weber, 2010). Third, teachers are found to be important, but there is no set of observable characteristics that define a good teacher or an intervention that transforms all teachers into excellent teachers. In the US, it has been found that a one standard deviation increase in average teacher quality for a grade raises average student achievement in the grade by at least 0.11 standard deviations (Rivkin, Hanushek, and Kain, 2005). Successful systems have recruited the best graduates by providing status and salaries

How and when resources are provided matters more than the amount of resources.

comparable to other professions, provided continuous training, and presented clear goals and expectations for teachers. Salaries, on the other hand, have been shown to have little effect (Bruns, Filmer, and Patrinos, 2011; Hanushek, 2011). Fourth, information and the ability to act on it are critical for accountability, which in turn improves test scores. Examples include school choice and parent report cards (Bruns, Filmer, and Patrinos, 2011; Banerjee et al., 2008).

This section has summarized the literature on why education is relevant for development, the main theories that determine investment in education, and evidence of inputs that matter in education, along with information that provides hints on what the human capital production function may look like. The section provided a general idea of what the literature has found and the main logic behind education determinants. While the contents in the section should allow for putting an evaluation in context, they should not replace a careful review of the specific intervention considered for evaluation. This section has also stated the limitations of what is known and why. Before starting the impact evaluation design, the next section provides background information to justify the choice of a randomized control trial, the assumptions, and the interpretation of the results.

1.4 How to Evaluate

The goal of an impact evaluation is to assess the effects of a program by comparing outcomes with and without the intervention. A **causal effect** for individual i is defined as the difference between an outcome with and without a program:

$$CE_i = Y_i(T) - Y_i(C), \tag{4}$$

How credible the results of the evaluation are depends on how valid the comparison group is.

Where $Y_i(1)$ and $Y_i(0)$ are potential outcomes if the individual had received the treatment or not, respectively. The fundamental problem of causal inference is that we can never observe outcomes under both scenarios for a given individual. A proposed solution is to identify average causal effects for a given population. A sample representative of such a population is exposed to treatment and another sample is not. Using this design, we can observe both the average outcome in the population under treatment and the population with no treatment. How to create these two groups is crucial. For example, in a teacher training program that aims to improve test scores of students, it is not possible to just allow teachers to volunteer for the training, because more motivated teachers are more likely to participate. Under that scenario, if we compared the average test scores of students with teachers who volunteer for the training to that of students with teachers who are not trained because they did not volunteer for the training, then we would not be able to tell how much of this difference can be explained by the program and how much by differences in teacher motivation. The proposed solution is to randomly assign teachers to a group to be trained (treatment) and others to serve as a comparison group (control). The role of each group is to provide information

under each scenario. The role of the comparison group is to provide information on what would have happened in the absence of the program. Given that random selection (for example, by a lottery) is not likely to assign more motivated teachers to one group or the other, we can then attribute differences in the final outcomes (in test scores, in our example) to the intervention (teacher training). This is true not only for motivation but for other characteristics of the teacher, the students, the school, or the environment that we cannot observe. How credible the results of the evaluation are depends on how valid the comparison group is. An evaluation is **internally valid** if it uses a valid comparison group. The **Average Treatment Effect (ATE)** is defined as

$$\begin{aligned} ATE &= E[Y_i(T) - Y_i(C)] \\ &= E[Y_i(T)] - E[Y_i(C)] \\ &= E[Y|T] - E[Y|C], \end{aligned} \tag{5}$$

Two assumptions are required to identify the average treatment effect:

1. Stable unit treatment value assumption (SUTVA). The treatment status of an individual does not affect the potential outcome of others. Noninterference assumes no spillovers, general equilibrium effects, and no variation in treatment. An example of a violation would be to provide vouchers to significantly more students than the private sector can accommodate, or to provide vouchers with greater value for poorer families[2]. Non-variation in treatment implies that individual behavior does not change treatment intensity. An example of a violation is a one-to-one tutor program where children with more motivated parents attend more and/or longer sessions.

2. Unconfoundedness (or strong ignorability): $Y(T), Y(C) \perp DT$. Treatment assignment is independent of the outcomes in either hypothetical scenario with or without program benefits[3].

Randomization guarantees ignorability. For SUTVA, you need to control for treatment variation and noninterference. In this framework, one should first take a random sample of the population of interest, then randomly assign treatment and control to the units, and then estimate the average treatment effect. When things in the original design do not go as planned, additional assumptions are necessary to try to assess the effects of a program. For example, individuals may move from

Design of the evaluation should rely on as few assumptions as possible.

[2]Examples of violations include training for too many people flooding the market with qualified job applicants (interference) and some patients getting extra-strength aspirin (variation in treatment) . The examples are taken from the handout on "Causal Inference" by Patrick Lam, available at www.people.fas.harvard.edu/~plam/teaching/methods/causal/causal.pdf

[3]One violation, for example, is omitted variable bias. The example is taken from the handout on "Causal Inference" by Patrick Lam, available at www.people.fas.harvard.edu/~plam/teaching/methods/causal/causal.pdf.

treatment to control groups and the other way around, or people may drop out of the sample. The design of the evaluation should aim to add as few assumptions as possible.

The guidelines presented here focus on experimental evaluation, which requires more plausible assumptions when compared to other evaluation approaches. There are other nonexperimental techniques such as regression discontinuity, instrumental variables, differences in differences, and matching, but discussion of these methodologies is beyond the scope of this document. A good starting point to explore these alternatives is Gertler et al. (2011).

The next sections provide steps to guide the impact evaluation of a program. Section 2helps to define the evaluation hypothesis. Section 3 helps in the selection of the appropriate sample and the creation of treatment and control groups. Section 4 provides a guide on what information should be collected. Section 5 presents the tools to carry out the analysis and deal with changes to the evaluation design.

2 STEP 2: Define the Questions to Be Answered

The objective of this section is to specify the main evaluation research question. The hypothesis needs to be very specific because the design of the evaluation evolves around the question posed. To achieve this goal, this section explains how to answer the six five W's: Why, Who, What, When, Where, and hoW. In practice the question that can be answered is a mix of what is desirable to learn and what is possible to evaluate. A single evaluation cannot answer all questions, so priorities should be defined. It is a good practice to set evaluation hypotheses before the evaluation starts in order to prevent bias in the findings by the researcher and data mining. If questions (or outcomes) are added after data were collected, a justification will be needed for these changes. Additionally, a clear evaluation question allows for planning to ensure collection of relevant data.

2.1 Establish WHY: For What Purpose Will the Evaluation Be Used?

The WHY question aims to establish evaluation goals. Ask yourself: *Why and in what way will I use the evaluation results?* What information is key for decision-making for future implementation? Impact evaluation quantifies the effects of a program on result indicators. It should also quantify changes in other dimensions that can improve the understanding of how these results were achieved and under what circumstances. Gertler et al. (2011) suggest impact evaluation should aim to:

1. Justify use of resources
2. Provide information to make an informed decision on budget allocation versus alternative uses
3. Provide information to adjust plans for expansion
4. Improve program components and assessments
5. Provide guidelines to other implementing agencies.

The authors also list the following criteria to determine whether or not an impact evaluation should be carried out. To be relevant to evaluate a program, an impact evaluation should be:

1. Innovative: It is testing a new, promising approach. (See the literature review in Section 1 for a starting point.)
2. Replicable: The program can be scaled up or applied in a different setting.

3. Strategically relevant: The program is a flagship initiative, requires substantial resources, covers or plans to expand to a large number of people, or could generate substantial savings.

4. Able to test the untested: Little is known about the effectiveness of the program globally or in a particular context.

5. Influential: The results will be used to inform key policy decision

Impact evaluation complements other assessments such as qualitative studies and cost-benefit analysis. Qualitative studies usually help to construct the story of why changes take place among different indicators included in the causal chain. If information on benefits associated with alternative uses of program funds is available, then cost-effectiveness analysis should be included.

2.2 Establish WHO: The Unit of Analysis

The WHO question aims to establish the unit of analysis. Ask yourself: *Who are the people that will be making key changes to obtain results?* The answer should aim to provide information useful for future policy implementation. Individual-level data are usually preferred because within-school test score variation is larger than across-school variation (Mizala, Romaguera, and Urquiola, 2007) and outcomes are likely to depend on initial performance (Lockheed and Hanushek, 1988). It is more likely that different responses will be found in individual test scores within a given classroom than across school averages. Sometimes it is not possible to collect data at the individual level. For example, the operative side may be concerned with privacy issues. In this case, evaluation is done answering the question "What is the effect of teacher training on the average mathematics test scores in the school?"

Individual-level data are usually preferred because within-school test score variation is larger than across-school variation.

2.3 Establish WHAT: Intervention and Randomization

The WHAT question should determine which features of the program can be evaluated. Ask yourself: *What is the key feature of the program that will bring results?* To answer this question, think about two program features.

First, programs usually are composed of multiple components implemented at different levels. Evaluation with causal attribution is possible for components for which treatment can be randomly differentiated. For example, it is common for a teacher training program to be accompanied by changes in curriculum and administrative processes. Teacher training may be phased in and reach schools at different times, but changes in curriculum and administrative processes are likely to benefit all schools at the same time. As a result, teacher training can be evaluated because some teachers will not benefit from the program at the time others will, allowing for

Give priority to program features that are expected to bring results and are costly.

a control group. On the other hand, changes in curriculum and administrative processes will benefit everyone at the same time, eliminating the possibility of creating a control group. As a result, the evaluation will be limited to the teacher training feature of the program.

Second, to evaluate a specific program feature it is necessary to create a specific group. Cross-cutting design is covered in the next subsection. As evaluation costs and sample size are usually limited, prioritize features that are expected to bring results. Costly features should also receive priority. Randomization should be done once the relevant program features to be evaluated are determined. In some cases it may not be possible to do this. Be aware of the implications of choice of unit of analysis for policy recommendations.

Be specific about the definition of program features. To illustrate how the WHAT changes with specific program features, let's continue with the example of the evaluation of the teacher training program. Suppose that teacher training is given monthly during the school year. Suppose teacher mobility is high and any teacher who arrives at the school joins the monthly training program. In this scenario, the evaluation question is: "What is the effect of providing monthly teacher training to the group of treated schools on standardized student test scores?" Another possibility is that the training is assigned to a specific teacher. If the teacher moves to a different school, the replacement teacher does not benefit from the training program. In this second scenario the question is: "What is the effect of training the initially assigned teacher on standardized student test scores?" The differentiation between these two questions is important in cases where teacher and/or student mobility is high.

The choice of program features to be evaluated determines the cost and use of the evaluation.

The operative aspects of the program play an important role in determining what questions can be answered. For example, if teacher training can only be provided to all teachers within a district, then operational issues only make it possible to differentiate treatment at the district level. In this case, randomization should be attempted at the most disaggregated level feasible: the district. More observations are needed when randomization is done at a more aggregated level. As a result, the cost of the evaluation increases. Section 3 discusses the number of observations needed, which depends on how many units are randomized.

Sometimes the population out of which the sample can be selected is restricted and it is not possible to have a large number of units to randomize. For example, consider an infrastructure intervention in a state where geographic proximity is associated with economies of scale. In this case the operative aspect limits implementation to all beneficiary schools within a municipality. First consider the relevance of the questions to be answered. Impact evaluation should shed light on information that supports better-informed decision making in the future. Keep in mind that a small sample may not be representative of the population of interest. If there will be no use for this information, then the investment in resources in impact evaluation may not be justified. If the question is worth answering, then alternative methods can be used to analyze data to maximize the power of the hypothesis tests.

Impact evaluation should shed light on information that supports better-informed decision making in the future.

Section 5.2.2 discusses some of these methods.

Duflo, Glennerster, and Kremer (2008) describe some methodologies to create treatment and control groups once the features to be evaluated have been chosen. First, randomization may be possible within schools. For example, Banerjee and Somanathan (2007) randomized groups within schools to ensure cooperation from school authorities. Every school in the study received a tutor in every year. However, some schools were asked to use the tutor in grade 3 and others in grade 4 based on random assignment. A second option would be to randomly phase in the program when possible in order to differentiate a treatment and control group based on the timing of the intervention. This is common when operational capacity does not allow for covering the whole population at once. A third choice is to randomly invite individuals to participate in a given program within the school. A fourth option is to accept the fact that the program cannot be evaluated as a whole and explore if there are other relevant policy questions that can be answered. An example is a program where all individuals must participate. For example, in an evaluation on how to better provide incentives to community instructors, we randomly chose to provide upfront payments every three months to a randomly selected group, versus the status quo group, which receives payments at the end of every month. All community instructors receive the same payment, but we expect to check if credit constraints are playing a role in the dropout rates of community instructors.

Randomize across groups or differentiate program features in order to introduce variation in key program features.

2.3.1 Ask More than One Question When Possible-Cross-cutting Design

Sometimes it may be relevant to ask more than one question. For example, assume it would be useful to ask about the effect on students of monthly teacher training and/or the provision of books. In this case, four groups can be created allocating teachers to each group randomly. Each group would allow the evaluator to observe how students respond on average to a given provision of program components. Table 4 shows the four groups and the averages calculated from each one. With estimation of expected test scores for each group, it would be possible to evaluate the effect of teacher training alone, book provision alone, and both. To achieve this design, first randomly assign half of the group to teacher training (treatment 1) and half to serve as controls (control 1). Second, take only teachers in the treatment 1 group and randomly assign half to receive books (treatment 1 and treatment 2). Those not chosen to receive books will receive teacher training but no books (treatment 1 and control 2). Third, take the control 1 teacher group and randomly assign half of them to receive books (control 1 and treatment 2). Teachers not chosen to receive books will not receive training and therefore will belong to a pure control group (control 1 and control 2). The creation of groups requires an increase in the sample size and therefore has an implication on evaluation costs. Usually the choice of groups is based on the amount of resources allocated to a given component.

Each group would allow for answering how students respond on average to a given provision of program components.

Table 4: **Cross-cutting Design**

	Books (B)	**No Books (NB)**	**Difference**
Teacher trained (A)	E[Y|A,B]	E[Y|A,NB]	Effect of books for students with trained teachers
Teacher not trained (NA)	E[Y|NA,B]	E[Y|NA,NB]	Effect of books for students without trained teachers
Difference	Effect of teacher training for students with books	Effect of teacher training for students without books	
Both components versus no program	E[Y|A,B]-E[Y|NA,NB]		

Source: Prepared by the author.

2.3.2 Account for Possible Contamination - Spillovers

Spillovers are typically viewed as effects that control units receive from the treated units. In our example, it may be that trained teachers share tools they find useful with other teachers during lunch, so non-trained teachers may adopt these tools as well. Therefore, if we use students from other groups within the school as comparison groups, these groups would be contaminated. We would not be assessing the effects of the program, as we would be comparing against students who are indirectly benefiting from the program. Miguel and Kremer (2004) propose creating three different groups:

- Group P: Pure treatment
- Group S: Spillover or contamination group (schools without a program in communities with treated schools)
- Group C: Control (schools in communities where no school receives treatment).

Therefore, the spillover effects can be calculated as $E[Y|S] - E[Y|C]$ and the program effects can be calculated as $E[Y|P] - E[Y|C]$.To create a pure control group, randomization needs to take place at a level such that spillovers are unlikely. There are some other interventions that are specific to one group. In these cases spillovers are not a concern and randomization can take place across groups. Gertler, Patrinos, and Rubio-Codina (2007) propose randomizing at the community level instead of the school level so as to estimate spillover effects. A second alternative is to

Randomize at the community level instead of at the school level so as to estimate spillover effects.

vary the intensity of treatment, therefore assessing the differential externalities, if present. A third alternative is to exploit exposure across groups that is naturally present. For example, data can be collected on friends of treatment individuals about changes in key behavior. A fourth alternative is to move away from randomization and structurally model interactions. If all individuals in the experiment are equally contaminated or if contamination also flows from control to treatment groups, then evaluation is problematic with any methodology (Duflo, Glennerster, and Kremer, 2008). Note that the level at which randomization is done depends on program characteristics and context. Therefore, it is important to formulate the following questions:

1. At what minimum level can treatment be differentiated?
2. Would there be contamination?
3. Who is likely to be contaminated?

2.4 Establish WHEN: Timing

The WHEN question seeks to determine the timing of the evaluation. Ask yourself: *When will the program deliver the results for which it was designed?* Effects are likely not to be constant in time. On the one hand, if the evaluation is conducted too soon, students may not have benefited for long enough to report tangible benefits from the program. On the other hand, they may forget if much time passes after the intervention. Weber (2010) states that 70% of gains in a given year are retained in the long run. For example, Banerjee and Somanathan (2007) find that a remedial education program with tutoring increased test scores by 0.28 standard deviations and a computer-assisted learning program increased math scores by 0.47 standard deviations in one school year. One year after the programs were over, initial gains faded to about 0.10 standard deviations. On the other hand, a program such as school-based management may take a few years to show results on the quality of education. In this case, costs related to changes in organization and dynamics may even bring a negative effect in the first year (Gertler, Patrinos, and Rubio-Codina, 2007). Chay, McEwan, and Urquiola (2005) find that a program that provided infrastructure, materials, teacher training, and tutoring did not improve test scores in the first year but improved test scores by 0.2 standard deviations by the second year. Students forget material over vacations and therefore long breaks between interventions and data collection should be avoided (Muralidharan and Sundararaman, 2011). In sum, it is important to time the evaluation based on when the program is expected to bring about changes. Schedule data collection accordingly. If data are collected too soon or too late, effects may not be found (Gertler et al., 2011). If it is relevant to estimate how effects evolve over time, plan on collecting data at several points in time.

Time the evaluation based on when the program is expected to bring about changes.

Keep in mind that during the time it operates, the program affects the outcomes of individuals that do not involve their behavior. For example, a teacher may not participate in another training program in order to attend the one being evaluated. Card, Ibarraran, and Villa (2011) emphasize the importance of distinguishing these "within-program" effects from any impact of the program on post-program outcomes[4]. Therefore, information on the participants is necessary after program completion in order to differentiate between these two types of outcomes. In our example, outcomes should be collected after the teacher training period is over in order to differentiate within-program effects.

2.5 Establish WHERE: The Sample

The WHERE question aims to determine from which groups the evaluation will draw its findings. Ask yourself: *What is the population that I want to describe?* WHERE does not necessarily refer to a place. In this context, it refers to the group relevant to the evaluation. The WHERE question is important because it determines the external validity of the evaluation. An evaluation is **externally valid** if it represents the eligible population of interest. A first step toward answering the WHERE question would be to establish the eligibility criteria for the program. Make this list as clear as possible, including age range, education levels, geographical extension, and any other restrictions the program may have. An evaluation should draw from the population to which the program could expand. Ideally, take a random selection of individuals from the eligible population. The number of individuals required is discussed in Section 3. Note that in an impact evaluation, the sample size is determined by the minimum number of observations needed to detect the minimum relevant effect. There is a cost and accuracy trade- off: more observations for a given group decrease the error with which the average outcome is calculated, but they also make the evaluation more costly. The relevant minimum effect to be determined is usually defined as that found to be obtained by a second-best alternative use of resources. The idea is to test if the program is at least as good as the second-best option. When no information is available on alternatives, review effects of similar programs in similar contexts in the literature.

Take a random sample of individuals from the eligible population.

When testing differences between two means, it is usually assumed that the underlying data be generated by a process that results in a normal distribution. Therefore the population size is irrelevant (or intuitively, assumed infinite). Under this assumption, the standard errors are determined by the sample size but the population size is not relevant to the calculations: $SE = \sqrt{\frac{1}{n}}\sigma$, where σ is the standard deviation of the population. In some cases it may be possible to estimate effects based on the entire population. Whether a sample or the whole population is

[4]In labor economics, **Ashnefelter's dip** refers to the temporal depression of labor market outcomes below their long-run or "permanent" levels as a result of program selection.

observed, the standard error provides information on the variation of the outcome variable.

When sampling students within schools, it is common to make a random selection of teachers and then choose a random sample of students of fixed size within the class. For example, suppose the plan is to observe 10 students per class in 200 schools for evaluation purposes, regardless of class size. If school size varies significantly across schools and data are collected on a fixed sample size by class of students, school size should be taken into account to weight student outcomes accordingly to calculate averages. An alternative is to control for school size when calculating averages (Duflo, Glennerster, and Kremer, 2008). Sample size is usually determined in order to calculate a given statistic such as the mean in other contexts. In this case, the standard error should improve when the number of observations increases. Also in this case, the number of observations in the universe plays a role. On the other hand, the standard deviation is a measure of the variation of the data in the sample. This should not change as the sample size increases if the sample is a random draw. Assume a small sample

Calculate the minimum number of observations needed to detect the differences found employing the best alternative use of resources.

$$SE = \sqrt{\frac{N-n}{n(N-1)}}\sigma, \tag{6}$$

Where SE denotes standard error, N is the population size, n is the sample size, and σ is the standard deviation of the population. For the case of Mexico, the average class size is 20. If a sample of five students is taken, then the estimate of the mean of that class based on those five observations has a standard error of $\sqrt{(20-5)/(5(20-1))}\sigma \approx 0.4\sigma$.

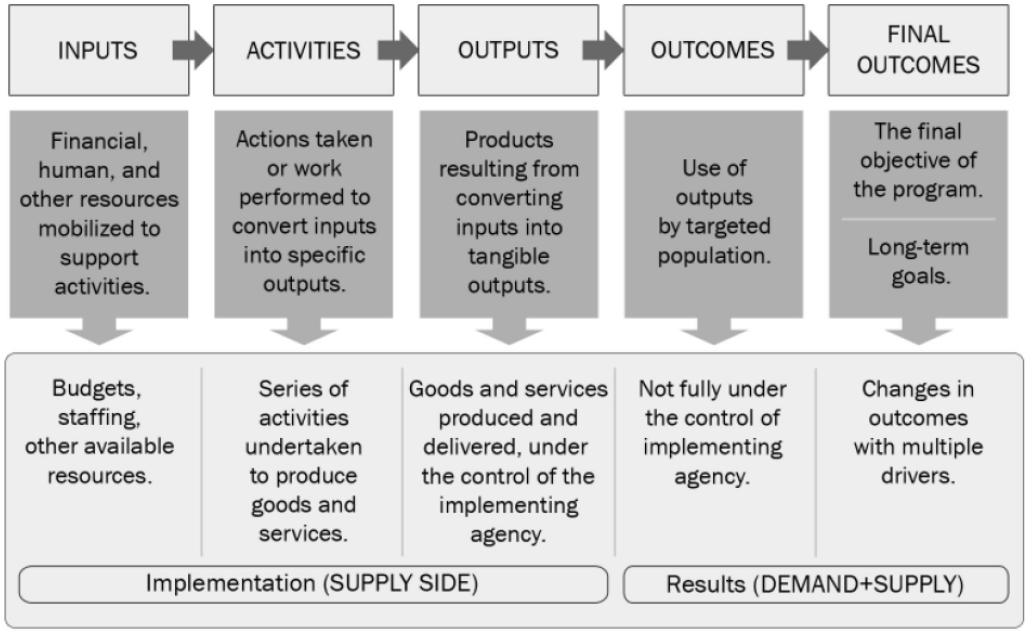
2.6 Establish HOW: The Causal Chain

The HOW question should help to better understand how the program leads to results. Ask yourself: *What is the underlying process that transforms inputs into results?* The basic theory of how the program is expected to lead to results should be established. The main tool proposed in impact evaluation is to create a causal chain that links inputs to final outcomes and allows for an explicit statement of the expected changes. Figure 2illustrates this idea. The chain has the following elements:

A causal chain links inputs to final outcomes and allows for an explicit statement of expected changes.

1. *Inputs*: Resources at the disposal of the project, including staff and budget.
2. *Activities*: Actions taken or work performed to convert inputs into outputs.
3. *Outputs*: The goods and services produced by project activities.
4. *Outcomes*: Results likely to be achieved once the beneficiary population uses the project outputs (usually in the short to medium term).

Figure 2: **Results Chain**



Source: Gertler et al. (2011)

5. *Final outcomes*: The project's final goals.

Inputs, activities, and output are in the control of the implementing agency, while the outcomes and final outcomes also depend on demand-side factors. (For a more detailed description, see Gertler et al. (2011)).The elements in the causal chain are indicators and should be based on the recommendations included in Section 4.

The construction of indices should be avoided when possible. Sometimes it is desirable to summarize the information of several variables into one. For example, suppose there is a set of questions regarding student attitudes toward education. A summary of the information in a single dimension is not ideal because the process implies reducing the amount of information available in the analysis. Therefore, information that may reveal specific aspects of program effects in the analysis should be included separately rather than put into an index. In some cases, the construction of indices may be unavoidable. There is no generalized method to construct indices, which are frequently based on assumptions behind the variables observed. Indices are usually a weighted average of a set of variables. For example, test scores are frequently the addition of right answers to a set of questions.

Include information that may reveal specific aspects of program effects in the analysis rather than putting them into an index.

A data-driven approach to the construction of an index is to use factor analysis. The idea is to find the dimensions over which most variation takes place[5]. The

[5]An intuitive way to think about this is to assume you want to reduce information from three

problem is that there is usually more than one combination of factor weights that reflect the same variation. Appendix 6 includes details on index construction with factor analysis. The key idea behind the construction of an index is to have a clear idea of what variation it attempts to capture. Remember that it may be more useful to have one variable that is relevant than a noisy index that is difficult to interpret.

2.7 Example of an Evaluation

Consider the evaluation of a teacher training program in the state of Puebla in Mexico. The program includes an investment in administrative capacity at the state level. Here is an example of such a hypothetical program and the answers associated with the six five W's:

1. WHY: Suppose the main question for the implementing agency is whether or not to expand the program. If the evaluation shows the program provides benefits, the agency would like to have information on how to improve it.

2. WHO: Suppose the goal of the program is improve student knowledge. The main question regards the distribution of students (not teachers or schools)

3. WHAT: Suppose the program provides teacher training and randomization can be implemented at this level. The main question regards changes at the teacher level, not the student level. It is not possible to include possible changes to administrative capacity that may accompany the teacher training program, as all teachers (treatment and control) would benefit from this feature.

4. WHEN: Suppose teacher training implements pedagogical changes expected to be brought about during the school year. The main question regards effects one year after the program starts.

5. WHERE: Suppose the program is implemented in the state of Puebla. Inferences about states other than Puebla or a specific municipality within the state would be out of the scope of the evaluation. Inferences about other populations involve planning and secondary analysis.

6. HOW: A goal of the evaluation should be to better understand how the program works and make relevant decisions to improve its efficacy and efficiency. Assume investment is US$200 per teacher (inputs). This investment leads to 100 hours of teacher training (activities). As a result the program delivers trained teachers (outputs). Teachers change classroom behavior measured

dimensions (you) to two dimensions (the wall). Factor analysis finds the shadow such that it can better predict your three dimensional shape with linear combinations (just rescaling and rotating placement of points on the wall).

by effective teaching time with Stallings (outcomes). These changes should improve mathematics test scores (final outcomes).

The information provided by the 6 W's define the following main impact evaluation question: What is the effect of teacher training on the mathematics standardized test scores of students after one year in the state of Puebla?

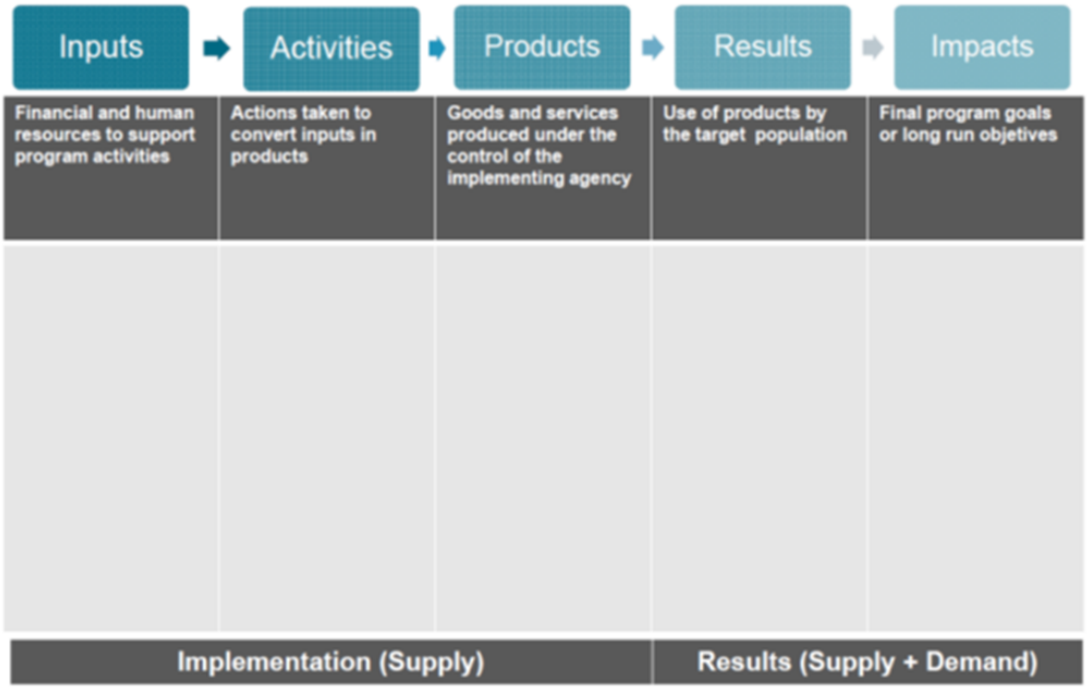
2.8 Checklist and Diagram to Define Evaluation Questions

Table 5 provides a checklist to define the evaluation questions, taking into account key program and population characteristics as discussed in this section. A blank causal chain is also provided in figure 3 to support the establishment of a theory to link inputs to final outcomes of the program.

Table 5: **Checklist for the six W's : Check Each Item for Which there Is an Answer**

Category	**Questions**	**Check mark**
WHY	Why should this program be evaluated? How would evaluation results be used?	
WHO	Who are the individuals or groups toward whom the intervention is targeted? What information is available and/or feasible to collect?	
WHAT	What program components could be differentiated? At what minimum level of differentiated treatment can the program be implemented? How many groups would it be possible to create? Would contamination be a concern?	
WHEN	When is the program expected to bring results?	
WHERE	What is the population from which it is relevant to draw inferences? What are the eligibility criteria?	
HOW	How will the program bring about changes? What is the theory of change behind the program?	
Main research question:		

Figure 3: **Theory of Change**



Source: Gertler et al. (2011)
Note: The grey area is intended to be used for notes by the evaluator.

3 STEP 3: Select the Sample and Create Treatment and Control Groups

Random assignment provides an ideal setting to identify the effects of a program. By randomly selecting individuals or schools and placing them into treatment or control groups, we rule out the possibility that differences in other characteristics that we do not observe will influence one group or the other. Therefore, differences after treatment between the two groups can be attributed exclusively to the program.

3.1 Determine How Many Observations Are Needed: Power Calculations

This section summarizes the main formulas and rationale for power calculations in Duflo, Glennerster, and Kremer (2008). As an example, suppose we want to evaluate the effect of a teacher training program on test scores and we can randomize at the teacher level. We would estimate the effect as:

$$\tau = E[Y|T] - E[Y|C], \tag{7}$$

Where Y denotes test scores, T denotes the treatment status, C denotes control status, and τ denotes the program effect. Since we randomly allocated teachers to treatment group T and control group C, we are confident that the treatment group is statistically equivalent to the control group before the intervention, and would have behaved as the control group in the absence in the program. Another way to estimate this is:

$$Y_j = \alpha + \beta DT_j + \epsilon_j, \tag{8}$$

Where Y denotes the test score for teacher j and DT is a dummy that equals 1 if the teacher was assigned to the treatment group and 0 otherwise. ϵ is an error term. If we estimate this equation by ordinary least squares (OLS), the coefficient beta will capture the average difference after the program between treatments and controls, that is, the average impact. Duflo, Glennerster, and Kremer (2008) state the formula for the minimum detectable effect (MDE) as

$$MDE = (t_{(1-\kappa)} + t_{\alpha/2})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}, \tag{9}$$

for a two-sided test, where $P = T/N$ is the proportion of teachers treated, T denotes the number of teachers treated, and C denotes the number of teachers in the nontreated group and $N = T + C$ is the total number of observations. For a power κ of 0.80, $t_{1-\kappa} = 0.84$ and a significance level α of 0.05, then $t_{\alpha/2} = 1.96$. Given that

in education it is common to refer to effects in terms of standard deviations we can state the MDE in terms of standard deviations as

$$MDE_{SD} = (2.8)(T+C)\sqrt{\frac{1}{TC(T+C)}}, \tag{10}$$

Where MDE_{SD} is $\frac{MDE}{\sigma}$. At the student level, power calculations must take into account the fact that observations are clustered at the classroom level. Assume the effect of the program is calculated with the following estimating equation:

$$Y_{ij} = \nu_j + \beta DT_{ij} + \epsilon_{ij}, \tag{11}$$

Power calculations must take into account the fact that students will be clustered.

Where i denotes a student and j denotes a classroom. ν_j denotes a classroom fixed effect. In this case, the formula for the MDE is

$$MDE = (t_{(1-\kappa)} + t_{\alpha/2})\sqrt{\frac{1}{P(1-P)J}}\sqrt{\rho + \frac{1-\rho}{n}}\sigma, \tag{12}$$

Where J is the number of teachers and n is the number of students per teacher. ρ denotes the intra-cluster correlation (ICC). Taking the same power and significance as before, and writing in terms of standard deviations, we have

$$MDE_{SD} = (2.8)(T+C)\sqrt{\frac{1}{TCJ}}\sqrt{\rho + \frac{1-\rho}{n}}, \tag{13}$$

The parameter ρ changes according to the context. Hedges and Hedberg (2007) provide an example of an ICC calculation[6].

Sometimes there are no data available on what effect size to expect, and therefore sample size calculation becomes difficult. Duflo, Glennerster, and Kremer (2008) maintain that an effect of 0.2 standard deviations is "small," 0.5 is "medium," and 0.8 is "large." The authors also suggest calculating a range with intraclass correlation coefficients ranging from 0.2 to 0.6 when it is not possible to obtain a reliable estimate of the ICC[7]. Gertler, Patrinos, and Rubio-Codina (2007) state that a rule of thumb among educational researchers for a cluster randomized trial is 40 to 50 schools with 40 to 60 students, with the same number of schools in treatment and control groups to detect MDEs of between 0.10 to 0.25 standard deviations.

A rule of thumb among educational researchers for a cluster randomized trial is 40 to 50 schools with 40 to 60 students.

The number of units in the treatment and control groups do not need to be equal. When the number of observations of a given group increases, the error with which the group is measured decreases. The same number of observations in treatment and control groups is optimal in the absence of priors on the errors expected from both

[6]The ICC can be calculated in STATA by estimating xtreg y T, fe i(j). The output reports the ICC as rho.

[7]The author's own calculations for ICC in Mexico for several years and for different states according to their marginality levels are between 0.24 and 0.40, without a clear pattern in time or marginality status.

groups and no operative restrictions or differentiated costs. It is best to increase the sample size of groups for which observations across rounds of data collection are expected. Usually, control units are less likely to be willing to participate than treated units. Attrition is discussed in Section 5.1.3. For details on the derivation of the formula with different costs for data collection in treatment and control groups, see Duflo, Glennerster, and Kremer (2008).

Some teachers may have a class size smaller than the estimated size as determined by the procedure described in equation 13. In these cases data should be collected on all students in the class. Omission of small classes or schools biases impact estimates toward the effect in larger schools. Information on the number of students in the class should be collected to weight results for estimations related to the student-level distribution.

Two considerations should be present when considering carrying out impact evaluations with a small number of observations. First, the rule of thumb is that at least 30 observations are needed to invoke the law of large numbers associated with the assumption of normality for rejection of the null. Normality is required for consistent estimators. Second, power calculations based on 15 treatment and 15 controls would lead to a minimum detectable effect of one standard deviation, which is roughly equivalent to about one school year in the United States (Hill et al., 2007). Given that education interventions take time to bring large results, it is unlikely that effects will be found with such a sample size. On the other hand, there have been successful impact evaluations with a limited number of observations. For example, Banerjee and Somanathan (2007) include 55 schools in year 2 and 56 schools in year 3 to evaluate a computer-assisted learning component. They find effects of 0.42 and 0.27 standard deviations for students in the bottom and top tertiles, respectively. They also find that the effect fell after a year to 0.09 standard deviations. Other examples are Basinga et al. (2010), who used randomization at the district level on eight pairs for health outcomes, and Bloom et al. (2011), who used randomization with 20 experimental textile treatment plants. The factors listed in Table 6 should be taken into account when thinking about the minimum number of observations required to detect a given effect. Section 5 includes a discussion on techniques that can be applied to evaluations with randomization using a small group.

3.2 Assigning Treatment and Control Groups: Strata Randomization

Block or stratified randomization is based on the idea of creating groups that are as similar as possible and then randomizing within these groups. An example is the paper by Banerjee and Somanathan (2007) that looks at the impact of a program on test scores. The researchers stratified according to class size, language of instruction, school, gender, and pre-intervention test scores. Blocking improves precision to the extent the variables used for blocking explain the variation in test scores (Duflo,

Table 6: **Factors that Influence the Number of Observations Needed in an Evaluation**

Factor	Importance
Size of the effect expected	Determines the number of units for which differentiated treatment is needed. Smaller effects require more observations.
Intra-cluster correlation (ICC)	Determines the number of observations needed within a group (school). A higher ICC indicates that fewer observations within a group and more observations across groups are useful.
Timing of measurement	Timing is related to effect size (Banerjee and Somanathan, 2007)
Whether or not the population is subject to group (school) shocks	Shocks common to all students in a school usually introduce more variation in estimates than student-level shocks (Bloom et al., 2011).
Measurement error in the dependent variable	Measurement error bias estimates toward zero (Hyslop and Imbens, 2001)

Source: Prepared by the author.

Glennerster, and Kremer, 2008). To see why this would pose an advantage, imagine that within a district, only schools with low marginality levels are assigned to treatment; none of these schools are assigned to the control group. In this case, schools in this district would not contribute to the comparison of low versus high marginality levels. The education literature highlights the following variables (in order of relevance): pre-test scores, student socioeconomic status, and teachers (when applicable). But several other relevant variables on the program specific context should be considered. For example, in Mexico there are several modalities of elementary school, including general, indigenous, and community-based. The populations across these types of schools are different, as are the test scores. Therefore, one would introduce the modality of the school as a dimension across which to block. In practice, the number of groups created using continuous variables (like pre-test scores or socioeconomic status) versus introducing more dimensions to stratify depends on data availability. When the size of the resulting strata is different across groups, large strata can be further broken down into smaller strata to improve precision. Hyslop and Imbens (2001) suggest not creating strata smaller than four individuals so that the variance within each strata can be calculated. The key is not to change the probability of treatment for any given teacher. If, for example, half of the teachers in the sample will be trained, then all teachers must have one-half probability of being trained.

It is important to stratify across the dimensions over which it may be relevant to contrast results. Examples of dimensions to build strata include initial level of

education (Glewwe, Kremer, and Moulin, 2009), socioeconomic status (Paxson and Schady, 2007), and gender (Hastings, Kane, and Staiger, 2006; Roland G. Fryer and Levitt, 2010). Stratification across dimensions of special interest has two advantages. First, it increases the power to contrast effects across groups. Second, it allows for explicitly determining ex ante the relevant dimensions to be explored and the design of those dimensions. It also heads off subsequent doubts about data mining. To summarize, randomization can be implemented using the following steps:

1. Identify key dimensions across which to randomize. Take into account data availability. Use dimensions deemed relevant by the literature: pre-test scores (Banerjee and Somanathan, 2007), socioeconomic status (Mizala, Romaguera, and Urquiola, 2007) and school characteristics (state, modality and/or level).

2. Decide on block size and create blocks. Blocks should not include fewer than four observations. If there are blocks with more than four observations, continue to break down a given block by separating the main continuous variable into two groups (usually pre-test scores or socioeconomic status above and below the median) within the block. Continue this process within blocks and stop when a given block has between four and seven observations. If the size is smaller than four, go up one level and group. If the size is eight or larger, break it down.

3. Create a random number for each individual. Note that Stata and Excel create quasi-random numbers[8]. The webpage www.random.org can be used to create random numbers but there are many web pages to create random draws.

4. Calculate the mean of the random numbers of the individuals within the strata. If the random number of the individual is above the mean of the strata, then the individual is assigned to treatment, otherwise to control. If half of the population is not assigned to treatment within a block, then, instead of the mean, generate the statistic that would allow for separating the group in the desired proportion, such as tertiles, quartiles, etc. Adjustments to block size may be unavoidable in some cases.

5. Ensure that the randomization was successful by checking that the two groups are balanced (statistically equivalent on average values) in the main characteristics at baseline (prior to the intervention).

Implementation of stratified randomization depends on data availability and operative considerations. In many countries census data are available and some proxy information can be used to determine the socioeconomic status of a given school.

[8]Quasi-random numbers differ from random numbers in that the draw uses information on previous results. Quasi-random numbers comply with a low discrepancy requirement. As a result, it is less likely to get more points in a certain range with repeated quasi-random draws. For a formal approach see Morokoff and Caflisch (1994).

3.3 Implementation

After the design of evaluation, documentation and coordination follow. An impact evaluation implies collaboration among many individuals, and it is always useful to have a document to guide everyone. Guidelines for evaluation are especially useful when many states or implementing units are involved. Guidelines usually involve an introduction with the main rationale of the evaluation, goals and limitations, and information on how implementation should differentiate between treatment and control and who should receive differentiated treatment. Guidelines also include activities such as data collection that require permission from and even direct participation by the implementing units. The document should include sections on how to register and substitute dropouts and how to deal with changes, when applicable. Guidelines should also include any information required from implementing units, including the format and time of delivery.

Guidelines are useful to coordinate evaluation efforts when multiple implementing units are involved.

A second useful document is one that enables communication with participants regarding the evaluation when it is likely that the participants will find out about the evaluation and/or differentiated treatment. Appendix C presents an example of a draft. Evaluation participants should ideally be blind to the evaluation. This means that to the extent possible participants should not know who is in the treated or control groups. The rationale behind keeping participants unaware is to avoid what are called "Hawthorne" and "John Henry" effects. According to the Hawthorne effect, school constituents may change their behavior because they know they are (or are not) participating in a program. Duflo, Glennerster, and Kremer (2008) demonstrate this effect using a treatment status elevating the morale of beneficiaries and a control status making individuals in the comparison group feel offended. School constituents may also change their behavior because they participate in an experiment, but not necessarily in the program. The resulting change is the John Henry effect. For example, teachers may change their classroom dynamics because they know they are being observed and are part of an evaluation. Duflo, Glennerster, and Kremer (2008) propose looking at long-run effects of programs to check whether short-run effects are likely to be a result of changes in behavior not related to the program. Those who implement and administer the program should be blinded when possible. Ideally, the person who analyzes the results should also be blinded so as to avoid any bias in the assessment.

The rationale behind keeping participants unaware is to avoid changes in behavior not related to the program.

3.4 Checklist for Power Calculations and Implementation of the Evaluation

The checklist in Table 7 should be completed to ensure inclusion of key considerations in calculating the sample size and creating treatment and control groups.

Table 7: **Sample and Randomization Checklist**

Sample	What is the number of potential beneficiaries?	Universe
	How many potential beneficiaries will participate in the evaluation?	Sample size N
	If grouped errors, what is the intra cluster correlation used for power calculation?	ρ
	What is the number of units within groups to be sampled?	n
Randomization	What is the level at which randomization is feasible? (School, teacher, student, etc.)	Level
	Is the randomization strategy described? (Simple, block, etc.)	Yes/No
	Is the creation of blocks documented?	Yes/No
	Are the variables used to block explicitly listed?	Yes/No
	Is the creation of the random sequence documented?	Yes/No
	What is the minimum detectable effect (MDE) of the main outcome variable? (Student or unit of observation level)	MDE
	What is the MDE of secondary variables of interest? (Teacher or groups)	MDE
Implementation	Is it possible to blind the experiment?	Yes/No
	Is there a document to coordinate evaluation efforts? (Concept note)	Yes/No
	Is there a need to explain the evaluation to the participants of the program?	Yes/No
	Are there guidelines for implementing units? (If applicable)?	Yes/No

Source: Prepared by the author.

4 STEP 4: Collect Information and Data

This section focuses on which data should be collected. The first subsection discusses how standardized test scores differ and lists some popular tests for the primary level. It focuses on standardized test scores because the guidelines presented here are designed for impact evaluation of programs that aim to improve the quality of education, which is usually proxied or defined by some standardized test. The second subsection discusses some variables that evidence shows can explain variations in test scores. Collecting data on this set of variables may improve estimation efficiency. The third subsection discusses timing, and the final subsection provides some references on the topic and useful links on how to document the evaluation. A checklist is provided at the end of the section. Operative aspects of data collection are not included in this guide. A complete reference on how to collect data-including how to hire the collection firm, format questionnaires, conduct fieldwork, and process and validate data-can be found in Gertler et al. (2011). The data collection process should aim to decrease errors when measuring a set of outcomes and collecting relevant information.

4.1 Collect Data on Test Scores

Thisdocument focuses on evaluations where test scores are the main outcome. Several considerations should be taken into account when deciding the best strategy to choose the instruments to measure education quality.

Standardized tests are designed to measure different aspects of learning.

First, be aware that standardized tests are designed to measure different aspects of learning. For example, some tests may be designed to measure how much of a given curriculum has been covered, as in the case of the national achievement tests (ENLACE) in Mexico. Other tests, like the Program for International Student Assessment (PISA) for 15-year olds, have been designed to measure what students can do with what they have learned. Some tests are designed to include mechanical and conceptual questions to better understand the effects of a given intervention. Others are diagnostic tests, such as the Early Grade Reading Assessment and Early Grade Math Assessment[9].Examples of instruments designed to measure test scores include standardized tests in Argentina (*Operativo Nacional de Evaluación* - ONE), Brazil (*Prova Brasil*), Chile (*Sistema de Medición de la Calidad de la Educacion* - SIMCE), Colombia (SABER), Mexico (ENLACE and *Exámenes de la Calidad y el Logro Ed-*

[9]These tests were developed in 2006 through a joint international effort by the Research Triangle Institute, the U.S. Agency for International Development, and the World Bank to provide an instrument to assess knowledge. The tests are designed to be applied to children in grades 1 to 3 at the end of the school year. For more information, go to http://toolkit.ineesite.org/toolkit/INEEcms/uploads/1038/Early_Grade_Reading_Assessment_Toolkit_EN.pdf and http://www.google.com/search?sourceid=ie7&q=EGMA+a+conceptual&rls=com.microsoft:en-us:IE-SearchBox&ie=UTF-8&oe=UTF-8&rlz=1I7SNNT_enUS344US344

ucativos - EXALE), and Peru (*Evaluación Censal de Estudiantes* - ECE), as well as the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS) administered internationally by the International Association for the Evaluation of Education Achievement, and the *Segundo Estudio Regional Comparativo y Explicativo* (SERCE) administered internationally by the United Nations Educational, Scientific and Cultural Organization (UNESCO). Some publishers offer standardized tests such as the Test of Early Mathematics Ability (TEMA)[10].

Second, developing a test and collecting data are costly and therefore it is worth considering the use of local or national tests when they are available. Many countries now have standardized test scores that can be used for evaluation.

Third, written tests are usually applicable to children in a small grade range and with specific characteristics. In some contexts where the educational level is very low, children may not be able to read instructions, and verbal application of tests may be necessary.

Fourth, care needs to be taken with the use of teacher-specific test scores and failure rates as the only instrument to measure education quality. This measure is positively correlated with test scores, but it is unlikely that all students are compared against a common scale. Teachers may tend to normalize test scores. For example, a grade of 7 out of 10 in a school where students tend to perform very well is unlikely to be equivalent to 7 out of 10 in a school where most students tend to perform poorly.

Fifth, potential sources of measurement error need to be taken into account and avoided. Muralidharan and Sundararaman (2011) test students twice in each round to include more material, reduce the impact of measurement errors specific to the day of the test, and reduce sample attrition due to student absence on the day of the test. The authors report on the use of repeated questions across rounds, as improvements may be a result of students remembering questions.

Finally, testing should be externally monitored. Cheating has been well documented both in the United States and other countries. This is a particular concern in programs with performance pay schemes.

To summarize, it is necessary to determine whether developing a test is necessary, choose a test that measures the outcome you intend to capture, ensure that the test is applicable to the specific student group, ensure that all students are measured by a common metric, and consider monitoring when collecting data.

[10] more information at http://www.psych-edpublications.com/arithmetic.htm#tema

4.2 Collect Information on Dimensions that Explain Variations in Test Scores

As discussed in Section 1, there are a number of factors that influence education. Information on these factors can facilitate the treatment of calculations among different groups or under specific circumstances. For the average treatment effect based on all individuals participating in the evaluation, no information on other factors that may affect test scores is necessary. In a randomized evaluation, we assume that there are no differences in observable or unobservable characteristics in the absence of the program.. As a result, program effects can be assessed by a simple comparison of means. There is no need to control for observable characteristics. Program effects can be estimated with a smaller standard error if data are collected on determinants of test score variation and introduced as controls to estimate differences in averages using OLS.

Duflo, Glennerster, and Kremer (2008) draw attention to cautious use of controls with OLS estimation. First, once you control for a covariate that is affected by treatment, then the covariate will capture some of the impact. Second, including variables that explain little variation will increase the standard error. Third, covariates make the estimated variance of the effect noisier. Fourth, the inclusion of covariates imposes the additional assumptions associated with OLS, including linearity. Section 1.3, referring to some evidence found in the United States, discusses why test scores are unlikely to be a linear function of initial test scores when more than one cohort is considered. A fifth reason to be careful is interpretation. Todd and Wolpin (2003) posit that it is not always ideal to include proxy variables because interpretation becomes difficult. For example, suppose the goal of the evaluation is to assess the effect of books on student test scores, and family income is used to proxy for other inputs to education. If family income is constant, then providing books to children may imply the household may be increasing expenditure on other items. If these items are other inputs to education, then the coefficient of books on OLS confounds the effect of more books with the increase in other inputs. For non-randomized evaluations, Todd and Wolpin (2003) state that the problem of including proxy variables cannot be solved because it involves a comparison between two unknown biases. In summary, there are several effects that result from including covariates:

1. Covariates reduce residual variance and therefore reduce variance of parameters.

2. Covariates may increase the variance of the estimate of the effect. This effect is not present with stratification because it makes treatment orthogonal to other covariates.

3. The estimated variance of the effect is noisier when controlling for covariates

than when not controlling for them, and it is unbiased.

4. The estimates adopt the assumptions for OLS.

5. Interpretation of the estimated effect changes.

The first point requires knowledge about the factors that explain test score variation. Section 1 provides a review of the literature. Based on this review, information on pre-intervention test scores, socioeconomic status (such as education of parents and income), and teacher and school characteristics should be relevant. Administrative data may be available and should be used for calculation. There are multiple household, teacher, parent, principal, and school surveys that can help build a specific evaluation instrument. Table 8 provides links to surveys that can be useful to adapt to the evaluation.

Collect data on education of parents, household income, and pre-intervention test scores.

A relevant piece of information to collect when evaluating a program aiming to effect changes on the supply side is classroom behavior. Teachers are likely the most important factor at the school level that influences student test scores. Several instruments have been developed to assess some of the behavior of teachers in the classroom:

Teachers are likely the most important factor at the school level that influences student test scores.

1. CLASS http://www.teachstone.org/about-the-class

2. Framework for Teaching: http://www.danielsongroup.org

3. Stallings: http://www.eddataglobal.org/embedded/stallings_snapshot.doc

4. TIMSS 1999 video study: http://timssvideo.com/

For an assessment of before- and after-school care, consider using the School Age Care Environment Rating Scale (SACERS) available at http://ers.fpg.unc.edu/ (Gertler et al., 2011). It provides information on the conditions necessary for the information to be useful for impact evaluation. Indicators for any aspect for which information is collected should be specific, measurable, attributable, realistic, and targeted, or SMART:

1. **S**pecific - Indicators must specifically address a particular skill to be tested. Be aware that test scores may reflect very different abilities. For example, TIMSS tries to assess "what students know," while PISA tries to assess "what students can do with what they know."

2. **M**easurable - There must be a common scale against which all students are compared. For example, creativity or leadership may have multiple dimensions that cannot be easily captured in a few characteristics. When individuals play a role in assigning a score, data collection teams must be trained so as to

Table 8: **Links to Surveys**

Student	ENLACE	http://www.enlace.sep.gob.mx/ba/
	PIRLS	http://timss.bc.edu/pirls2011/international-contextual-q.html
	PISA	http://pisa2009.acer.edu.au/downloads.php
	TIMSS	http://timss.bc.edu/TIMSS2007/context.html
Parent	PISA	http://pisa2009.acer.edu.au/downloads.php
Household	MXFLS	http://www.ennvih-mxfls.org/
	PIRLS	http://timss.bc.edu/pirls2011/international-contextual-q.html
Teacher	ENLACE	http://www.enlace.sep.gob.mx/ba/
	PIRLS	http://timss.bc.edu/pirls2011/international-contextual-q.html
	TIMSS	http://timss.bc.edu/TIMSS2007/context.html
Principal	ENLACE	http://www.enlace.sep.gob.mx/ba/
	MXFLS (Mexican Family Life Survey)	http://www.ennvih-mxfls.org/
	PIRLS	http://timss.bc.edu/pirls2011/international-contextual-q.html
	PISA	http://pisa2009.acer.edu.au/downloads.php
	TIMSS	http://timss.bc.edu/TIMSS2007/context.html
Authorities	PIRLS	http://timss.bc.edu/pirls2011/international-contextual-q.html

Source: Prepared by the author.

homogenize the way in which units of observation are graded. For example, in an oral exam an individual applier may be prone to help the student by providing extra motivation or reading the instructions once again, while a second applier may not do so. It is likely the student with the first applier would get a higher score.

3. **A**ttributable - The index must reflect the changes we are expecting out of the intervention. For example, if we are introducing mathematics software, it would make sense to apply a mathematics test that includes the concepts that the software emphasizes.

4. **R**ealistic - Tests must be administered in a relatively short time. A student may not be able to concentrate for more than an hour at a time. Students in elementary school with poor performance may not even be able to understand a written test. In this case, it is unrealistic to capture information without verbal support from the data collection team.

5. **T**argeted - Tests must be designed for the appropriate age group and context.

The bottom line of indicators is credibility. Choose impact indicators such that any noise introduced by other factors is reduced to a minimum.

To summarize this subsection, consider including observable characteristics that explain variation in test scores, but be careful with measurement and interpretation. If controls are included, consider collecting data on education of parents, household income, initial test scores, and teachers. Make sure your indicators are SMART.

4.3 Other Relevant Information

Data collection should include information that uniquely identifies and locates each unit of observation. Student, teacher, and school information should all include a unique identifier. Collect the following information to guarantee that the necessary information is collected to carry out the analysis:

1. Include a unique identifier for each student, teacher and school. Do not rely on names, birth dates, and location. These variables are prone to errors. The identifier should include other useful pieces of information. For example, in Mexico the school identifier is two numbers that identify the state, one letter with the financing source, two letters for school modality, four spaces with progressive numbering, and a verifier letter.

2. Collect re-contact information. This information allows for tracking students and dealing with dropouts and late enrollments to the sample that may lead to sample bias.

3. Collect location data. This information ensures that the program is being implemented where it is supposed to be and the information is being collected where it is supposed to be collected. When implementation takes place at several sites, this information allows for separation of the effects of the site from those of the program. Global Positioning System technology is helpful to identify remote locations and locate participants in the future.

4. Collect program monitoring data. This will help ensure equal fractions of treatment and control individuals when program implementation takes place at several sites. These data provide information on the extent of cross-overs and are of special interest regarding implementation characteristics. The information is also very helpful to explain effects or the absence of effects.

5. Collect follow-up data on as many individuals included in the original design as possible, regardless of whether they complied with their program treatment assignments. This information allows for accounting for cross-overs that lead to selection bias.

6. For those individuals not present in the original design, document any available information on why they are not there. Although replacements are not desirable, if replacements are possible at no cost, randomly choose from those present. Resources should prioritize finding dropout reasons over finding replacements. For example, if randomization is done to evaluate a teacher training program with 200 teachers before the school year begins, but 10 teachers do not show up when the school year starts, it is necessary to find out why they did not show up and replace them by randomly selecting among those present.

7. In some contexts it may be relevant to determine whether the program increases school access. This can be measured by enrollment or attendance rates. Not taking these effects into account may bias results. Gertler, Patrinos, and Rubio-Codina (2007) suggest analyzing dropout rates, as these may be an indicator of how successful the school is at keeping students. Dropout rates may also indicate how parents perceive the quality of education. The proportion of over-age students is also of interest because those students are more likely to drop out because their opportunity cost is higher and resources allocated to students in normative age in school must be shared with over-age students.

8. Visit a few treatment and control units. Visits enhance understanding of actual intervention and provide important clues to interpret program effects.

4.4 Timing

There are usually at least two rounds of data in impact evaluation:

1. Baseline. The role of the baseline is to:
    - Collect pre-intervention test scores. This information allows for testing for heterogeneous effects and reducing standard errors
    - Test for pre-intervention differences and validate the experimental design.
2. Follow-up. The role of the follow-up is to calculate program effects.

More rounds of data allow for the calculation of dynamic effects. Panel data are preferred in these cases. A discussion on timing is included in Section 2.4. In practice, the baseline takes place early in the school year and the first follow-up at the end of the first school year. This design allows for a higher likelihood of obtaining a panel. Follow-up data are usually collected at the end of the school year because national curricula are general designed to complete established learning goals by then.

4.5 Document the Evaluation Process

A part of the evaluation often neglected is documentation. Documentation should be the minimum necessary that allows for full replication of findings. Table 9 presents a data collection checklist to ensure that critical documentation is included in the evaluation report. The evaluation process should document where and how data were collected, including the following information:

Documentation should be the minimum necessary that allows for full replication of findings.

1. Why, who, what, when, where, and how (see Section 2). Remember to document changes in implementation that may occur during the evaluation period.
2. Definition of the sample and trial design (see Section 3). Document changes in both sample and treatment status that may occur during the evaluation period. The relevance of these changes is discussed in Section 5.
3. How and where data were collected. Include any changes and justify them (covered in this section).
4. Details on the analysis of the data and adjustments (discussed in Section 5).

4.6 Data Collection Checklist

The checklist in Table 9 should be completed to ensure that all of the elements of data collection are included in the evaluation design and procedures.

Table 9: **Checklist for Data Collection**

Data Collection Item	**Included?**
Is there a strategy to identify individuals uniquely?	Yes/No
Is there a strategy to link students, teachers, and schools?	Yes/No
Re-contact information?	Yes/No
Are ALL indicators of interest included? (Defined in Figure 3)	Yes/No
Are ALL indicators SMART? (See Subsection 4.2)	Yes/No
Are location data included?	Yes/No
Is there a strategy to obtain monitoring data?	Yes/No
Has information on socioeconomic status been collected?	Yes/No
Has information on individual characteristics (age, gender, education) been collected?	Yes/No
Is there a strategy to document the evaluation process?	Yes/No
Did you double check that data on program goal indicators have been collected?	Yes/No

Source: Prepared by the author.

5 STEP 5: Analyze Data

This section provides guidance on how to calculate program effects and the corresponding confidence intervals. The first subsection discusses the methodology related to estimation of program effects. It starts by discussing how to correctly weight observations, followed by a discussion of the analysis of groups of interest, how to interpret program effects, and what to do when actual implementation is different from program evaluation design. The second subsection discusses the methodology related to the estimation of confidence intervals.

This section will use the example of an evaluation of the effects of teacher training on standardized student test scores. To simplify, suppose that teacher training is designed for sixth grade teachers and that randomization was done at the school level. Once a teacher is trained, all students benefit. Assume one group per school. This example will allow for illustrating how specific program design features play a role in data analysis.

5.1 How to Calculate Program Effects

The average treatment effect of the program can be calculated as the difference between the expected value of the outcome of interest in treatment group T and control group C. The average treatment effect is defined in equation 5 in Section 1.4. The average should correspond to the population the evaluation seeks to provide information about under each treatment scenario. Sampling and deviations from sampling should be taken into account to calculate program effects. For example, the most common case in education is to take a random sample of students of constant size across schools of different sizes. If weights are ignored, estimates may correspond to an effect on the school distribution of test scores as opposed to the student distribution of test scores. In this case, students in small schools would have a larger weight in the calculation of program effects. If the outcome of interest is student test scores, then the expected value should be calculated using school size X as a weight:

The average treatment effect should correspond to the extent possible to the population the evaluation aims to provide information about under each treatment scenario.

$$\hat{\beta} = E_X\{E[Y_i|X,T] - E[Y_i|X,C]\}, \tag{14}$$

Where β denotes program effect, Y denotes test scores or indicators of interest, T denotes assignment to treatment status, and C denotes assignment to control status. An alternative is to control for X (school size) in the regression of test scores on treatment(Duflo, Glennerster, and Kremer, 2008).

Reversion to the mean is not a problem for randomized evaluations but it is common in education. Unlucky students who perform poorly on one test will tend to score better on subsequent exams. This phenomenon, known as reversion to the mean, is not causal. This means that other factors that influence test scores

do not play a role in how lucky a student is on a given test. Program selection based on test scores may then lead to effects where mean reversion may play a role. Randomization solves the problem, as we would expect reversion to the mean to equally affect both treatment and control groups. The most common approach with other methodologies is to control for pre-program trends in test scores (Gertler, Patrinos, and Rubio-Codina, 2007). For a detailed discussion on the topic, see Smith and Smith (2005).

5.1.1 Analyze Groups of Interest

There may be some groups that are of special interest to explore. For example, it may be of interest to differentiate the effects of teacher training on students who perform poorly. To do so, choose the lowest tertile of the test score distribution in the baseline year. The best practice to explore effects in subgroups is to pose the questions after the evaluation in order to avoid data mining. Planning ex ante will allow for stratifying across groups to exploit variation in the experiment. Of special interest may be differences in changes according to pre-intervention test-scores (as in Banerjee et al. (2008)), socioeconomic status, and initial endowment of inputs when applicable. Other common subgroups of interest involve gender, age, and ethnic or cultural groups.

Stratify across groups to exploit variation in the experiment.

If the program is implemented at several sites and if the treatment effect is likely to vary substantially across these sites, calculate differences in means between treatment and control groups at each site. Then calculate a weighted average of these differences using the number of individuals in the treatment group as weights. Also use the sample size from each site as weights. The result may differ from OLS. For more details, see Card, Ibarraran, and Villa (2011).

5.1.2 How to Interpret Program Effects

Suppose the average treatment effect has been calculated. The interpretation is that one change in treatment status changes the outcome in the magnitude of the average treatment effect. This magnitude may be placed in context by comparing it to values at baseline or to benchmarks[11]. Interpretation of the results should not stop there. First, one change in schools causes individuals to adjust for other inputs. For example, Datar and Mason (2008) find that an increase in class size from 15 to 20 students is associated with a decrease in interactions between parents and students of 0.07 standard deviations and an increase in parent-financed activities of 0.03 standard deviations in kindergarten and first grade. Duflo, Dupas, and Kremer (2011) find that the effects of hiring short-term teachers in Kenya reduced teacher effort as a response to the lower pupil-teacher ratio and influenced parent-teacher

[11] Some studies collect data on non beneficiary individuals who perform well to serve as benchmarks.

associations to hire their relatives. They conclude that the reactions of stakeholders in a weak institutional environment reduced the impact of the program. Second, the change depends on the input levels A_0 and B_0. Therefore the change is specific to individual characteristics. An example is the study by Glewwe, Kremer, and Moulin (2009) where textbooks increased the scores of the best students, who could better understand them, but had little effect on other students. Lavy (2011) finds that a traditional teaching style (memorization) has a greater benefit on beginners (girls, low socioeconomic status, below median test scores), while a modern teaching style (critical thinking) benefits those more advanced (boys, high socioeconomic status, above median test scores). Additionally, we would expect households to react in a different way to a given program. For example, Elacqua and Fabrega (2007) argue that high-income parents tend to value student satisfaction, while low-income parents tend to value test scores over other factors such as school proximity or social connections.

Suppose the education production function depends on two inputs A and B. Assume the objective is to measure the effect of a change of one unit of input A in output $Y = Y(A, B)$. Assume the given value for a student is $Y_0 = Y(A_0, B_0)$. Suppose the student is allocated an extra unit of A. Then the observed change would be

$$\left.\frac{dY}{dA}\right|_{A_0,B_0} = \left.\frac{\partial Y}{\partial A}\right|_{A_0,B_0} + \left.\frac{\partial Y}{\partial B}\frac{\partial B}{\partial A}\right|_{A_0,B_0}, \tag{15}$$

Two conclusions can be drawn from this equation. First, equation 15 shows the direct change of test scores as a result of the change in input A plus the effect of the change of A on test scores via adjustment of input B. If A and B are substitutes, then increasing one may decrease the other. Changing a given input may cause changes in other inputs, therefore making it difficult to assess the effect of a change in a single input on test scores.

Changing a given input may cause changes in other inputs, thus making it difficult to assess the effect of a change in a single input on test scores.

Going back to the policy question of changing a given input to improve the quality of education, impact evaluation will provide an estimate for the expected value of the change of test scores on the population of interest. Assuming random allocation of input A:

$$E\left[\left.\frac{dY}{dA}\right|_{A_0,B_0}\right] = E\left[\left.\frac{\partial Y}{\partial A}\right|_{A_0,B_0}\right] + E\left[\left.\frac{\partial Y}{\partial B}\frac{\partial B}{\partial A}\right|_{A_0,B_0}\right], \tag{16}$$

Note that the effect of a change on the input depends both on the characteristics of the population as defined by the distribution of A and B and on the adjustments that agents make to other inputs as a result of a change in A. Two conclusions follow:

The effect of a change on a given input depends both on the initial endowment of inputs and adjustments.

1. Estimates of changes of a given input on test scores are specific to the population.

2. Estimates include adjustments to other inputs.

Todd and Wolpin (2003) discuss these issues in detail. The first point adds to the idea that with more information in terms of context, policies can be better fitted to specific groups. Therefore, reports on average effects on a given policy should clearly set parameters to make that policy comparable to other populations. Cross-cutting design discussed in Section 2.3.1 should be implemented taking into account other relevant inputs when possible. This would allow for a better understanding of how a given input affects other inputs and outcomes[12].

External Validity

Section 2.5 includes a discussion on how to select the sample. An evaluation is **externally valid** if it represents the eligible population of interest. Ideally, take a random sample of individuals who meet program participation requirements and evaluate based on that sample. The evaluation would provide information on the population toward which the program could expand. The evaluation should be designed to describe the population of interest for program expansion and/or policy recommendations. There are some features that could be introduced to the evaluation to try to make inferences about other populations and not leave the evaluation separate from those other populations. Rubin (1992) discusses how to make use of several studies to make inferences about the effects we would expect in a population not previously considered. He introduces the concept of an effect-size surface. Let $E[\tau|A_0, B_0] = E\left[\left.\frac{dY}{dA}\right|_{A_0,B_0}\right]$. In our example, an effect-size surface is defined as:

An effect-size surface is a function that gives the expected effect as a function of a given input.

$$E[\tau|A_0, B] = f(A_0, B) = f_0(B), \tag{19}$$

An effect-size surface is a function that gives the expected effect as a function of input B. Rubin (1992) generalizes B to the inclusion of relevant factors. He also emphasizes the problem resulting from high dimensionality, interactions, and curvilinear relationships, suggesting hierarchical and hyperparametrical models. This

[12] An alternative is to impose some structure as to how we believe the data are generated. Suppose that the data-generating process is:

$$Y_i = \beta_1 + \beta_A A_i + \beta_B B_i + u_i, \tag{17}$$

$$B_i = \alpha_1 + \alpha_A A_i + \epsilon_i, \tag{18}$$

Where $A \perp u_i, \epsilon_i$ and $B \perp u_i$. Then the elements of equation 16 can be identified as direct effect, $E\left[\left.\frac{\partial Y}{\partial A}\right|_{A_0,B_0}\right] = \beta_A$ and total effect $E\left[\left.\frac{dY}{dA}\right|_{A_0,B_0}\right] = \beta_A + \beta_B\alpha_A$. The model can be extended to condition on a set of controls X. An input B of special interest is the investment of the individual in participation.

idea is used in practice in Cruces and Galiani (2007). The working paper version of that study discusses in more detail how to assess the causal effect of fertility on labor supply (Cruces and Galiani, 2003). The authors conclude that the effects found in the United States can be generalized both qualitatively and quantitatively to the populations in Argentina and Mexico. They estimate a response surface of the estimated effects against real GDP per capita and find evidence that seems to suggest that the effect of fertility on the female labor supply increases with the wealth of the country. The authors assert that this statement assumes too much extrapolation, and they do not reject the null at standard levels of no differences across countries. In the case of education, it may be relevant to take into account that demand- and supply-side provision of inputs may behave as substitutes. In this sense it is of interest to collect data on other inputs to education that may change as a response to a given intervention. To estimate an effect-size surface in education, it may be relevant to stratify by input and then randomize within strata. With enough sample size per strata, it may be possible to estimate differentiated effects across different statuses of initial input. Stratification should extend to relevant inputs (such as B in my example) whenever possible. Duflo, Glennerster, and Kremer (2008) emphasize the importance of this idea when interpreting results. Variations in program implementation and target population are likely to cause different outcomes.

There are two final points to consider when reading results. First, question the presence of any Hawthorne or John Henry effects (discussed in Section 3). Second, consider effects that are present when a program expands but that are not present when a limited number of schools participate. For example, Hsieh and Urquiola (2003) posit that vouchers may increase the sorting of students by ability, preferences, and race when a program expands. However, this is not testable, as control schools may not experience significant competition, which may be present if the program becomes universal.

5.1.3 Deviations from the Original Design

This section discusses how to account for deviations from the original evaluation design. For example, teachers in participating schools may leave and new teachers may enroll. Teachers may also try to migrate to treated schools to be trained. As a result, the evaluation will require some additional considerations. These situations can be avoided to the extent possible through the use of operative guidelines. This section will provide guidance on how to deal with the unavoidable cases. To ease presentation, the section includes two subsections for two types of threats to evaluation design:

1. Individuals move in or out of the sample.

2. Individuals move in or out of the treatment/control groups.

Whether these changes are random or not is key to the analysis of data. The random case is discussed here and the nonrandom case is discussed in the subsections that follow.

If individuals move in or out of the sample randomly, then the design would have a change in power, but there is no need to make further adjustments besides documenting. One way to check whether individuals move in or out of the sample randomly is to first, calculate whether attrition rates are similar in the treatment and control groups. A t-test on a dummy variable which equals one for each individual present at baseline but not at follow up and zero otherwise would test if attrition rates are equal among treatment and control groups. Second, test if individuals who move out of the sample (this may include nonresponse) are different than those that did not migrate. For this you can focus on key outcomes of interest at baseline. Compare the mean of outcomes for the baseline sample to the sample of individuals observed at follow up. If the null of differences is rejected between the two, then it is likely same sampling bias is present (Card, Ibarraran, and Villa, 2011). If it is known before the program is implemented that there will be a number of individuals who are likely to drop out of the sample by treatment status, then adjust the sample size to achieve a given minimum detectable effect. The formula given by Duflo, Glennerster, and Kremer (2008) is

$$MDE = (t_{(1-\kappa)} + t_{\alpha})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}\frac{1}{c-s}, \tag{20}$$

Where $t_{(1-\kappa)} + t_{\alpha} \approx 2.8$ is a test with power of 0.80 and significance of 0.05, c is the share of subjects initially assigned to the treatment group that actually receive it, s denotes the share of individuals initially assigned to the comparison group who receive the treatment, $P = T/N$ is the proportion of treated individuals, $N = T + C$ is the total number of observations, and σ is the standard deviation of the outcome of interest. If program operation requires filling open slots, prepare a waiting list to use when the pull of applicants allows for doing so. The waiting list can similarly be created by randomly assigning individuals to a treatment or control group with a random order of entrance. In cases of programs that are continuously recruiting, randomize for each "cohort" of applicants. Note that individuals can differ across cohorts, and therefore cross-cohort comparison will reflect not only differences related to time of program participation but also to participant characteristics (Card, Ibarraran, and Villa, 2011). In the absence of substitution for individuals who leave, estimate whether or not those who remain are a representative sample of the baseline sample, and account for losses in external validity.

If individuals move in or out of the sample or among treatment or control status for random reasons, then use the new allocation and evaluate. Changes of individuals across treatment and control groups randomly is usually known to the evaluator because the assignments did not depend on individuals' behavior or characteristics. The assignment process should be random. To validate the new allocation of individ-

uals, document the process that generated the new allocation and test for balance. Sometimes data are missing because collection is such that it is not possible to obtain all answers in a given questionnaire. Appendix 6 includes a brief discussion on how to deal with missing data for a given individual under the assumption these nonresponse items are not systematic. The next two subsections discuss strategies to deal with nonrandom changes to the original evaluation design. The first subsection examines how to deal with changes in the sample and the second looks at how to deal with changes in program participation.

How to Deal with Changes in the Original Sample

Sometimes individuals move in or out of the sample for reasons beyond the controls of the implementing agency or the evaluator. **Sample selection** takes place when those who benefit from a given program are oversampled. **Attrition** is the loss of observations between successive rounds of data collection. If information is available on factors that influenced inclusion in the sample, then this situation is defined as **selection on observables**. In this case, the solution to the bias introduced by ignoring the problem would be to control for characteristics that make individuals drop out or match (Cameron and Trivedi, 2005, p. 869). When there is no information on factors that influenced inclusion in the sample and it is not possible to observe the characteristics of individuals that dropout, then the sample suffers of **selection on unobservables**.

Control for variables that explain dropouts from the program.

In cases of selection on unobservables there are two options. First, you can introduce structural models to correct for attrition bias. In this case, you may follow the Roy Model and do Heckman correction (Cameron and Trivedi, 2005). This approach requires the adoption of stronger assumptions than the second approach suggested. Second, you may consider two methods to address attrition proposed by Duflo, Glennerster, and Kremer (2008). The first is to use nonparametric Manski-Lee bounds. To understand how this works, suppose there is a higher nonresponse rate in the treatment group than in the control group. Drop the observations within the treatment group with the lowest values of the outcome variable at baseline until the fraction retained is equal to that of the control group. A comparison of the resulting groups will provide an upper bound. To obtain a lower bound, drop observations in the treatment group with the highest values. The bounds will be wider when the difference in response rates is large or the outcome has a large variation. The second method discussed by Duflo, Glennerster, and Kremer (2008) draws on a method proposed by Angrist, Bettinger, and Kremer (2006). First, find the share of individuals in the control group with zero values on a variable that is zero for those that leave. Call this variable Y. Once this percentage is calculated, subtract from the estimation the percentage of individuals in the treatment group with the lowest values of Y. The resulting sample will provide an upper bound on the average treatment on the treated (ATOT) conditional on those who remain when the

program is not offered. This method assumes that treatment status is never harmful and that those offered treatment status are at least as likely to remain in the sample as those who are not and rank preservation. The rank preservation restriction states that when the potential outcome in the comparison state is above a certain quantile in its own distribution, then the potential outcome in the treatment state is also above that quantile in its own distribution. An alternative is to separate the outcome in quantiles and estimate bounds. For more details, see Card, Ibarraran, and Villa (2011). Gertler, Patrinos, and Rubio-Codina (2007) state that best-practice impact evaluations aim to keep nonresponse and attrition below 5 percent.

How to Deal with Changes in Original Treatment and Control Allocation - Imperfect Compliance

Suppose that the most motivated teachers in the control group lobby the implementing agency and manage to receive training. An evaluation is said to have **imperfect compliance** when there is a difference between planned and actual treatment status. **Self-selection** happens when the individual choice of whether or not to participate is determined by characteristics that also determine the outcome, like motivation. Individuals are likely to decide to participate or not based on their assessment of likely gains from participating. In this case, a comparison would not only provide the effects of the program but also the difference between motivated and unmotivated teachers. **Selection bias** takes place when we compare two groups that are different even in the absence of program participation. In our example, the two groups will differ in teacher motivation. The definition of treatment is not always straightforward when treatment can be received partially. An example is a teacher training course where dropouts are possible. It may be that a dropout takes place too soon to consider a teacher or a student as having participated in the program. The definition of **participation** is made on a case-by-case basis, but it is usually defined in terms of costs when costs are determined regardless of program completion. Individuals who do not show up at the start of the program will not be considered treated (Card, Ibarraran, and Villa, 2011).

Estimate the share of teachers and students in treatment schools within a municipality to check for migration.

For interventions where the unit of treatment is not the individual, one may consider **migration** . For example, a school-level program successful at improving learning and/or teaching conditions may create an incentive for teachers or students to move to those schools. Better students and teachers try to enroll in better performing schools. As a result, test scores at the school level would improve without a role played by the program (Hsieh and Urquiola, 2003). To test for the presence of migration, Galiani, Gertler, and Schargrodsky (2008) estimate the share of students and teachers enrolled in treatment schools to confirm whether or not migration takes place:

$$Share_s = \alpha + \beta DT_s + \epsilon_s, \tag{21}$$

Where *Share* is the percentage of students in a given municipality enrolled in benefited schools. The authors cannot reject the null of β being zero. Therefore they conclude that migration is unlikely to influence their estimates. Gertler, Patrinos, and Rubio-Codina (2007) suggest controlling for parental education as a proxy for parental motivation to move their children in or out of treatment schools. They also suggest to bound effects. Calculate the upper (lower) bound by adding (subtracting) the estimated probability of being treated (or control) given a set of observable characteristics. If teachers, principals or school characteristics change, then it may be necessary to control for changes in school characteristics in the evaluation.

What if not all of the individuals comply with the program treatment status that was assigned to them? It may not be possible to directly calculate the average treatment effect defined in equation 5. The evaluation can provide much information about the program by using the original allocation to estimate the **Intention-to-Treat (ITT)**. The ITT will provide an estimate of the benefits of implementing a program regardless of whether the program was actually used. How useful it will be will depend on the program characteristics. For example, suppose a program provides vitamins to parents to improve nutrition in children in order to improve test scores. But it is not known if parents actually give the vitamins to children because it is not possible to collect information in their homes to observe if it happens. In this case, ITT would provide the average effect on the population of *providing parents with vitamins*. Estimation will not allow for estimation of the effects of *the children taking the vitamins* on test scores. The **Treatment on the Treated (TOT)** is the ITT divided by the difference in participation rates of the treatment and control groups: $TOT = ITT/D$ (Card, Ibarraran, and Villa, 2011).

Use the original allocation to estimate the intention to treat.

The local average treatment effect (LATE) estimates can be recovered using an instrumental variable approach. Local average treatment effects refer to the effect of people in the margin of switching program participation. Suppose there are three types of students: compliers, always-takers, and never-takers. Assume there are no defiers. Compliers participate according to assignment, always-takers participate regardless of assignment, and never-takers do not participate regardless of assignment to treatment or control groups. The local average treatment effect is a measure of the effect of treatment on the subgroup of those at the margin of participating, denoted as compliers. The treatment-on-treated estimates are a case when never-takers are excluded (Cameron and Trivedi, 2005). Gertler et al. (2011) propose extending the instrumental variables approach to other contexts in which other methods may apply through:

Use individuals in the margin of switching program participation to calculate local average treatment effects.

1. **Regression discontinuity** : Using an instrumental variable that would be a 0/1 variable that indicates if the unit is located on the ineligible side.

2. **Differences in differences and selective migration** : Using the location of the individual before the announcement of the program as an instrument for location of the individual after the start of the program.

The local average treatment effect will be different than the average treatment effect if individuals do not react in the same way to treatment. For instrumental variable estimation with the heterogeneous treatment effect, see Cameron and Trivedi (2005), chapter 25.

Selection models can also be applied to attempt to correct for imperfect compliance. Maddala (1983) discusses models of sequential self-selection where individuals first choose to participate in the experiment and then whether or not to accept treatment status. In either case, regardless of compliance status, information is required to improve estimates. Information to track people at baseline should be collected when possible.

5.2 Estimate the Confidence Interval for Program Effects

Program impacts are usually heterogeneous and therefore will present some variance. As a result, the calculated average treatment effect must be accompanied by some information regarding variation in program effects among students. The standard way to account for variation in the data is to assume that the difference of individual test scores after considering treatment status $y_i - E[Y_i|T]$ is determined by other factors that are a random draw from a normally distributed process. Given this setting, the question is, where would the calculation of program effects fall most of the time? This is the rationale to calculate confidence intervals. The confidence interval where 95% of estimated differences of theoretically infinite random draws would fall is given by

$$[\beta - 1.96\frac{\sigma_\beta}{\sqrt{n}}, \beta + 1.96\frac{\sigma_\beta}{\sqrt{n}}], \tag{22}$$

Where β denotes the estimation of the program effect, σ_β denotes the standard deviation of the estimation, and n denotes the number of observations. For a derivation of this formula and a more rigorous discussion, see Cameron and Trivedi (2005). To get an intuition on results by looking at program effects and standard errors, multiply the standard error $SE = \frac{\sigma_\beta}{\sqrt{n}}$ by two (an approximation of 1.96) and add and subtract to the estimate of the program effect β. Reject the null hypothesis of no effect when zero is not in the confidence interval.

The calculation of standard errors depends on the assumptions of the process that generate the error term $\epsilon_i = y_i - E[Y_i|T]$. An incorrect specification for a model for the data-generating process results in incorrect and/or larger standard errors. The subsections that follow discuss the scenario in which within-group errors are correlated but not across groups. This is equivalent to assuming that the error between two students in the same school is likely to be correlated, but not the error between two students in different schools. If the part of the outcome that is not explained by the model does not follow this structure, see Greene (2003) for other structures.

The standard way to account for variation in test scores after considering treatment status is to consider that differences are a random draw from a normally distributed process.

5.2.1 Standard Errors for Clustered Data

In education, errors are likely to be clustered by teacher, school, and/or states. Assume that errors are correlated within groups but not across groups. Bertrand, Duflo, and Mullainathan (2004) find that ignoring correlated outcomes may lead to the calculation of smaller-than-actual confidence intervals. As a result, analysis may erroneously lead to conclude the program led to an effect when in fact it did not; that is, zero falls with the confidence interval of the program effect. This is over-rejection of the null of no effect. The authors suggest assuming that the correlation of errors within a given group can be approximated by the average of the correlation within groups. In order to have a good approximation, Bertrand, Duflo, and Mullainathan (2004) suggest allowing for an arbitrary variance-covariance matrix within clusters when the number of clusters is larger than 50. Therefore, consider the use of the White-like formula to compute the standard errors[13].

Ignoring correlated outcomes may lead to the calculation of smaller-than-actual confidence intervals.

5.2.2 Standard Errors with Small Samples

As discussed in Section 3, sometimes it is not possible to randomize over a large number of groups. For example, consider the case of an evaluation with less than 60 schools participating in the evaluation. Bertrand, Duflo, and Mullainathan (2004) show the cluster-correlated Huber-White estimator performs poorly when the number of clusters is small (< 50) and leads to over-rejection of the null hypothesis of no effect. An alternative approach is proposed by Rosenbaum (2002), who suggests randomly generating placebo treatments and replacing the treatment dummy to calculate standard errors. The intuition is as follows: suppose the effects found are a result of a lucky draw of students into the treatment group. If this lucky draw is common, it is likely the results are just that. On the other hand, if a draw of students showing the effects found is rare, then it is most likely that the effect found is a result of the program. The way to implement this idea is to take equation 8 and replace by a randomly generated placebo DP[14]. DP should replace at the level at which treatment was randomized. If all students within a school have a given status, the placebo should also be homogeneous for that group. After generating placebo status, estimate effects for placebo treatment as follows:

Randomly generate placebo treatments, and replace the treatment dummy to calculate standard errors.

$$Y_{ij} = \delta + \beta_P DP_j + \nu_{ij}, \tag{23}$$

Let $F(\hat{\beta}_P)$ be the empirical cumulative density function of $\hat{\beta}_P$ for all elements of DP_j. Suppose the effect calculated for the program with actual treatment status DT is β_T. Now perform an hypothesis test by checking if the measured treatment effect is in the tails of the distribution of placebo treatments. Reject $H_0 : \hat{\beta}_T = 0$ with a

[13] In Stata, this is equivalent to xtreg y d2 dtd2,robust cluster(teacher_id)

[14] DT was defined in equation 8 as a dummy that equals 1 if the teacher was assigned to the treatment group and 0 otherwise.

confidence level of $1-\alpha$ if $\hat{\beta}_T \leq F^{-1}(\frac{\alpha}{2})$ or $\hat{\beta}_T \geq F^{-1}(1-\frac{\alpha}{2})$. Since the placebo assignments DP_j vary only across clusters, this method takes intra-cluster correlations into account. The authors warn that this technique can result in larger confidence intervals when the true effect is large because it does not put even minimal structure on the error term. Gertler et al. (2011) extend this idea in two ways, first by incorporating the stratification structure in the calculation and second by incorporating conditioning on the covariates that were not balanced at baseline in their experiment. Randomization makes it possible to permute treatment status within blocks without affecting the differences between treatment and control groups. Treatment is not related to individual characteristics within a block because the assignment of individuals was random. As a result, treatment status is exchangeable within blocks. If there is no effect in reality, differences in outcomes should not change. Therefore, means should not be statistically different. Gertler et al. (2011) break strata further by including balanced variables. The authors argue that since treatment is exchangeable within a partition of individuals who share the same characteristics $\mathbf{X}$, it is also exchangeable for any finer partition with $\mathbf{Z}$ as conditioning variables. The authors only keep $\mathbf{Z}$ to the set of variables that are not balanced at baseline, because as the blocks are reduced in size the number of participants decreases. as does the number of possible permutations. At the limit, each participant would be in one block and no permutation would be possible. Gertler et al. (2011) also only condition on imbalanced baseline variables if their impact on target outcomes is statistically significant. Consistent with previous intuition, stratification works to the extent that the variables explain variation in the outcome. This nonparametric solution is flexible, as it can be applied to any data-generating process of outcomes. For examples of studies with applications of permutations, see Allegretto and Reich (2009); Chetty, Looney, and Kroft (2009); Conley and Taber (2011); Ho and Kosuke (2008). An alternative to permutations is to order individuals according to the outcome of interest from best to worst. The Wilcoxon-Mann-Whitney test allows for determining if a program caused improvements of treatment units relative to the control units. It is not possible with this test to estimate the magnitude of the effect.

5.2.3 Multiple Outcomes

Experiments with multiple outcomes are more likely to find an effect in some of them even if there are no such effects. For example, suppose program effects are measured on standardized tests scores in two academic subjects, student failure and dropout rates, teacher and student assistance, student and parental participation, and teacher and student motivation. Duflo, Glennerster, and Kremer (2008) calculate that a test of 10 independent outcomes will yield finding an effect in at least one with probability of about 40% when there is no effect in reality. Testing for multiple hypotheses can also contribute to improving accuracy. Bloom and Institution (2006)

Experiments with multiple outcomes are more likely to find an effect in one of the outcomes.

state that to the extent that outcomes within clusters are not perfectly correlated, individuals within clusters may increase power. Assume the goal is to test the effect of a program for k different outcomes π:

$$H_o : (\pi_1, ..., \pi_k)' = \mathbf{0}, \tag{24}$$

, versus

$$H_a : (\pi_1, ..., \pi_k)' \in O^+, \tag{25}$$

Where O^+ is the positive orthant. Normalize each outcome by its standard deviation and take the average of

$$\tau = \frac{1}{K} \sum_{k=1}^{K} \frac{\pi_k}{\sigma_k}, \tag{26}$$

O'Brien (1984) showed that τ could be used to test a restricted version of hypothesis of 24 under the additional assumption of a constant treatment effect across outcomes within the family. Let t_K be the $KX1$ vector per-comparison t ratio for each of the treatment effects. Let R be a KXK matrix with elements $\rho_{kl} = corr(\pi_k, \pi_l)$ and j be a $KX1$ vector of ones and j^T denote its transpose. Then the t-ratio for τ in testing hypothesis 24 is given by equation 27:

$$t_\tau = \frac{j^T t_K}{\sqrt{j^T \hat{R} j}}, \tag{27}$$

O'Brien (1984) showed that the ratio t_τ is t distributed with $n-2$ degrees of freedom.When outcomes are defined over different groups, outcomes can be aggregated to a common level (e.g., different age groups add to the village level) to obtain a consistent unit of observation. For example, Bloom and Institution (2006) define 22 health outcomes, including provision of vitamin A to children under age 5, treatment of diarrhea in children under 5 who have symptoms, and private health spending. They aggregate outcomes to the village level (e.g., children 12-23 months old, women who have given birth in the past year) to obtain a consistent unit of observation. Once they have a consistent number of observations across outcomes, they proceed to test for hypothesis 24.

5.3 Analysis Checklist

The checklist in Table 10 should be completed to ensure that all of the elements of data analysis are included in the evaluation design and procedures.

Table 10: **Checklist for Data Analysis**

Category	**Question Checklist**	**Strategy if Answer is Yes**
Program effects	Is a sampling strategy necessary to weight observations?	Re-estimate using appropriate weights (equation 14)
	Are there changes in the sample as a result of student enrollments and dropouts?	Check if dropouts are related to the program (Subsection 5.1.3)
	Is migration of teachers and students between treatment and control schools a possibility?	Check with equation 21
	Is there potential for contamination?	Assess extent of contamination (Subsection 2.3.2)
	Are there cases of imperfect compliance?	Calculate for intention to treat or late average treatment effect (Section 5.1.3)
	Is analysis of a specific group relevant?	Consider pre-intervention test scores, initial endowment of input when applicable, socioeconomic status, gender, age, and ethnic or cultural group (Subsection 5.1.1).
Confidence intervals	Are errors likely to be correlated within some groups?	Recalculate using standard errors for clustered data (Subsection 5.2.1)
Yes/No Checklist Based on Subsection 4.5		
Report	Is there a table showing baseline demographic characteristics and outcomes of interest such as test scores for each group?	Yes/No
	Is enough information available to provide a clear idea of the context in which the evaluation was set up?	Yes/No
	Are statistical methods used to compare groups clearly stated?	Yes/No
	Are estimates of confidence intervals for each outcome of interest included?	Yes/No
	Are limitations included (such as interpretation of indexes, potential threats to identification of effects, etc.)?	Yes/No
	Is external validity discussed?	Yes/No
	Are sources of funding reported?	Yes/No
	Is other relevant documentation included (such as the individuals involved, when applicable)?	Yes/No
	Are databases, code, and ancillary files (i.e., results of other analysis performed) available and referenced?	Yes/No

Source: Prepared by the author.

6 Conclusions

The goal of these guidelines has been to summarize evidence and provide references in a single document so as to save time for those implementing impact evaluation in education through randomized control trials. The guidelines aimed to provide a starting point to pose questions and prepare evaluation strategies based on the work done to date. The document has laid out the following steps to set up impact evaluation of education programs:

1. First, determine the relevance of the proposed evaluation. Determine the main variables that explain variation in the outcome of interest, or the test scores in the population of interest. Prepare a theory on how the program is expected to change the resources available to school constituents and their behavior. The guidelines have provided a thorough review of the literature on the determinants of investment in basic education.

2. Second, specify the main evaluation and questions derived from the causal chain. Be clear about who is being treated, what changes are expected as a result of the program, and when and how those changes will manifest themselves. The document has provided checklists with key questions to answer before designing the evaluation

3. Third, set up the treatment and control status in the sample. The document has provided the formula to get started.

4. Fourth, collect data. The document has provided a checklist and multiple links for tests and questionnaires. It also includes a list of suggested variables to collect. The importance of baseline and monitoring data is emphasized. The guidelines also suggest tracking the individuals chosen to participate in the study.

5. Fifth, analyze data. These guidelines have provided the basics of impact evaluation and the intuition behind different calculations. Common methodologies to deal with changes to evaluation design were also presented, as well as a checklist to review for inclusion of analysis considerations in the evaluation report.

This document includes multiple references and links to analyze a given topic in more detail. However, it is not exhaustive and methodologies are not covered in depth. The main contribution of the guidelines is to put forth a methodology to design impact evaluation, and to provide multiple references. Future versions of the guidelines should continue to incorporate advances in the field and any new resources that are developed. These tools should have as their common goal to support our shared effort to learn how to provide good-quality and equitable education through impact evaluation.

Appendix A How to Do Variance Decomposition

The main idea behind variance decomposition is to take into account how much of the variation in certain variables is explained by differences across averages within a given group. For example, let the observation of test scores be denoted by Y_{ig} for student i in group g. Define the sum of squares of test scores across all students in the population as

$$SS_{TOT} = \sum_g \sum_i (Y_{ig} - \bar{Y})^2, \tag{28}$$

Where $\bar{Y} = \sum Y_{ig}$ is the average over the whole population.

The between-group sum of squares is defined as

$$SS_{BTW} = \sum_g (\bar{Y}_g - \bar{Y})^2, \tag{29}$$

and the within-sum of squares is defined as $SS_{WTH} = SS_{TOT} - SS_{BTW}$. Assume the share of SS_{TOT} explained by variance within a group and across groups is of interest. Divide $SS_{TOT} = SS_{WTH} + SS_{BTW}$ over SS_{TOT} to get $1 = \eta_0 + \eta_1$, where η_0 is the share of variation within groups and η_1 is the share of variation across groups. Use the Stata command *oneway* to obtain the calculations of the sum of squares and calculate η_0 and η_1. For a similar procedure on variance decomposition, see Ramírez (2007). Another work using regressions with fixed effects is Mizala, Romaguera, and Urquiola (2007).

Now consider more than one level. For example, assume calculation of variation across students, classroom, grades, schools, and states is relevant. Follow the steps below:

1. Take student-level data and perform decomposition of the sum of squares grouping by classrooms. Get $SS_{WTH1} = SS_{TOT} - SS_{BTW1}$. The share of variation by groups is $\eta_1 = \frac{SS_{BTW1}}{SS_{TOT}}$. Calculate $\eta_0 = 1 - \eta_1$.

2. Average over classes to obtain classroom-level data and perform decomposition again taking grade levels as groups. Obtain $SS_{WTH2} = SS_{TOT2} - SS_{BTW2}$. The share of variation across groups is $\eta_2 = \frac{SS_{BTW2}}{SS_{TOT2}}$. The η_2 share of the variation in η_1 can be attributed to differences across grades. Therefore, the variation by classrooms is $\eta_1^* = \eta_1 - \eta_2$

3. Average over grades to obtain grade-level data and perform decomposition again taking schools as groups. Obtain $SS_{WTH3} = SS_{TOT3} - SS_{BTW3}$. N η_3 is the share of the variation in η_2 that can be attributed to differences across grades. Therefore, the variation by grades is $\eta_2^* = \eta_2 - \eta_3$.

4. Average over schools to obtain school-level data and perform decomposition again taking states as groups. Obtain $SS_{WTH4} = SS_{TOT4} - SS_{BTW4}$. η_4 is the share of the variation in η_3 that can be attributed to differences across grades. Therefore, the variation by schools is $\eta_3^* = \eta_3 - \eta_4$.

Therefore, decompose variance as

$$1 = \eta_0 + \eta_1^* + \eta_2^* + \eta_3^* + \eta_4, \tag{30}$$

where η_0 denotes the share of variation across students in the classroom, η_1^* denotes the share of variation across classrooms within a given grade, η_2^* denotes the share of variation across grades within a school, η_3^* denotes the share of variation across schools within a state, and η_4 denotes the share of variation across states.

Consider ENLACE, the standardized mathematics test in the state of Puebla in Mexico. ENLACE is a diagnostic tool to measure how much knowledge students have to successfully learn from what is programmed in the curricula for the next school year. It is a standardized test that has been applied to all students at the end of every school year since 2006. Test scores range from 200 to 800. The test was designed to have an average of 500 points and a standard deviation of 100 in 2006 for every subject grade[15].

ENLACE information constitutes an individual-level balanced panel for 86,713 elementary-level students in public schools of the general type (excludes indigenous and community based) the years from 2009 to 2011. Information on test scores is missing for at least one of the three years for 34% of students. Estimates are likely to represent those individuals more likely to not migrate in or out of school and between schools that do not participate in ENLACE. Individuals attending the fourth grade in 2009 are tracked to fifth grade in 2010 and sixth grade in 2011. The sample includes 3,565 schools that are located in 1,969 localities and 213 municipalities. On average, there are 24 students in a school, two schools in a locality, and nine localities in a municipality. The average test scores increase from one year to the next, with a standard deviation of 122. In our sample, 7% of students migrate to a different school.

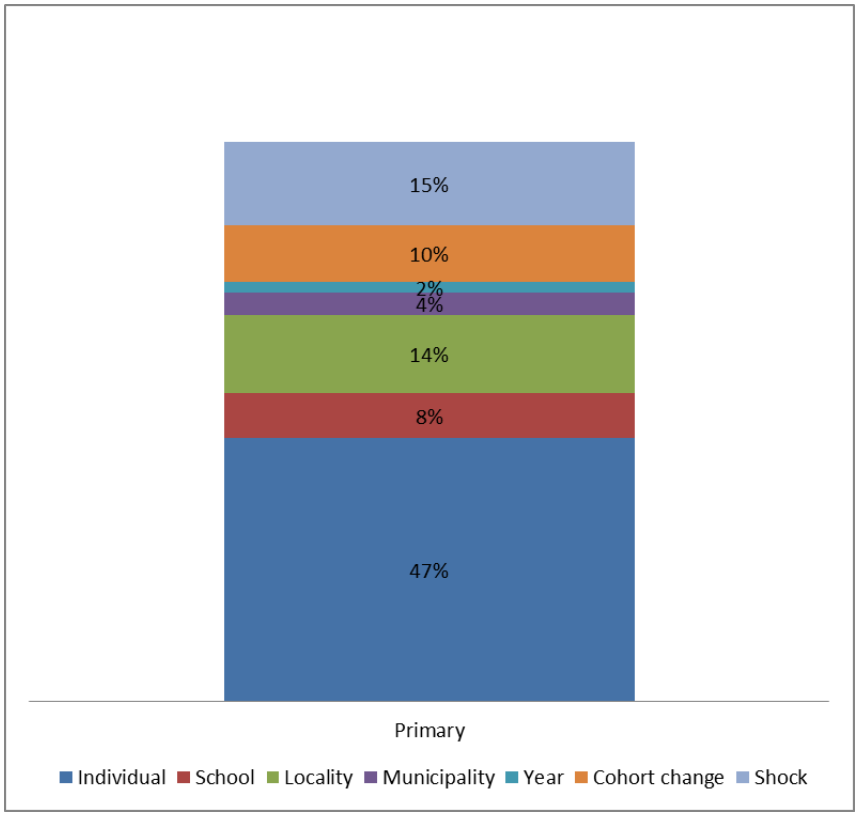
Variance decomposition results are summarized in Figure A.1. Figure A.2 includes standard deviations of test scores at different levels to extend the interpretation of variance decomposition. Three important conclusions can be drawn from these two figures:

1. The best student in a bad school performs better than a bad student in a good school. This idea extends from schools to groups, localities, and municipalities.

2. When tested over several years, the lowest performance of a good student is better than the best performance of a bad student.

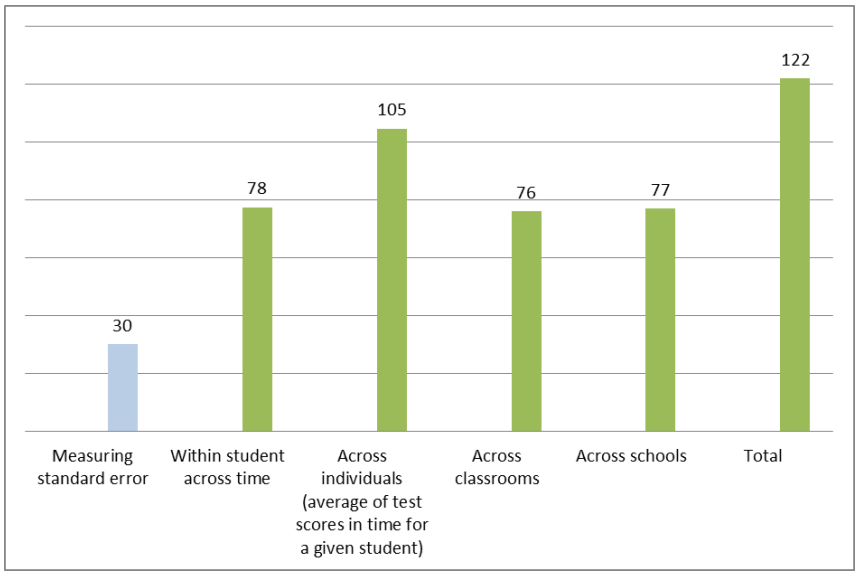
[15]For more information on the ENLACE test, visit www.enlace.sep.gob.mx

Figure A.1: **Variance Decomposition of Test Scores for Students Who Do Not Migrate**



Source: Prepared by the author.

Figure A.2: **Standard Deviations for Math ENLACE Test Scores in Puebla, Mexico**



Note: ENLACE = Evaluación del Logro Académico en Centros Escolares (Mexico).
Source: Prepared by the author.

3. Most of the variation takes place at the individual level and there is variation at every level.

Variance decomposition for several school types and for schools at the secondary level leads to similar conclusions.

Appendix B Factor Analysis

Factor analysis identifies the eigenvectors of the correlation matrix. The reconstruction of the original variables is therefore a linear combination of the eigenvectors and an error term. More specifically, factor analysis finds q common factors that can reconstruct the p original variables:

$$y_{it} = f_{i1}b_{1j} + f_{i2}b_{2j} + ... + f_{iq}b_{qj} + \epsilon_{ij}, \tag{31}$$

where y_{ij} is the value of the ith observation on the jth variable, f_{ik} is the ith observation on the kth common factor and b_{kj} are factor loadings. The residual term is denoted ϵ. The interpretation of the index should be made with care, as possible solutions exist to loadings. Factor loadings are the weights that are given to a given factor[16].

[16]To estimate factor loadings in Stata, do the following: factor y_1 y_2 y_3 ... y_p, pct. predict index.

Appendix C Example of Explanation to Evaluation Participants

This appendix presents an example of a letter to community instructors. In the evaluation, it is very likely those instructors will communicate with each other, so it is important to provide information. Note the example here includes an evaluation with two different components (see Section 3 on spillovers).

Dear Community Instructor,

The goal of this letter is to share with you our efforts to improve ways to support your work in the future. We are in the pilot stage of two differentiated schemes for community instructors: upfront payments every three months and a recognition scheme. Payments have traditionally been made every month and we would like to know if payments every three months would work better for most community instructors. A second point we would like to learn about is the "perseverance sheet." which is a diploma that recognize your effort at the end of the school year.

Before we make any changes in our policies we need to carry out a pilot program to evaluate whether or not these alternative components work. This year we are starting with 375 community instructors receiving payments under the alternative scheme and 660 community instructors receiving the "perseverance sheet". We will monitor how these components work this year and we will then improve these alternatives before we extend them to all community instructors.

We appreciate your collaboration with our efforts. It would not be possible for us to learn and to make improvements without your support. We especially appreciate your understanding if you were not selected to be in the group you would have preferred. We made a random choice of participants to be fair to everyone and to make possible a comparison to assess the effects of changes. Your help will contribute to improving the support we will provide next year to you and to future generations. Our hope is that better service will result in better learning of our children.

Our priority is to support each one of our community instructors. We are committed to creating and improving the best way to support and recognize your work and improve your lives. We sincerely hope we can count on you, and we hope to provide you with improved support next year.

If you would like to express your views or if you have questions or comments regarding the evaluation, you can email us at XXXXX.com by XX/XX/XXX.

Please see the Stata manual for more details. A good interpretation of Stata output can be found at: http://dss.princeton.edu/training

Appendix D How to Deal with Missing Data

If information is missing on the outcome variable, it is usually a bad idea to fill in missing values of , because that would require you to predict values, which in turn would end up affecting estimation. (See Greene (2003) p.427, for a deeper discussion.) If there are missing values of the explanatory variable, X, replacing the missing values with the average is equivalent to dropping the observation. This happens because the regression line goes through the mean and basically you are adding a dot on top of an already-existing line. The cost is that the R^2 will be lower. Computing fitted values for the missing X's using values of y and X may not be a good idea either. The proposed solution is to use the information available on X to fill the gaps, but to not use the values of y. To provide an idea of how this works, assume it is relevant to include school size and rural/urban status in program estimation, but information is missing for some values for the rural/urban status. Use the school size in a linear regression to predict the values on rural/urban and predict the values. Then use the predicted values to fill in the gaps and estimate the full equation. With this procedure, all the variation in school size and test scores is exploited without losing the observations without information on rural/urban status (Greene, 2003). For example, assume two sets of variables Z and X are available and the goal is to estimate

$$y = \beta X + \gamma Z + \epsilon, \tag{32}$$

where data are missing for Z for a set of observations. Suppose that there is a linear relationship between X and Z:

$$X = \delta Z + u, \tag{33}$$

Do the following:

1. Estimate $X = \delta Z + u$ with the observations that have data for both X and Z.
2. Predict $\hat{X}$ using $\hat{\delta}$ and the observed values for Z.
3. Create a new variable $X2 = X$ if data are available or $X2 = \hat{X}$ if X is missing.
4. Estimate $y = \beta X2 + \gamma Z + \epsilon$

For more details, see (Greene, 2003, p. 431). For practical help, see the Stata command *predict*.

References

Allegretto, Arindrajit Dube, Sylvia and Michael Reich. 2009. "Spatial Heterogeneity and Minimum Wages: Employment Estimates for Teens Using Cross-State Commuting Zones." Working paper, institute of industrial relations. university of california at berkeley, Institute for Research on Labor and Employment. URL http://ideas.repec.org/p/cdl/indrel/qt1x99m65f.html.

Andrabi, Tahir, Jishnu Das, and Asim Ijaz Khwaja. 2009. "Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets." *Bureau for Research and Economic Analysis Development Working paper* (226). URL http://ipl.econ.duke.edu/bread/abstract.php?paper=226.

Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." *American Economic Review* 92 (5):1535–1558. URL http://ideas.repec.org/a/aea/aecrev/v92y2002i5p1535-1558.html.

Angrist, Joshua, Eric Bettinger, and Michael Kremer. 2006. "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia." *American Economic Review* 96 (3):847–862. URL http://www.aeaweb.org/articles.php?doi=10.1257/aer.96.3.847.

Arulampalam, Wiji, Sugato Dasgupta, Amrita Dhillon, and Bhaskar Dutta. 2009. "Electoral Goals and Center-State Transfers: A Theoretical Model and Em-pirical Evidence from India." *Journal of Development Economics* 88 (1):103–119. URL http://ideas.repec.org/a/eee/deveco/v88y2009i1p103-119.html.

Banerjee, Abhijit and Rohini Somanathan. 2007. "The Political Economy of Public Goods: Some Evidence from India." *Journal of Development Economics* 82 (2):287–314. URL http://ideas.repec.org/a/eee/deveco/v82y2007i2p287-314.html.

Banerjee, Abhijit V., Rukmini Banerji, Esther Duflo, Rachel Glennerster, and Stuti Khemani. 2008. "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India." Policy Research Working Paper Series 4584, The World Bank. URL http://EconPapers.repec.org/RePEc:wbk:wbrwps:4584.

Barrera-Osorio, Felipe. 2007. "The Impact of Private Provision of Public Education: Empirical Evidence from Bogotaâs Concession Schools." Policy Research Working Paper Series 4121, The World Bank. URL http://ideas.repec.org/p/wbk/wbrwps/4121.html.

Basinga, Paulin, Paul J. Gertler, Agnes Binagwaho, Agnes L.B. Soucat, Jennifer R. Sturdy, and Christel Vermeersch. 2010. "Paying Primary Health Care Centers for Performance in Rwanda." Tech. rep. URL http://ssrn.com/abstract=1543049.

Becker, Gary. 1964. "Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education." *NY: Golumbia U. Press* .

Berlinski, Samuel, Sebastian Galiani, and Paul Gertler. 2009. "The Effect of Pre-primary Education on Primary School Performance." *Journal of Public Economics* 93 (1-2):219–234. URL http://ideas.repec.org/a/eee/pubeco/v93y2009i1-2p219-234.html.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *The Quarterly Journal of Economics* 119 (1):pp. 249–275. URL http://www.jstor.org/stable/25098683.

Bloom, E. and Brookings Institution. 2006. *Contracting for Health: Evidence from Cambodia*. Brookings Institution. URL http://books.google.com/books?id=I-CSNwAACAAJ.

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts. 2011. "Does Management Matter? Evidence from India." NBER Working Papers 16658, National Bureau of Economic Research, Inc. URL http://ideas.repec.org/p/nbr/nberwo/16658.html.

Bruns, Barbara, Deon Filmer, and Harry Anthony Patrinos. 2011. *Making Schools Work: New Evidence on Accountability Reforms*. World Bank Publications. URL http://siteresources.worldbank.org/EDUCATION/Resources/278200-1298568319076/makingschoolswork.pdf.

Bullock, Will. 2004. *Fair Share, Share Fair: Representation, Power, and the Distribution of Federal Funds*. Ph.D. thesis, Public Policy Program, Standford University.

Cameron, A. Colin and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge University Press.

Card, David E., Pablo Ibarraran, and Juan Miguel Villa. 2011. "Building in an Evaluation Component for Active Labor Market Programs: A Practitionerâs Guide." SPD Working Papers 1101, Inter-American Development Bank, Office of Strategic Planning and Development Effectiveness (SPD). URL http://ideas.repec.org/p/idb/spdwps/1101.html.

Chay, Kenneth Y., Patrick J. McEwan, and Miguel Urquiola. 2005. "The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools." *American Economic Review* 95 (4):1237–1258. URL http://ideas.repec.org/a/aea/aecrev/v95y2005i4p1237-1258.html.

Chen, Yuyu and Hongbin Li. 2009. "Motherâs Education and Child Health: Is There a Nurturing Effect?" *Journal of Health Economics* 28 (2):413–426. URL http://ideas.repec.org/a/eee/jhecon/v28y2009i2p413-426.html.

Chetty, Raj, Adam Looney, and Kory Kroft. 2009. "Salience and Taxation: Theory and Evidence." *American Economic Review* 99 (4):1145–77. URL http://www.aeaweb.org/articles.php?doi=10.1257/aer.99.4.1145.

Clotfelter, Charles T., Helen F. Ladd, and Jacob Vigdor. 2010. "Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects." *Journal of Human Resources* 45 (3). URL http://EconPapers.repec.org/RePEc:uwp:jhriss:v:45:y:2010:iii:1:p655-681.

Coleman, James S. 1966. *Equality of Educational Opportunity (COLEMAN) Study*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], icpsr06389-v3icpsr06389-v3 ed. URL http://dx.doi.org/10.3886/ICPSR06389.v3.

Conley, Timothy G. and Christopher R. Taber. 2011. "Inference with "Difference in Differences " with a Small Number of Policy Changes." *The Review of Economics and Statistics* 93 (1):113–125. URL http://ideas.repec.org/a/tpr/restat/v93y2011i1p113-125.html.

Constantine, J., D. Player, T. Silva, K. Hallgren, M. Grider, J. Deke, and E. Warner. 2009. "An Evaluation of Teachers Trained through Different Routes to Certification." Tech. Rep. NCEE 20094043, Institute of Education Science. URL http://ies.ed.gov/ncee/pubs/20094043/pdf/20094044.pdf.

Cruces, Guillermo and Sebastian Galiani. 2003. "Generalizing the Causal Effect of Fertility on Female Labor Supply." William Davidson Institute Working Papers Series 2003-625, William Davidson Institute at the University of Michigan. URL http://ideas.repec.org/p/wdi/papers/2003-625.html.

———. 2007. "Fertility and Female Labor Supply in Latin America: New Causal Evidence." *Labour Economics* 14 (3):565–573. URL http://ideas.repec.org/a/eee/labeco/v14y2007i3p565-573.html.

Datar, Ashlesha and Bryce Mason. 2008. "Do Reductions in Class Size Crowd Out Parental Investment in Education?" *Economics of Education Review* 27 (6):712–723. URL http://ideas.repec.org/a/eee/ecoedu/v27y2008i6p712-723.html.

Dee, Thomas S. 2004. "Are There Civic Returns to Education?" *Journal of Public Economics* 88 (9-10):1697–1720. URL http://ideas.repec.org/a/eee/pubeco/v88y2004i9-10p1697-1720.html.

Díaz-Cayeros, Alberto. 2008. "Electoral Risk and Redistributive Politics in Mexico and the United States." *Studies in Comparative International Development* 43 (1).

Dixit, Avinash and John Londregan. 1996. "The Determinants of Success of Special Interests in Redistributive Politics." *The Journal of Politics* 58 (4):pp. 1132–1155. URL http://www.jstor.org/stable/2960152.

Duflo, Esther. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review* 91 (4):795–813. URL http://ideas.repec.org/a/aea/aecrev/v91y2001i4p795-813.html.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review* 101 (5):1739–74. URL http://ideas.repec.org/a/aea/aecrev/v101y2011i5p1739-74.html.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2008. *Using Randomization in Development Economics Research: A Toolkit*, *Handbook of Development Economics*, vol. 4, chap. 61. Elsevier, 3895–3962. URL http://ideas.repec.org/h/eee/devchp/5-61.html.

Elacqua, G. and R. Fabrega. 2007. *El consumidor de la educación: El actor olvidado de la libre elección de escuelas en Chile*. Santiago, Chile: PREAL: Uso e impacto de la información educativa en Amárica Latina. URL http://www.ulavirtual.cl/courses/PSIC101305/document/consumidor_educacion.pdf.

Eskeland, Gunnar S. and Deon Filmer. 2002. "Autonomy, Participation, and Learning in Argentine Schools - Findings and Their Implications for Decentralization." Policy Research Working Paper Series 2766, The World Bank. URL http://EconPapers.repec.org/RePEc:wbk:wbrwps:2766.

Ferraz, Claudio and Frederico Finan. 2008. "Exposing Corrupt Politicians: The Effects of Brazilâs Publicly Released Audits on Electoral Outcomes." *The Quarterly Journal of Economics* 123 (2):703–745. URL http://ideas.repec.org/a/tpr/qjecon/v123y2008i2p703-745.html.

Foster, Andrew D. and Mark R. Rosenzweig. 1995. "Learning by Doing and Learning from Others: Human Capital and Technical Change in Agriculture." *Journal of Political Economy* 103 (6):1176–1209. URL http://ideas.repec.org/a/ucp/jpolec/v103y1995i6p1176-1209.html.

Fuchs, Thomas and Ludger Woessmann. 2007. "What accounts for international differences in student performance? A re-examination using PISA data." *Empirical Economics* 32:433–464. URL http://dx.doi.org/10.1007/s00181-006-0087-0. 10.1007/s00181-006-0087-0.

Galiani, Sebastian, Paul Gertler, and Ernesto Schargrodsky. 2008. "School Decentralization: Helping the Good Get Better, but Leaving the Poor Behind." *Journal of Public Economics* 92 (10-11):2106–2120. URL http://EconPapers.repec.org/RePEc:eee:pubeco:v:92:y:2008:i:10-11:p:2106-2120.

Garet, M., A. Wayne, F. Stancavage, J. Taylor, M. Eaton, K. Walters, M. Song, S. Brown, S. Hurlburt, P. Zhu, S. Sepanik, and F. Doolittle. 2011. "Middle school mathematics professional development impact study: Findings after the second year of implementation." Tech. rep., Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gertler, Paul, Harry A. Patrinos, and Marta Rubio-Codina. 2007. "Impact Evaluation for School Based Management Reform." *Doing Impact Evaluation. The World Bank.* (10).

Gertler, Paul J., Sebastian Martinez, Patrick Premand, Laura B. Rawlings, and Christel M. J. Vermeersch. 2011. *Impact Evaluation in Practice*. Washington, D.C.: World Bank, pap/psc ed.

Glazerman, Steven, Sarah Dolfin, Martha Bleeker, Amy Johnson, Eric Isenberg, Julieta Lugo-Gil, Mary Grider, and Edward Britton. 2008. "Impacts of Comprehensive Teacher Induction: Results from the First Year of a Randomized Controlled Study. Mathematics Policy Research Report." Mathematica policy research reports, Mathematica Policy Research. URL http://EconPapers.repec.org/RePEc:mpr:mprres:5666.

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. "Teacher Incentives." *American Economic Journal: Applied Economics* 2 (3):205–27. URL http://ideas.repec.org/a/aea/aejapp/v2y2010i3p205-27.html.

Glewwe, Paul and Hanan G. Jacoby. 2004. "Economic Growth and the Demand for Education: Is There a Wealth Effect?" *Journal of Development Economics* 74 (1):33–51. URL http://ideas.repec.org/a/eee/deveco/v74y2004i1p33-51.html.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. 2009. "Many Children Left Behind? Textbooks and Test Scores in Kenya." *American Economic Journal: Applied Economics* 1 (1):pp. 112–135. URL http://www.jstor.org/stable/25760149.

Glewwe, Paul William, Hanan Jacoby, and Elizabeth M. King. 2001. "Early Childhood Nutrition and Academic Achievement: A Longitudinal Analysis." *Journal of Public Economics* 81 (3):345–368. URL http://EconPapers.repec.org/RePEc:eee:pubeco:v:81:y:2001:i:3:p:345-368.

Gould, Eric D., Victor Lavy, and Marco Daniele Paserman. 2003. "Immigrating to Opportunity: Estimating the Effect of School Quality Using a Natural Experiment on Ethiopians in Israel." CEPR Discussion Papers 4052, C.E.P.R. Discussion Papers. URL http://ideas.repec.org/p/cpr/ceprdp/4052.html.

Greene, William H. 2003. *Econometric Analysis*. Upper Saddle River, NJ: Prentice Hall.

Hanushek, Eric A. 2011. "Valuing Teachers: How Much is a Good Teacher Worth?" *Education Next* 3 (11):40–45.

Hanushek, Eric A. and Javier A. Luque. 2003. "Efficiency and Equity in Schools Around the World." *Economics of Education Review* 22 (5):481–502. URL http://ideas.repec.org/a/eee/ecoedu/v22y2003i5p481-502.html.

Hanushek, Eric A. and Ludger Woessmann. 2007. "The Role of Education Quality for Economic Growth." Policy Research Working Paper Series 4122, The World Bank. URL http://EconPapers.repec.org/RePEc:wbk:wbrwps:4122.

———. 2012. "Schooling, Educational Achievement, and the Latin American Growth Puzzle." *Journal of Development Economics* 99 (2):497–512. URL http://ideas.repec.org/a/eee/deveco/v99y2012i2p497-512.html.

Harris, Douglas N. and Tim R. Sass. 2011. "Teacher Training, Teacher Quality and Student Achievement." *Journal of Public Economics* 95 (7-8):798–812. URL http://ideas.repec.org/a/eee/pubeco/v95y2011i7-8p798-812.html.

Hastings, Justine S., Thomas J. Kane, and Douglas O. Staiger. 2006. ". Gender and Performance: Evidence from School Assignment by Randomized Lottery." *The American Economic Review* 96 (2):pp. 232–236. URL http://www.jstor.org/stable/30034648.

Heckman, James J., Jora Stixrud, and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24 (3):411–482. URL http://ideas.repec.org/a/ucp/jlabec/v24y2006i3p411-482.html.

Hedges, L. V., R. D. Laine, and R. Greenwald. 1994. "Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes." *Educational Researcher* 23 (3):5–14.

Hedges, Larry V. and E. C. Hedberg. 2007. "Intraclass Correlation Values for Planning Group-Randomized Trials in Education." *Educational Evaluation and Policy Analysis* 29 (1):pp. 60–87. URL http://www.jstor.org/stable/30128045.

Hill, Carolyn J., Howard S. Bloom, Alison Rebeck Black, and Mark W. Lipsey. 2007. "Empirical Benchmarks for Interpreting E?ect Sizes in Research." Tech. rep., MDRC Organization.

Ho, Daniel E. and Imai Kosuke. 2008. "Estimating Causal effects of Ballot Order from a Randomized Natural Experiment: California Alphabet Lottery, 1978-2002." *Public Opinion Quarterly* 72 (2):216–240.

Hsieh, Chang-Tai and Miguel Urquiola. 2003. "When Schools Compete, How Do They Compete? An Assessment of Chileâs Nationwide School Voucher Program." NBER Working Papers 10008, National Bureau of Economic Research, Inc. URL http://ideas.repec.org/p/nbr/nberwo/10008.html.

Hyslop, Dean R. and Guido W. Imbens. 2001. "Bias from Classical and Other Forms of Measurement Error." *Journal of Business & Economic Statistics* 19 (4):475–81. URL http://ideas.repec.org/a/bes/jnlbes/v19y2001i4p475-81.html.

Kielstra, Paul. 2012. "The Learning Curve: Lessons in Country Performance in Education." Report, Pearson. URL www.thelearningcurve.pearson.com.

Krueger, Alan B. and Mikael Lindahl. 2001. "Education for Growth: Why and For Whom?" *Journal of Economic Literature* 39 (4):1101–1136.

Krueger, Alan B. and Diane Whitmore. 2000. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR." NBER Working Papers 7656, National Bureau of Economic Research, Inc. URL http://EconPapers.repec.org/RePEc:nbr:nberwo:7656.

Lavy, Victor. 2011. "What Makes an Effective Teacher? Quasi-Experimental Evidence." NBER Working Papers 16885, National Bureau of Economic Research, Inc. URL http://ideas.repec.org/p/nbr/nberwo/16885.html.

Lockheed, M.E. and E.A. Hanushek. 1988. *Improving Educational Efficiency in DeÂ¬velopingCountries : WhatDoWeKnow?WorldBankreprintseries.*

Luo, Renfu, Max Kleiman-Weiner, Ai Yue, Rey Martorell, Michelle Lee, Scott Rozelle, Linxiu Zhang, Chengfang Liu, Yaojiang Shi, and Brian Sharbono. 2010. "Anemia in Rural Chinaâs Elementary Schools: Prevalence and Correlates in Shaanxi Provinceâs Poor Counties." *Ecology of Food and Nutrition* 49 (5):357–372. URL http://www.ncbi.nlm.nih.gov/pubmed/21888576http://www.ncbi.nlm.nih.gov/pubmed/21888576.

Maddala, Gangadharrao S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge UK: Cambridge University Press.

Mas-Colell, A., M.D. Whinston, and J.R. Green. 1995. *Microeconomic Theory*. Oxford University Press New York.

McEwan, P. J. and Lucrecia Santibanez. 2005. "Teacher and Principal Incentives in Mexico." In Incentives to Improve Teaching: Lessons from Latin America. Washington, DC: World Bank Press, 213–53.

McEwan, Patrick J. and Joseph S. Shapiro. 2008. "The Benefits of Delayed Primary School Enrollment: Discontinuity Estimates Using Exact Birth Dates." *Journal of Human Resources* 43 (1). URL http://EconPapers.repec.org/RePEc:uwp:jhriss:v:43:y:2008:i:1:p1-29.

Metzler, Johannes and Ludger Woessmann. 2010. "The Impact of Teacher Subject Knowledge on Student Achievement: Evidence from Within-Teacher Within-Student Variation." IZA Discussion Papers 4999, Institute for the Study of Labor (IZA). URL http://EconPapers.repec.org/RePEc:iza:izadps:dp4999.

Miguel, Edward and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72 (1):159–217. URL http://ideas.repec.org/a/ecm/emetrp/v72y2004i1p159-217.html.

Mizala, Alejandra, Pilar Romaguera, and Miguel Urquiola. 2007. "Socioeconomic Status or Noise? Tradeoffs in the Generation of School Quality Information." *Journal of Development Economics* 84 (1):61–75. URL http://ideas.repec.org/a/eee/deveco/v84y2007i1p61-75.html.

Morokoff, William J. and Russel E. Caflisch. 1994. "Quasi-Random Sequences and Their Discrepancies." *SIAM J. Sci. Comput* 15:1251–1279.

Muralidharan, Karthik and Venkatesh Sundararaman. 2010. "The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India." *The Economic Journal* 120 (546):F187–F203. URL http://dx.doi.org/10.1111/j.1468-0297.2010.02373.x.

———. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy* 119 (1):39 – 77. URL http://ideas.repec.org/a/ucp/jpolec/doi10.1086-659655.html.

O'Brien, Peter C. 1984. "Procedures for comparing samples with multiple endpoints." *Biometrics* 40 (4):pp. 1079–1087. URL http://www.jstor.org/stable/2531158.

Patrinos, Harry Anthony, Cris Ridao-Cano, and Christos Sakellariou. 2006. "Estimating the Returns to Education: Accounting for Heterogeneity in Ability." Policy Research Working Paper Series 4040, The World Bank. URL http://EconPapers.repec.org/RePEc:wbk:wbrwps:4040.

Paxson, Christina and Norbert Schady. 2007. "Cognitive Development among Young Children in Ecuador: The Roles of Wealth, Health, and Parenting." *The Journal of Human Resources* 42 (1):pp. 49–84. URL http://www.jstor.org/stable/40057298.

Pritchett, Lant and Deon Filmer. 1997. "What Educational Production Functions Really Show: A Positive Theory of Education Spending." Policy Research Working Paper Series 1795, The World Bank. URL http://ideas.repec.org/p/wbk/wbrwps/1795.html.

Ramírez, María José. 2007. "Diferencias Dentro de las Salas de Clases. Distribución del Rendimiento en Matemáticas." *Puntos de Referencia* 284. URL http://www.cepchile.cl/dms/lang_1/doc_3941.html.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73 (2):417–458. URL http://ideas.repec.org/a/ecm/emetrp/v73y2005i2p417-458.html.

Rogoff, Kenneth. 1990. "Equilibrium Political Budget Cycles." *American Economic Review* 80 (1):21–36. URL http://ideas.repec.org/a/aea/aecrev/v80y1990i1p21-36.html.

Roland G. Fryer, Jr. and Steven D. Levitt. 2010. "An Empirical Analysis of the Gender Gap in Mathematics." *American Economic Journal: Applied Economics* 2 (2):pp. 210–240. URL http://www.jstor.org/stable/25760212.

Romeo, Isabella and Emanuela Raffinetti. 2012. *School Tracking and Equality of Opportunity in a Multilevel Perspective*. Cleup in Proceedings of the 46 scientific meeting of the Italian statistical society.

Rosenbaum, Paul R. 2002. *Observational Studies*. New York: Springer-Verlag, 2nd ed.

Rubin, Donald B. 1992. "Meta-Analysis: Literature Synthesis or Effect-Size Surface Estimation?" *Journal of Educational Statistics* 17 (4):pp. 363–374. URL http://www.jstor.org/stable/1165129.

Sawada, Yasuyuki and Andrew Ragatz. 2005. *Decentralization of Education, Teacher Behavior and Outcomes: The Case of the El Salvador EDUCO Program*. Incentives to Improve Teaching: Lessons from Latin America. The World Bank., 255.

Smith, Gary and Joanna Smith. 2005. "Regression to the Mean in Average Test Scores." *Educational Assessment* 10 (4):377–399.

Spence, A Michael. 1973. "Job Market Signaling." *The Quarterly Journal of Economics* 87 (3):355–74. URL http://ideas.repec.org/a/tpr/qjecon/v87y1973i3p355-74.html.

Thomas, Duncan. 1996. "Fertility, Education and Resources in South Africa." Tech. Rep. 96-15, RAND - Labor and Population Program. URL http://ideas.repec.org/p/fth/randlp/96-15.html.

Todd, Petra E. and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *The Economic Journal* 113.

Vegas, Emiliana and Jenny Petrow. 2008. *Raising Student Learning in Latin America: The Challenge for the Twenty-First Century*. The World Bank. URL http://EconPapers.repec.org/RePEc:wbk:wbpubs:6802.

Vegas, Emiliana and Ilana Umansky. 2005. *Improving Teaching and Learning through Effective Incentives. What Can We Learn from Education Reforms in Latin America?* The World Bank, Washington DC. URL https://openknowledge.worldbank.org/handle/10986/8694.

Weber, Karl, editor. 2010. *The Difference is Teacher Quality*. Waiting for Superman: How We Can Save Americaâs Failing Public Schools. New York: Public Affairs. 81–100.

Woessmann, Ludger. 2003. "European Education Production Functions: What Makes a Difference for Student Achievement in Europe?" European Economy - Economic Papers 190, Directorate General Economic and Monetary Affairs, European Commission. URL http://ideas.repec.org/p/euf/ecopap/0190.html.

Woessmann, Ludger and Thomas Fuchs. 2005. "Families, Schools, and Primary-School Learning: Evidence for Argentina and Colombia in an International Perspective." Policy Research Working Paper Series 3537, The World Bank. URL http://ideas.repec.org/p/wbk/wbrwps/3537.html.