GENERATING EVIDENCE IN EDUCATION

Impact Evaluations

Strong evidence is of central importance in informing policy and programming decisions across all agencies and organisations working with education systems in developing countries. Robust research and evaluation generates the evidence required to form judgments, deliberate options and make intelligent decisions about how to spend scarce financial resources. It is therefore vital that the evidence generated is based on the best available research derived from both observation and experimentation. Investments in what works in education are urgently needed. Programs taken to scale should be based on rigorous evidence.

This guide is the first in a series of publications on evaluations in education prepared for the Building Evidence in Education (BE^2) working group. It provides an introduction to: the importance of sound research to inform education policy, the ways in which to design impact evaluations and issues to consider when generating them.

The BE^2 Steering Committee*

Cover Design: JBS International, Inc.

Cover Photo: Curt Carnemark/The World Bank Group

* The Building Evidence in Education (BE^2) working group is led by a Steering Committee composed of the Department for International Development (DFID), United Nations International Children's Emergency Fund (UNICEF), United States Agency for International Development (USAID) and The World Bank Group.

Generating Evidence in Education – Impact Evaluations

Foreword

The Building Evidence in Education (BE^2) donor working group was launched in 2012 with the aim to engage bilateral and multilateral donors and foundations committed to:

- increasing the quality of education research;
- promoting the use of evidence in education programming; and
- strengthening donor research collaboration.

The working group is led by a Steering Committee composed of the Department for International Development (DFID), United States Agency for International Development (USAID), The World Bank Group and a rotating representative of the United Nations (UN) organizations, currently the United Nations International Children's Emergency Fund (UNICEF).

This series of Guidance Notes, prepared for the BE^2 working group by its respective members, provides tools and guidance for generating better evidence and leveraging existing evidence more effectively and efficiently. These Guidance Notes have benefited from the advice of BE^2 member organizations and are valuable tools for researchers and commissioners of research.

Chris Whitty	Josephine Bourne	Christie Vilsack	Claudia Costin
Chief Scientific Officer	Associate Director	Senior Advisor for	Senior Director,
DFID	Education	International Education	Education
	UNICEF	USAID	The World Bank Group

Prepared by Harry Anthony Patrinos (Manager, The World Bank Group) and Jess Cross (Junior Professional Associate, The World Bank Group) on behalf of the Building Evidence in Education (BE^2) working group. Based in part on Impact Evaluation for School-Based Management Reform by Paul Gertler, Harry Patrinos and Marta Rubio-Codina (World Bank), 2007, and Impact Evaluation of Private Sector Participation in Education by Laura Lewis and Harry Patrinos (CfBT and World Bank), 2012.

This page intentionally left blank.

Contents

1. Introduction 1
   a) Why Strong Evidence Matters 1
   b) Purpose of This Guide 1
   c) Education Challenges 1
   d) Assessing the Strength of Evidence Leads to Gaps 2
2. Deciding on Your Research 3
   a) Defining the Treatment 3
   b) Type of Research Question 4
   c) Unit of Analysis 5
   d) Indicators 5
   e) Data Sources 6
   f) Sample Size 8
   g) Timing and Duration of Evaluation 8
3. Designing an Intervention 9
   a) Methods 9
   b) Issues to Consider 9
      i. Non-experimental Designs 9
      ii. Quasi-experimental Designs 10
      iii. Experimental Designs 13
4. Conclusion 15

References 18

This page intentionally left blank.

1. Introduction

a) Why Strong Evidence Matters

Investments in what works in education are urgently needed. Strong evidence is of central importance in informing policy and programming decisions across all agencies and organisations working with education systems in developing countries. Robust research and evaluation generates the evidence required to form sound judgements, deliberate available options and make intelligent decisions about how to spend scarce financial resources. It is therefore vital that the evidence generated is based on the best available research derived from both observation and experimentation. Programmes taken to scale should be based on rigorous evidence and evaluation.

b) Purpose of This Guide

This guide provides a thorough introduction for donors, funders and practitioners of research to:

1) The importance of sound research to inform education policy,
2) The ways in which to design impact evaluations and
3) Issues to consider when generating evidence.

Several technical issues and terms will be referenced throughout this guide in order to provide an accurate portrayal of the processes and challenges involved with the production of sound evidence. Additional definitions and clarifications of many of the terms referenced in this guide can be found in Part I of the two-part guide, *Assessing the Strength of Evidence in the Education Sector* (Hinton 2015). Although there is no one-size-fits-all evaluation method or general hierarchy of methods, some methods are better than others at addressing certain areas and issues. It is recommended that this guide be used to inform conversations with evaluators and research partners, and that the expertise of a professional be utilized to ensure the evaluation meets all needs and expectations while producing rigorous evidence.

c) Education Challenges

Education is a critical driver of economic growth and poverty reduction as education systems help expand knowledge and promote skills that propel individual labour productivity. In order to encourage such growth, it is imperative to bridge the gap in access to schooling. Education affects people's lives on various levels, their participation in economic activities, and overall economic development in various ways. A person without basic literacy and numeracy skills finds it difficult to master the skills of everyday life; the lack of basic education has always been accepted as one of the major components of any multidimensional concept of poverty.

In 1999 more than 105 million children were out of school. As of 2012, this number had been reduced to 58 million (United Nations Educational, Scientific and Cultural Organization [UNESCO] Institute for Statistics 2014). Countries have come a long way in improving access to

education. Still, enrolment rates remain low in several developing regions, and access to education is proving to be elusive for many. Often, low-income families, girls, indigenous peoples and disadvantaged groups have limited access to schooling, and several African and Asian countries have yet to achieve universal primary coverage.

Furthermore, the quality of education globally, as measured by student performance on standardised tests, is low and represents a major challenge and a learning crisis. Most students from developing countries who participate in international assessments score poorly. For example, developing countries that participate in the Programme for International Student Assessment (PISA) almost always rank at the bottom (http://www.oecd.org/pisa/). On average, only 50 per cent of students in developing countries demonstrate proficiency in reading that is at or above the baseline, needed to be effective and productive in life. By comparison, 81 per cent of students in the Organisation for Economic Co-operation and Development (OECD) countries performed at or above the baseline level in reading.

These numbers demonstrate that the average child from a developing country who took part in PISA is 40 to 50 points – half of a standard deviation, or two full years in terms of learning – behind the *worst students* in the economic superstar countries. Even the best performers from developing countries – the top 5 per cent in science – were almost 100 points behind the average child in Singapore, 83 points behind the average Korean and a staggering 250 points behind the best of the best.

The benefits and long-lasting effects of schooling are many, including poverty reduction, equity enhancement, promotion of rights, gender equity, increased child education, improved child and own health, informed fertility decisions, job search efficiency, ability to adapt to technological change, social cohesion and crime reduction, among many others. It is a universal fact that the more education a child receives, the higher his or her earnings will be, with the global average rate of returns to schooling estimated at 10 per cent (Montenegro and Patrinos 2014). However, the positive benefits of education fail to reach children if they are excluded from school or if they do not learn while in school. There are multiple sources of disadvantage that make it difficult for many children to benefit from a high-quality education. These disadvantages may include gender, socio-economic status, geographic location, disability and ethno-linguistic background. For example, in Guatemala, the illiteracy rate among indigenous women stands at 60 per cent, 20 percentage points above indigenous men and twice the rate of non-indigenous women (World Bank 2012). In Pakistan, a half-kilometre increase in the distance to school decreased female enrolment by 20 percentage points (World Bank 2012).

d) Assessing the Strength of Evidence Leads to Gaps

Reliable and well-conducted evaluations of programmes that can lend empirical support to the various claims on the advantages of education interventions are needed, and more so as more countries are adopting these reforms. Development agencies, policymakers and government officials need to know and understand what works in education and why. Rigorous programme evaluations can serve several purposes. First and foremost, evaluations determine whether a

programme has had an impact on the target population and how large that impact truly is. Evaluations also enable researchers to determine whether a programme design was effective, or which specific areas of the design could benefit from improvements and which could be scaled up. Third, they hold policymakers, development agencies and schools accountable. Lastly, evaluations confirm whether resources were allocated and spent as originally intended.

The world of international education development lacks rigorous, quality research capable of informing policy and decision makers, investors, development agencies and even schools themselves. Research in education is critical to ensure that investments made at the school level are both successful and cost-effective. Programmes that do not work to improve access to education or the quality of education should not be brought to scale. However, this information can only exist after extensive evaluation is performed. Due to the current lack of quality research, there is a need for original research to be performed and disseminated so that children worldwide can continue to benefit from education system reforms.

A reliable impact evaluation is challenging to design and implement. This guide recommends three key foundations:

1. An *appropriate model of behaviour or theory of change* that provides a guide towards the development of hypotheses on the expected benefits of the intervention in question.
2. *Detailed data* over an *appropriate period of time* that measures the response of beneficiaries to the programme.
3. An *identification strategy* that allows the measurement of a *counterfactual* – what would have happened to beneficiaries without the programme – to enable changes in outcomes to be attributed to the programme (Gertler et al. 2007).

2. Deciding on Your Research

An intervention in education development responds to a pressing problem, such as low enrolment, low levels of learning, poor school management and so on. The programme is then designed to combat those issues with an intended positive effect. Education programmes often appear promising at their outset, but over time they may not generate the expected or intended outcomes. Thus, an impact evaluation can be implemented to determine whether programmes are bringing about the anticipated effects and to better understand what works, what does not work and how the changes observed can be attributed to a particular project, individual policy or intervention subcomponent. Impact evaluations should determine the instruments through which beneficiaries are responding to the intervention (Khandker et al. 2010).

a) Defining the Treatment

The first step is to determine what intervention will be evaluated. This becomes the treatment. The treatment variable is also known as the independent variable, which in an evaluation of a

programme is the variable being manipulated or changed in order to observe effects. For example:

- A World Bank project in Indonesia helped to meet the persistent challenge of stunted growth in children, which then effected the likelihood of completing basic education, with an intervention that included raising community awareness of the importance of early childhood education and development.
    - *Treatment:* Community awareness programme

- Bangladesh has recently experienced high demand for skilled labour in an economy where the skill base of workers is limited. An intervention seeks to increase the skills of the workforce through a quality vocational training intervention.
    - *Treatment:* Vocational training for students

- In Nigeria, where in 2011 it was estimated that 11 million children were still out of school, education management information systems were targeted and strengthened as part of a sector-wide intervention. (http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTEDUCATION/0,,contentMDK:20270003~menuPK:4679417~pagePK:210058~piPK:210062~theSitePK:282386,00.html)
    - *Treatment:* Education management information system strengthening programme

The treatment is only given to participants in the experimental group (the treatment group), which will be further explained.

In defining the treatment, it is critical to understand what it means for a community, school, parent, principal, teacher or student to participate in a programme. Various statistical interactions can be used to analyse results that take into account treatment status (under treatment or not under treatment) and length of treatment (has the subject received treatment for all years of the programme or just some, was subject phased in, etc.), and determine whether the intervention changed in any ways over the course of the programme cycle.

b) Type of Research Question

The next step in conducting a rigorous evaluation of an education intervention is to determine whether the research question will focus on a programme evaluation, a process evaluation or a combination of the two types of evaluations.

Programme evaluations examine the effect of the intervention as a whole on outcomes as opposed to not receiving any intervention. They are beneficial in determining the overall success of an intervention and provide basic information regarding the magnitude of impact. A more detailed understanding of which particular aspects of an intervention are affecting which outcomes is sometimes necessary and can be examined through process evaluations.

Process evaluations attempt to identify the individual mechanisms through which the programme intervention is affecting outcomes to understand the programme more fully (Lewis and Patrinos 2012). This approach requires the identification of distinct programme components and examines the causal relationships of each subcomponent separately. Although process evaluations require greater data and the identification of a valid counterfactual (explained in greater detail below) for each subcomponent, they are more comprehensive than programme evaluations and allow policymakers to determine which aspects of the intervention should be scaled up or replicated in other areas. Several treatment variables are defined through this approach.

The identification of a counterfactual is used to measure what would have happened to the beneficiary population in the absence of an intervention (Lewis and Patrinos 2012). Counterfactuals are critical to understanding the effects of receiving the intervention relative to a population that did not receive the intervention. A valid counterfactual should be as similar to the treatment population as possible in order to most accurately determine causal linkages and dismiss the possibility of other variables having an effect on outcomes.

c) Unit of Analysis

Once a treatment and related programme goals have been defined, it is necessary to determine the unit of analysis – that is, who or what will be studied to determine the effects of an intervention. For example, the natural unit of analysis of a conditional cash transfer programme is the student, who will be directly impacted by those cash transfers. In examining the engagement of the private sector, the unit of analysis for a school voucher programme would also be the student. However, the unit of analysis in the evaluation of a school-based management intervention might be the school (or the teacher or director or student as well). In this case, all members of the school community – students, teachers, principals, teacher aides and parents – are likely to benefit from the reform, some more directly than others, and thus a further analysis can be conducted at each of those levels.

d) Indicators

The unit of analysis, whether it be the school, the student or the teacher, will determine the indicators through which to measure the effects of the intervention. It is important that an indicator be measured both before and after treatment to accurately measure any impacts of the programme. It is critical that an indicator reliably monitor progress toward the intended outcomes, and thus it will vary from programme to programme. Depending on the unit of analysis, indicators may focus on teacher absenteeism or motivation, principal/teacher/student relationships, community attitudes toward education, school transition rates, gender ratios, or other factors. One common indicator used to measure the effects of an intervention is student learning, which can be measured through standardised exams.

Student achievement outcomes are the most common and obvious of indicators through which to measure the effects of a programme. For example, student achievement is examined in a study

on the effects of increasing choice in schooling (Loeb et al. 2011). Standardised exam results can be used to gather this information, both for individual students and for entire schools. Caution should be taken when using existing student scores as a measure of progress. Student learning may be difficult to use as an indicator in areas where schools do not administer exams or where data is not readily available. Additionally, the data available is often not reliable or comparable, as test scores may not be representative of the sample under study or the data is not comparable over time if exams have changed over the years. Reading and math scores may be collected as part of the evaluation, though this process can be costly and time-consuming.

Intermediate quality of education outcomes can be measured to determine impacts on students, such as repetition and drop-out rates, failure, acceleration and retention. Drop-out rates, which were examined in a study by Barrera-Osorio (2007) on the effects of Bogotá's Concession Schools, are an indicator to measure the ability of a school to retain students, while transition rates from one grade to the next can also measure the ability of a school to serve its students' needs and ensure that they are learning what is required to pass a grade level. Grade repetition indicates that students are not learning, as failures indicate as well. Lastly, researchers can examine data that details the age of students, as students who are older than their peers at grade level are often more likely to drop out. These indicators can be measured at the school level or disaggregated to examine impacts by gender, socio-economic status, age or ethnicity.

School access outcomes can be measured to examine the rates at which students are enrolling and attending school. This indicator can be measured through several variables, such as total school enrolment rates, or by disaggregating the data to examine effects by gender, proximity to school and so on. School attendance can also be measured, as those who are enrolled in school may not actually be attending.

Teacher performance outcomes demonstrate impacts of an intervention on educators, which may directly impact students in the classroom. Bruns and Luque (2014) examined the effects of several reforms on teacher policies on teacher performance in their recent study. Several variables can be measured to examine programme effects on teachers, such as teacher effort and overall performance. Surveys are an effective means of measurement and often ask respondents to detail teacher–student interactions, teacher–teacher interactions, teacher motivation and drive, as well as absenteeism. As these areas will likely be unreliably reported through self-measurement, it is possible to ask principals, parents and students themselves to respond to these questions. For example, because teacher absenteeism is likely to be underreported by the teaching staff, other members of the school system can respond to this question by answering how many days of school the student missed over a period of time due to teacher absence (Jimenez and Sawada 1999).

e) Data Sources

For the purposes of this guide, data sources will be described as being in one of two categories, quantitative or qualitative. There are often quantitative and qualitative methods for data collection that can play equally important roles in evaluation. Mixed methods studies are being

used more frequently, though it is understood that the different approaches to evaluation need to be better integrated (Baker 2000; Adato 2011).

Evaluations can be retrospective, occurring after the conclusion of the intervention, or prospective, designed prior to the implementation of the intervention. If using a retrospective evaluation, the existence of sound data will be critical to the validity of the estimation of effects and will also dictate which method(s) to apply. Since original data is not gathered when using retrospective evaluation, this method can be cost-saving. Prospective evaluations may not be as cost-effective, though a treatment and control group can be defined at the outset of the programme and will allow for the collection of original data.

Quantitative data can be expressed numerically and used in statistical analyses to illustrate trends or to explore causal or correlational relationships. For the purpose of an evaluation of a programme, existing data can often be a useful and cost-effective way to measure impacts, though original data can also be collected. Regardless of the data sources being used, it is critical to cull administrative data on other related interventions, especially if those programmes are active in the same schools or districts, so that they can be controlled for prior to the intervention. It is additionally important to ensure that all data sources can be pulled together cohesively and that they are able to link to one another.

School and population data are often collected through national censuses and may be available for use. Existing school data can contain information on repetition, drop-out, failure, retention, students with a disability, school history and basic demographic information of the student population. Student exam scores may also be available, both at the individual level and at the school or district level.

If using original data, researchers should collect those data prior to, during and following the intervention. Relying on and building local capacity for data collection is important, regardless of the intervention or evaluation type. Original data can be gathered through a series of surveys that are administered to school staff (including principals, teachers, aides), the students themselves, parents and community members (Gertler et al. 2012). Questionnaires typically include sections on basic demographic information (such as age, ethnicity, gender, disability, socio-economic status), educational attainment and school history, parental education attainment, self/household labour force participation, basic information about the school and/or household infrastructure, time use and health, among other areas – all dependent on whom the questionnaire is targeting and for what purpose.

Data collected during the intervention should include information on treatment and control groups, including relevant criteria for treatments groups and total take-up, and also document any treatment-linked disbursements (monetary or not). Timing should also be noted, including for various phases of the project implementation and data collection.

Qualitative data are descriptions used to categorise or classify and do not involve numerical values. Qualitative data allow researchers to explore why an effect was found and do not rely on

a counterfactual to make a causal inference (Bamberger et al. 2012). Qualitative data allow researchers to better appreciate the nuances of an intervention as understood by those who are being impacted by the intervention or perhaps those receiving no benefits. Qualitative methods include interviews, focus groups and observation (direct and participant).

Direct observation requires the systematic recording of all witnessed activities, behaviours and changes in a setting while being as unobtrusive to others and to the environment as possible. The recorder of these observations is not involved in the study and therefore can provide an unbiased snapshot of the intervention. This can be done by keeping record of activities while physically present or by video recording the classroom proceedings to be viewed at a later time, as is done in many schools in Latin America (Bruns and Luque 2014). This is a cost-effective and relatively quick way to obtain primary data.

f) Sample Size

Sample size is a critical determinant of the original research design. The goal is to select a sample, or portion of the population, to study that is large enough to have significant statistical power and to be representative of the population as a whole. Because large sample sizes can be costly, it is important to strike a balance between representativeness and cost.

It is largely agreed upon that a common method, the cluster-based randomised control trial with a programme at the school level, will require a sample size of 40 to 50 schools (unit of treatment) with 40 to 60 students at each school on whom impacts can be measured, with impacts measured on their teachers and families as well (Gertler et al. 2007). This will allow for large enough treatment and control groups to see an effect and ensure power and significance. A sample of this size will allow for the observation of differences in student achievement exam scores between 0.10 and 0.25 standard deviations (Bloom et al. 1999; Raundenbush et al. 2004).

g) Timing and Duration of Evaluation

As with all aspects of designing original research, the timing and duration of an evaluation will depend upon several factors and may vary from evaluation to evaluation. These factors include the intervention itself, data sources and sampling units, and indicators being measured, among others.

If basing timing on the indicators chosen, student achievement outcomes should be measured pre-intervention and then again after two complete school cycles so as to ensure that test scores have adequate time to react to the new programming – for example, school-based management, which can take some time to produce a visible effect (King and Behrman 2009; Borman et al. 2003). Often schools themselves will capture what a student has learned over the course of each year, which data can also be utilised.

Intermediate quality of education outcomes, such as drop-outs and grade repetition, can and should be measured twice each year during which the intervention is implemented, once at the

beginning of the school year and again at the end of the school cycle. School access outcomes can also be measured at the beginning and end of the school cycle, though attendance should be measured at several consistent intervals throughout the year. This information is often collected by schools themselves.

3. Designing an Intervention

a) Methods

A research design is a framework in which a study is undertaken. It employs one or more research methods to (a) gather data and (b) analyse data and can include both quantitative methods and qualitative methods. There are three key research designs employed in a programme evaluation: non-experimental (observational), quasi-experimental and experimental. The design will ultimately determine the targeting of beneficiaries as well as the measures of impact which, in the case of experimental and quasi-experimental designs, are determined by the presence of a counterfactual. Non-experimental designs do not rely on a counterfactual for this purpose. When referring to a counterfactual, the term 'comparison group' is used in quasi-experimental designs; the term 'control group' is used in experimental designs.

Beneficiaries can be targeted several ways, depending on the research design, funding available, information readily available and several others factors. The overarching issues to consider in discussing the strengths and weaknesses of research design category are described below.

b) Issues to Consider

i. Non-experimental Designs

Non-experimental designs (also known as observational designs) encompass a wide range of valid empirical methods. Some non-experimental designs aim to explore causal relationships. The key distinction of non-experimental designs, relative to experimental designs, is that the researcher does not assign subjects to a treatment or control group to determine the effects of an intervention on a group (Hinton 2015).

Self-selection bias occurs in any evaluation in which an individual, school or teacher makes the decision to participate in the programme being offered. In theory, those who decide to participate can then be compared with those who did not choose to participate to determine the effects of the intervention. However, there are many reasons why someone may choose to participate in a programme or not that could present a bias and render a comparison of participants to non-participants an invalid evaluation (Berk 1983). For example, schools that are wealthier or better informed, parents with a higher socio-economic status or teachers with high levels of education themselves will be more likely to participate in an intervention than others. Additionally, the characteristics associated with participation are also associated with outcomes in that those who

self-select into a programme are often more likely to accept changes, perform well compared to others and be high achievers who will invest additional time, effort and (where available) resources to ensure that success is achieved.

Non-random programme placement can be found when the programme participants are not randomly selected, introducing an important source of endogeneity. Rather, participants are specifically chosen to receive an intervention. This can be seen at both ends of the spectrum, in low-performing schools in need of additional support and in high-functioning schools that will be quick to respond and show improvements. Whatever the reason for endogenous programme placement, biases will persist and the estimates of effects will be skewed. Programme effects might still be estimated by controlling for the biases associated with endogenous programme placement. One such method for estimating causal relationships in the absence of randomisation is the instrumental variables (IV) method, which relies on a variable that has an effect on participation in an intervention but not an effect on the outcomes of the programme. This method then controls for the endogeneity in the instrumental variable, though valid instrumental variables are not easily found (Heckman 1979).

ii. Quasi-experimental Designs

Quasi-experimental designs assign participants to either a treatment group (which receives the intervention) or a control group (which does not receive the intervention). They do not use randomisation for assignment into either group, reducing the confidence with which the effects of a programme can be accurately determined. These studies use statistical analyses to control for potential biases resulting from non-randomisation. They use non-beneficiary populations similar to those treated to create a valid counterfactual. They often require data on both the treatment and comparison groups for this purpose (Hinton 2015).

When an intervention has partial coverage – that is, not everyone in the community, school or population is receiving treatment – a comparison group can be created from those who do not participate in the programme that are most similar to the beneficiary group. This sorting can either be done prospectively, prior to the start of the intervention, or retrospectively, following the administration of the intervention. However, biases will arise as take-up of the intervention will not have been random and may affect observed outcomes.

There are research methods that attempt to negate the effects of non-randomisation for an evaluation. A few of these are described below.

Randomised promotion or encouragement design is a special case of an experiment that can be used in situations with little control over subjects' compliance (Gertler et al. 2011). The main idea is that instead of randomising the application of the intervention itself, what is randomised is encouragement to receive the treatment. By randomising encouragement and carefully tracking outcomes for all those who do and do not receive the encouragement, it is possible to obtain reliable estimates of both the encouragement and the intervention itself (Diamond and Hainmueller 2007). Encouragement may take the form of information that is additional to

whatever is already part of programme implementation and targeted at student or parent level. Some subjects receiving encouragement may not follow through with the programme. Others who do not receive encouragement may nevertheless access the programme. All that is required is that the encouragement increases the likelihood that units will follow through with what they are being encouraged to do.

Encouragement Designs
Promoting Education Infrastructure Investments in Bolivia (Gertler et al. 2011)

In 1991, Bolivia institutionalised and scaled up a successful Social Investment Fund (SIF), which provided financing to rural communities to carry out small-scale investments in education, health and water infrastructure. The World Bank, which was helping to finance SIF, was able to build an impact evaluation into the programme design.

As part of the impact evaluation of the education component, communities in the Chaco region were randomly selected for active promotion of the SIF intervention and received additional visits and encouragement to apply from programme staff. The programme was open to all eligible communities in the region and was demand driven in that communities had to apply for funds for a specific project. Not all communities took up the programme, but take-up was higher among communities where it was promoted.

Newman and others (2002) used the randomised promotion as an instrumental variable. They found that the education investments succeeded in improving measures of school infrastructure quality, such as electricity, sanitation facilities, textbooks per student and student–teacher ratios. However, they detected little impact on educational outcomes, except for a decrease of about 2.5 per cent in the drop-out rate. As a result of these findings, the Ministry of Education and the SIF now focus more attention and resources on the 'software' of education, funding physical infrastructure improvements only when they form part of an integrated intervention.

Propensity score matching (PSM) uses non-beneficiary characteristics to create treatment and control groups, though participants are not randomised. The propensity score is the probability of participation in the treatment on the basis of observed characteristics (Dehejia and Wahba 2002). Participants in a programme are then matched to non-participants on the basis of this propensity score to create a treatment group and a comparison group (Krueger and Zhu 2004). There are several challenges and limitations with matching methods such as PSM. The first issue arises when participation in treatment is determined by factors not observed by researchers. Those selection biases cannot be controlled for, leading to an overestimation of programme impact. Secondly, the researcher needs detailed data for both participants and non-participants, as well as information on observable characteristics for each. The more information and data collected, the greater the validity, but also the more difficult it becomes to find a match that meets all required criteria.

Differences-in-differences, or double difference (DD), is another research method used in quasi-experimental designs not requiring matching techniques. DD methods compare baseline and follow-up data for treatment and comparison groups pre- and post-intervention, as was done in the case of academy schools in England (Machin and Vernoit 2011). The mean difference between the 'after' and 'before' values of the outcome indicators for each of the treatment and comparison groups is calculated, followed by the difference between these two mean differences. The second difference is the estimate of the impact of the intervention. The advantage of DD methods is that empirical techniques can be used to difference out biases arising from all time-invariant observed and unobserved factors that could determine participation in a programme. One issue that may arise from DD methods relates to time-varying characteristics that may affect participation in the programme. To minimise this possible bias, these time-variant factors must be controlled for as much as possible, with individual time trends for treatment and comparison groups in the estimation of effects (Bertrand et al. 2004; Bell et al. 1999; Athey and Imbens 2006).

Programme phase-in in schools and districts can create a valid counterfactual. With this methodology, all schools within a community or all districts within a region will eventually receive treatment as determined by a regimented timeline that allows some participants to receive the treatment first, while others join at a later time (Marcus and Berman 2013; Khandker et al. 2010). The schools that are not yet receiving treatment thus become the counterfactual, so long as there are no differences in characteristics or political influences that determined which schools are phased in when and why. Matching methods or, if longitudinal information is available, DD methods can be applied for the analysis.

Phase-in Over Time
Cambodia: Challenges in Scaling Up Preschools (Marcus and Berman 2013)

With assistance from the World Bank and the Education Fast Track Initiative Catalytic Fund (now the Global Partnership for Education), the Government of Cambodia sought to improve early childhood programmes for the rural poor by expanding and evaluating three early childhood development options: formal preschools run by the Ministry of Education, informal community-based preschools and home-based programmes. The latter two were being piloted by United Nations Children's Fund (UNICEF) and Save the Children Norway in a few provinces.

To scale up the programme, the government decided that communities with an existing primary school that needed to be upgraded and/or expanded would receive a formal preschool as part of the renovations. Communities that didn't qualify for renovation work but had a high poverty rate and a large number of children under age five would get a community-based preschool or a home-based programme aimed at improving parenting practices.

The impact evaluation was designed to help the government determine which preschool model worked best, and researchers relied on randomisation to identify impact. Randomisation was implemented separately for the formal and informal preschool programmes because the criteria used to select participating communities differed. In practice, the experimental design included

five groups: three treatment groups (for each of the three interventions) and two control groups (one for the formal sample and one for the informal sample). Randomisation made it possible to compare outcomes in the different communities to determine the impact of each intervention.

To create a control group for the formal preschools, researchers relied on the programme's phase-in timeline. Not all schools could be renovated at the same time, so 19 were picked to be upgraded in the third year of the programme, allowing them to be used as a control group. Baseline surveys in December 2008 and endline surveys in June 2011 were conducted in 26 treatment villages and the 19 control villages, for a total of 1,553 households. To evaluate the informal preschool models, 450 villages were randomly selected in 10 provinces and were equally divided between the control group and the two informal models. Researchers surveyed 32 randomly-selected villages in each group, for a total of 3,807 households. The baseline was collected in May 2008 and the endline, in January 2011.

Randomised Phase-in:
Mexico: Programa de Educación, Salud y Alimentación (Khandker et al. 2010)

PROGRESA (originally known as *Oportunidades*) combined regional and village-level targeting with household-level targeting within these areas. Only the extreme poor were targeted, using a randomised targeting strategy that phased in the programme over time across targeted localities. One-third of the randomly targeted eligible communities were delayed entry into the programme by 18 months, and the remaining two-thirds received the programme at inception. Within localities, households were chosen on the basis of a discriminant analysis that used their socio-economic characteristics (obtained from household census data) to classify households as poor or non-poor. On average, about 78 per cent of households in selected localities were considered eligible, and about 93 per cent of households that were eligible enrolled in the programme.

Regarding potential ethical considerations in targeting the programme randomly, the phased-in treatment approach allowed all eligible samples to be targeted eventually, and it permitted the flexibility to adjust the programme if actual implementation was more difficult than initially expected. Monitoring and operational evaluation of the programme were also key components of the initiative, as was a detailed cost-benefit analysis.

iii. Experimental Designs

The best mechanism of guaranteeing a proper counterfactual and unbiased evaluation is randomisation. This method gives all an equal chance of being in the control or treatment group. It guarantees that all factors and characteristics will be on average equal between the two groups. The only difference is the intervention. Experimental research designs randomly assign subjects to either a treatment or control group in order to determine the effect of a programme. Those in the treatment group receive the intervention, while those in the control group do not. Random allocation to either group helps to ensure validity and increases the probability that effects are

due only to the intervention itself, eliminating other variables. Effects are then statistically calculated by comparing the observed outcomes of the two groups. This allows for a robust counterfactual (Hinton 2015).

Randomised control trials are often considered the 'gold standard' of programme evaluations and are one example of experimental design. Causal inference is possible in these studies where beneficiary and non-beneficiary groups are almost identical and randomly assigned. Differences in outcomes between the treatment and control groups can be simply calculated to determine the impacts of the intervention. However, there are several limitations that must be considered.

A randomised control trial has three key stages:

- The identification of a group of beneficiaries with similar characteristics
- Random assignment of subjects to treatment groups (who will receive the intervention) and control groups (who will not receive the intervention)
- The identification and manipulation of an independent variable (such as providing safe transportation to school, increasing the school day)

Randomised designs can assign individual subjects (such as students or teachers) to treatment or control groups, or they can assign entire groups (such as schools or districts) to treatment or control groups. This is known as cluster-based randomised design, and it is often the most effective way to minimise potential biases.

Though randomisation attempts to prevent bias, spillover effects, partial compliance and randomisation bias may still occur. Partial compliance, for example, occurs when those who were offered the intervention as part of the treatment group decline to take the treatment. However, biases occurring under randomisation are often more easily controlled than biases occurring from nonrandomised designs (Gertler et al. 2007).

Randomised control trials may be costly. They require the design of a rigorous evaluation pre-intervention, which may involve lengthy negotiations regarding the design and implementation of the intervention itself, the design of the study, the timing and the duration of the evaluation. External validity is another concern, as randomised designs frequently require large sample sizes that should be observed over long periods of time in order to accurately and reliably determine treatment effects. Estimates of the impacts of an intervention based on a randomised control trial may not necessarily be generalisable to the larger population. It is possible that a widespread reform may change the economic environment enough to invalidate the predictions of the experimental setup.

Randomised control trials are not exempt from sample selection bias (Shadish et al. 2002). This can occur when the randomisation of subjects is not upheld throughout the study. Initial allocation to treatment and control groups is random, but the actual evaluation is not random in practice. As a result, the average effect of the intervention on the randomly assigned treatments differs from the average treatment effect on those who actually participated in the intervention.

This can occur when a school not initially intended to participate and assigned to the control group faces pressure from the local government to be involved in the programme, or when a parent decides not to allow a child initially chosen as part of the treatment group to participate in a programme. Schools in the treatment group may also simply decide not to take up the intervention due to perceived benefits (self-selection).

Attrition bias, whereby the number of participants in a treatment or control group decreases over the course of the intervention, can have unintended effects on overall intervention outcomes. If, due to the aforementioned reasons, a teacher, student or school administrator chooses to withdraw from a treatment or control school during the course of the programme, and the reason for withdrawal is the school's treatment status, then an attrition bias is present. These biases can be controlled for by modelling selection into the sample group as a function of past observable characteristics and using the predicted probabilities to weigh observations in the outcome equation (Heckman 1976; Moffitt et al. 1999).

Spillover effects can be seen when an intervention inadvertently has an effect on the control population, such as when a family leaves one school district and moves to another, or when a treatment inadvertently helps the control group as well. In these situations, the comparison group is no longer a valid comparison group. This can be controlled for by ensuring that treatment and control schools are far enough apart to avoid frequent migration of students and staff and by randomising treatment across communities rather than simply across schools.

Substitution bias is well documented in actual experiments by Heckman and Smith (1995). The randomisation bias occurs when members of the control group receive some form of treatment, skewing the outcomes of an evaluation. For example, subjects in a control group may receive a conditional cash transfer as part of a separate intervention or receive nutritional supplements from a different programme. It is important to understand which other school programmes are occurring in an area and which may undermine the validity of an evaluation. The effects of this bias can be mitigated by controlling for other existing programmes from the outset.

4. Conclusion

Strong evidence is of central importance in informing policy and programming decisions across all agencies and organisations working with education systems in developing countries. Robust research and evaluation generates the evidence required to form judgements, deliberate options and make intelligent decisions about how to spend scarce financial resources. It is, therefore, vital that the evidence generated is based on the best available research derived from both observation and experimentation. Investments in what works in education are urgently needed. Programmes taken to scale should be based on rigorous evidence.

Other issues to consider when generating evidence are those of ethics. Randomised designs may present ethical constraints as interventions often provide increased benefits to a school, teacher or student. Government officials may be reticent to withhold an intervention from schools

equally in need, though some programmes do favour randomisation better than others. Randomised phase-in, wherein all schools in a district or districts in a region gradually receive the intervention, can provide a potential solution to this issue. With this design, schools are phased in over time so that an appropriate control can be established with some groups while others receive treatment. Eventually, once enough time has passed that effects can be observed with those initially treated, control groups also benefit from the intervention.

Programme evaluations might raise ethical concerns when they involve the collection of sensitive personal data on students. Denying benefits on methodological grounds is a sensitive issue and should be handled carefully and done only in situations where there is great benefit and where the control group will eventually receive the treatment, and only if the treatment actually has a positive benefit. In such cases data collection procedure requires clearance from a protection of human subjects board that guarantees protection of the subject's identity and the consent of the subject to participate in the study.

Perhaps one of the most infamous of all ethical cases is the Tuskegee syphilis study. In 1932, 623 African American men enrolled in a study on the effects of syphilis, and more than half of them were infected with the sexually transmitted disease. In exchange for their participation in the study, the men were given free medical exams and treatment, as well as payments to cover their burial expenses following their death. The study continued for 40 years, until 1972, when it was found to be unethical on many grounds. Though penicillin was found to be a viable treatment for syphilis in the late 1940s, it was not administered to the patients, several of whom died. Only African American men were asked to participate in the study, despite the fact that this disease was in no way attributed to race. Participants, many of whom were uneducated and poor, were not well informed of their disease, treatment options or even the purpose of the study (Gray 1998). Such cases led to the creation of courses, certification and other controls that guarantee the safety and identify of people in studies (see, for example, https://ethics.od.nih.gov/training.htm).

When conducting rigorous research, one cannot ignore political economy or politics. There are strong stakeholder groups with vested interests in education. Governments, which are often the funders, main providers and regulators of education programmes, have strong interests. Finding systematic ways to overcome these political economy hurdles is challenging (Grindle 2004). Information can play a critical role in paving the way for reform and evidence-based research, however (Khemani 2005, 2007; Majumdar et al. 2004). The policy process can be fed by credible public information on inputs and outcomes so that progress can be monitored transparently (Bruns et al. 2011).

Lastly, context plays an in important role in any evaluation as it will drive each aspect of the process, from the initial programme design to definition of the treatment, to determining the proper indicators and evaluation size. The implementation of an evaluation will thus vary from intervention to intervention depending on the local context, which will ultimately underlie the study and its results. Sound evaluations must take this element into consideration before, during

and following the study and should include reference to the local context in all formal write-ups and discussions.

Rigorous studies can help improve student learning outcomes. A greater number of rigorous studies will increase our understanding of the types of interventions that improve educational outcomes. Randomised studies require fewer assumptions and reduce biases, thus allowing researchers to produce robust findings. In cases where a full or national pilot randomised trial cannot be undertaken, the information from small-scale evaluations can be used to provide a case for roll-out and inform the design of future large-scale evaluations.

Governments around the world are striving to improve educational outcomes. It is important to ensure an impact evaluation is in place early in the intervention to capture the effect of the innovative approach. Creating a culture of evaluation will ensure that all stakeholders understand that the evaluation results will be used to demonstrate impact. Impact evaluations can facilitate evidence-based best practice sharing.

References

Adato, M. 2011. *Combining Quantitative and Qualitative Methods for Program Monitoring and Evaluation: Why Are Mixed-Method Designs Best?* Washington, DC: World Bank.

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

Barrera-Osorio, F. 2007. *The Impact of Private Provision of Public Education: Empirical Evidence from Bogota's Concession Schools*. Policy Research Working Paper No. 4121. Washington, DC: World Bank.

Baker, J. L. 2000. *Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners*. Directions in Development. Washington, DC: World Bank.

Bamberger, M., J. Rugh and L. Mabry. 2012. *Real World Evaluation: Working Under Budgets, Time, Data and Political Constraints*. Thousand Oaks, CA: Sage.

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

Berk, R. A. 1983. An Introduction to Sample Selection Bias in Sociological Data. *American Sociological Review* 48(3): 386-398.

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

Bloom, H., J. Bos and S. Lee. 1999. Using Cluster Random Assignment to Measure Program Impacts. *Evaluation Review* 23(4): 445-469.

Borman, G., G. Hewes, L. Overman and S. Brown. 2003. Comprehensive School Reform and Achievement: A Meta-analysis. *Review of Educational Research* 73(2): 125-230.

Bruns, B., D. Filmer and H. A. Patrinos. 2011. *Making Schools Work: New Evidence on Accountability Reforms.* Washington, DC: World Bank.

Bruns, B. and J. Luque. 2014. *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean.* Washington, DC: World Bank.

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

Diamond, A. and J. Hainmueller. 2007. *The Encouragement Design for Program Evaluation*. Washington, DC: International Finance Corporation (IFC) and Cambridge, MA: Harvard University.

Gray, F. D. 1998. *The Tuskegee Syphilis Study: The Real Story and Beyond.* Montgomery, AL: NewSouth Books, Inc.

Gertler, P., H. A. Patrinos and M. Rubio-Codina. 2007. *Impact Evaluation for School-Based Management Reform*. Washington, DC: World Bank.

Gertler, P., H. A. Patrinos and M. Rubio-Codina. 2012. Empowering Parents to Improve Education: Evidence from Rural Mexico. *Journal of Development Economics* 99(1): 68-79.

Gertler, P. J., S. Martinez, P. Premand, L. B. Rawlings and C. M. J. Vermeersch. 2011. *Impact Evaluation in Practice*. Washington, DC: World Bank.

Grindle, M. 2004. *Despite the Odds: The Contentious Politics of Educational Reform.* Princeton, NJ: Princeton University Press.

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

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

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

Hinton, R. 2015. *Assessing the Strength of Evidence in the Education Sector*. London, England: United Kingdom Department for International Development.

Jimenez, E. and Y. Sawada. 1999. Do Community-managed Schools Work? An Evaluation of El Salvador's EDUCO Program. *World Bank Economic Review* 13(3): 415-441.

Keefer, P. and S. Khemani. 2005. Democracy, Public Expenditures, and the Poor: Understanding Political Incentives for Providing Public Services. *World Bank Research Observer* 20(1): 1-28.

Khandker, S., G. Koolwal and H. Samad. 2010. *Handbook on Impact Evaluation: Quantitative Methods and Practices.* Washington, DC: World Bank.

Khemani, S. 2007. Can Information Campaigns Overcome Political Obstacles to Serving the Poor? in S. Devarajan and I. Widlund, eds., *The Politics of Service Delivery in Democracies.*

Better Access for the Poor. Stockholm, Sweden: Expert Group on Development Issues Secretariat, Ministry for Foreign Affairs.

King, E. M. and J. R. Behrman. 2009. Timing and Duration of Exposure in Evaluations of Social Programs. *World Bank Research Observer* 24(1): 55-82.

Krueger, A. B. and Zhu, P. 2004. Another Look at the New York City School Voucher Experiment. *American Behavioral Scientist* 47(5) 658-698.

Lewis, L. and H. Patrinos. 2012. Impact Evaluation of Private Sector Participation in Education. Reading, Berkshire, England: CfBT and Washington, DC: World Bank, 2012.

Loeb, S., Valant, J. and Kasman, M. 2011. Increasing Choice in the Market for Schools: Recent Reforms and Their Effects on Student Achievement. *National Tax Journal* 64(1): 141-164.

Machin, S. and Vernoit, J. 2011. *Changing School Autonomy: Academy Schools and Their Introduction to England's Education*. Centre for the Economics of Education Discussion Paper 123. London, England: London School of Economics.

Majumdar, S., A. Mani and S. Mukand. 2004. Politics, Information and the Urban Bias. *Journal of Development Economics* 75(1): 137-165.

Marcus, A. (ed.) and D. Berman. 2013. *Cambodia: Challenges in Scaling Up Preschools.* Washington, DC: World Bank Group.

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

Montenegro, Claudio E. and H. A. Patrinos. 2014. *Comparable Estimates of Returns to Schooling Around the World*. Policy Research Working Paper 7020. Washington, DC: World Bank.

Newman, J., M. Pradhan, L. B. Rawlings, G. Ridder, R. Coa and J. L. Evia. 2002. An Impact Evaluation of Education, Health, and Water Supply Investments by the Bolivian Social Investment Fund. *World Bank Economic Review* 16 (2): 241–74

Raudenbush, S. W., J. Spybrook, X. Liu and R. Congdon. 2004. *Optimal Design for Longitudinal and Multilevel Research: Documentation for the Optimal Design Software, Version 1.76.* Ann Arbor, MI: University of Michigan.

Shadish, W., T. D. Cook and D. T. Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston, MA, and New York, NY: Houghton Mifflin Company.

United Nations Educational, Scientific and Cultural Organization (UNESCO) Institute for Statistics and Education for All Global Monitoring Report. 2014. *Progress in Getting All Children to School Stalls but Some Countries Show the Way Forward.* Policy Paper 14/Fact Sheet 28. Montreal, Quebec, Canada: UNESCO Institute for Statistics.

World Bank. 2012. *World Development Report 2012: Gender Equality and Development.* Washington, DC: World Bank.

BUILDING EVIDENCE IN EDUCATION