Improving Education Through Assessment, Innovation, and Evaluation

Henry Braun, Anil Kanjee, Eric Bettinger, and Michael Kremer

© 2006 by the American Academy of Arts and Sciences
All rights reserved.

ISBN: 0-87724-058-2

The views expressed in this volume are those held by each contributor and are not necessarily those of the Officers and Fellows of the American Academy of Arts and Sciences or its Project on Universal Basic and Secondary Education.

Please direct inquiries to:
American Academy of Arts and Sciences
136 Irving Street
Cambridge, MA 02138-1996
Telephone: (617) 576-5000
Fax: (617) 576-5050
Email: aaas@amacad.org
Visit our website at www.amacad.org

Contents

Preface

How is progress toward the goal of universal basic and secondary education measured? If measurable progress is being made, to what can it be attributed? How can we identify effective strategies for increasing access to schooling or for improving the quality of what is taught? As part of the American Academy's project on Universal Basic and Secondary Education, we asked these questions of Henry Braun, Anil Kanjee, Eric Bettinger, and Michael Kremer.

Although assessment is often seen as a tool to measure the progress of individual students, it also allows individuals, communities, and countries to track the quality of schools and educational systems. In theory, if policymakers have access to reliable information on educational quality in specific schools and make this information available to the aware public, then students and parents may be better able to choose among educational options and demand education of higher quality.

Educational assessment must overcome a central dilemma, as Braun and Kanjee observe. If there are no consequences attached to a test, then it will do little to motivate healthy change within the educational system; however, if the result of an assessment is highly consequential, then it may engender unproductive or undesirable outcomes such as narrowing the curriculum, "teaching to test," and weakening the role of teachers. When assessments are tied to funding decisions, those responsible for the quality of education—teachers, administrators, and state officials—may oppose the release or even the creation of such data. Braun and Kanjee describe the factors preventing better assessment and review promising national, regional, and international initiatives for improving current practices and resolving this dilemma.

One recommendation they offer is to encourage developing countries to participate in international assessments as "associates," without requiring that the results be released internationally. This interim arrangement, they argue, would promote the generation of much-needed data, give developing countries access to expertise, and build local capacity to develop, administer, and analyze tests, while avoiding the political consequences to participating countries of possible poor performance.

Many aspects of traditional educational practice have not been evaluated rigorously. Would students learn arithmetic or history less effectively if they were not required to be in their seats by the time the school bell rang? Does a student who learns touch-typing from a computer learn any better, or at a significantly lower cost, than a student who learns from a traditional teacher or by self-instruction from a printed book? Few innovations in education

have been rigorously compared with traditional practices to measure quantitatively what they contribute to educational outcomes. Although traditional assessments are limited in the types of data they can provide to evaluate educational systems and practices, there are other means to ensure that educational practices achieve the desired ends.

As Bettinger and Kremer each discuss, a reliable means of getting answers to questions like these—namely, randomized controlled experimentation, the gold standard for evaluating treatments in medicine—is now finding use in education. Such experiments make possible valid comparisons among pedagogical techniques and systems of management because randomization establishes equivalent participant and non-participant groups for comparison. Randomized controlled experiments can, therefore, produce the most credible evaluation of programs, including their cost-effectiveness. With more reliable information from such experiments, education reformers can focus efforts and resources on the programs that have been found to be most effective.

Kremer's paper examines low-cost means of increasing enrollment. He reviews the findings from randomized evaluations of a number of education initiatives, including school-based health programs. Kremer reports on a program that provided de-worming medication and iron and vitamin A supplements to preschool children in Delhi (at a cost of $1.70 per student per year). The treatments were phased in at random to 200 schools over a two-year period, enabling a comparison of treatment and non-treatment groups. Researchers found that the treatment had the effect of reducing absenteeism by 20 percent, making it an extremely low-cost means of increasing the number of days students are in school. Similar results were found in a randomized, controlled, school-based de-worming program in Kenya.

In his overview of what has been learned through randomized evaluations in education, Bettinger explains why these experiments, though they provide highly credible results, remain underutilized guides for policy. Randomized experiments can be expensive and time-consuming. They require technical sophistication to plan, implement, and analyze properly. He notes, however, that certain types of experiments are no more expensive or time-consuming than other rigorous data collection activities. A more formidable problem is the political justification of delivering a program to only a small set of students or schools while withholding it from a comparison group of students or schools. However, when budgetary constraints make it difficult or impossible to reach all members of a population in a given year, randomly selecting which groups receive the program in each of subsequent years may be the fairest way to implement the program and simultaneously permit measurements of the program's impact.

Versions of the three chapters that follow were discussed at workshops held at the American Academy in Cambridge, Massachusetts. A workshop on "Educational Assessment," was held on October 28–29, 2002, and was attended by Albert Beaton (Boston College), David E. Bloom (Harvard University), Henry Braun (Educational Testing Service), Joel E. Cohen (Rockefeller and

Columbia Universities), Juan Enrique Froemel (UNESCO), Rangachar Govinda (National Institute of Educational Planning and Administration), Stephen Heyneman (Vanderbilt University), Anil Kanjee (Human Sciences Research Council), Denise Lievesley (UNESCO), Marlaine Lockheed (World Bank), George Madaus (Boston College), Martin Malin (American Academy), and Laura Salganik (American Institutes for Research). We thank the other participants in this workshop for their extremely valuable comments. Braun and Kanjee also thank Pai Obanya, Rangachar Govinda, Sidney Irvine, and Christopher Modu for their generous comments, Juan Guzman for assisting with the country profiles, and South Africa workshop participants for their comments on the framework presented in their paper.

A workshop on "Means and Technology" was held on February 1–2, 2004. Participants included: Farida Allaghi (Agfund), Leslie Berlowitz (American Academy), Eric Bettinger (Case Western Reserve University), David E. Bloom (Harvard University), Chip Bury (International Christian Supportfund), Joel E. Cohen (Rockefeller and Columbia Universities), James DiFrancesca (American Academy), Kira Gnesdiloff (Monterrey Tech), Donald Green (Yale University), Margaret Honey (Center for Children and Technology), Michael Kremer (Harvard University), Stanley Litow (IBM International Foundation), Colin Maclay (Harvard University), Martin Malin (American Academy), Lynn Murphy (William and Flora Hewlett Foundation), Laura Ruiz Perez (Monterrey Tech), Ryan Phillips (New York), Robert Spielvogel (Center for Children and Technology), Daniel Wagner (University of Pennsylvania), Robin Willner (IBM), and Cream Wright (UNICEF). We thank the other participants for their contributions. Bettinger also thanks Richard Murnane for his helpful comments and Erin Riley for her research assistance. Kremer extends his thanks to Heidi Williams for her excellent research assistance.

Each paper was read by two or more anonymous reviewers. We join the authors in thanking their respective reviewers for their written comments. A special thanks is due to Helen Curry at the American Academy, whose intellectual contributions, project coordination, and copy-editing have been indispensable. Leslie Berlowitz's vision and leadership as chief executive officer of the American Academy made this project possible.

The UBASE project focuses on the rationale, the means, and the consequences of providing the equivalent of a primary and secondary education of quality to all the world's children. This monograph is one in a series of the UBASE project published by the American Academy. Other papers examine related topics, including:

- basic facts about education, and the nature and quality of the data that underpin these facts;
- the history of efforts to achieve universal education, and political obstacles that these efforts have encountered;
- the goals of primary and secondary education in different settings;

- the costs of achieving universal education at the primary and secondary levels;
- health and education; and
- the economic and social consequences of global educational expansion.

The complexity of achieving universal basic and secondary education extends beyond the bounds of any single discipline and necessitates disciplinary rigor as well as interdisciplinary, international, and cross-professional collaboration. By focusing on both primary and secondary education, paying attention to access, quality, and cultural diversity, and encouraging fresh perspectives, we hope that the UBASE project will accelerate and enrich educational development.

This project is supported by major funding from the William and Flora Hewlett Foundation, and by generous grants from John Reed, the Golden Family Foundation, Paul Zuckerman, an anonymous donor, and the American Academy of Arts and Sciences. The project also benefits from the advice of a distinguished advisory committee, whose names are listed on the inside front cover.

As with all Occasional Papers of the American Academy, responsibility for the views presented here rests with the authors.

Joel E. Cohen
Rockefeller and Columbia Universities

David E. Bloom
Harvard University

Martin Malin
American Academy of Arts and Sciences

CHAPTER 1

Using Assessment to Improve Education in Developing Nations

HENRY BRAUN AND ANIL KANJEE

The Universal Basic and Secondary Education (UBASE) initiative has examined from a number of perspectives the challenge of providing every child in the world with a good basic and secondary education (Bloom and Cohen, 2005), exploring the premise that the rapid expansion of high quality education is essential to the economic, social, and political well being of developing nations. The goals of the UBASE initiative complement those of the Education For All (EFA) initiative, as introduced in the World Declaration on Education for All in Jomtein, Thailand (UNDP/UNESCO/UNICEF/World Bank, 1990) and later revised and reaffirmed in the Dakar Framework for Action (UNESCO, 2000a).

This paper provides a framework for conceptualizing the various roles assessment plays in education, as well as an overview of educational assessment in the developing world. It undertakes an analysis of some assessment-related issues that arise when planning to expand dramatically educational access and quality. In particular, it suggests how assessment practices and systems can generate relevant and timely information for the improvement of education systems, presents case studies of a number of nations, describes some international efforts, and proposes next steps.

The issues raised in this paper are especially relevant to the EFA initiatives; in particular, Goal 6 of the Dakar Framework (UNESCO, 2000a: 17) calls for "improving all aspects of the quality of education, and ensuring their excellence so that recognized and measurable learning outcomes are achieved by all, especially in literacy, numeracy and essential life skills." The Dakar Framework also suggests various approaches countries may adopt to attain the goals outlined and proposes that countries "systematically monitor progress towards EFA goals and strategies at the national, regional and international levels" (UNESCO, 2000a: 21).

Our intention in this paper is not simply to describe how assessment-related initiatives can be extended to the secondary-education sector, but to offer a comprehensive analysis of the roles assessment can—and could—play in educational improvement. This effort is undertaken with humility: most of the sensible things that can be said have been said; nevertheless, many of the sensible things that can be done, have not been done—at least not on a large scale.

The education landscape in many, if not most, developing countries is characterized by a number of patterns:

- There exist substantial disparities in the distribution of opportunity to learn and in achievement. These disparities are associated with factors such as geographic location, race/ethnicity, language, social class, and gender, among others.
- In a particular region (e.g., Latin America or Sub-Saharan Africa), disparities within a country are usually much greater than average differences among countries.[1]
- In general, achievement levels are low, both with respect to a country's own standards and in comparison to the norms established by developed nations.
- There are many impediments to progress, including limited facilities and resources, insufficient capacity, inefficient allocation of available resources, and wastage due to high rates of grade repetition and attrition.

The solutions to these problems are varied and extremely complex, and certainly cannot be addressed only, or even chiefly, through assessment. However, assessment policy and practices are critical to any successful educational improvement strategy; assessment data are essential to teaching and learning and are needed to monitor, evaluate, and improve the education system. Although some assessments serve learners, teachers, parents, and policymakers by providing them with useful information, others focus educational efforts by virtue of the consequences that are attached to learner performance. This dual role leads to the paradox of "high-stakes" assessment as an instrument of change. In the absence of serious consequences, it is difficult for assessment to exert much influence on an education system; however, if performance on an assessment entails serious consequences, it can lead to activities that are educationally unproductive and may actually undermine the integrity of the system.

This paradox is at the heart of the controversy over assessment in educational circles. To some, assessment is a fair and objective way to set and maintain standards, to spearhead reform at the levels of both policy and practice, and to establish a basis for meaningful accountability. To others, it is an instrument for maintaining the status quo, grossly unfair and educationally unproductive. There are of course more balanced positions, such as those of Little (1990) and Noah and Eckstein (1992b). Whatever their position, most observers would agree that assessment is neither an end in itself nor a panacea for the ills of education. They would likely also accept the proposition that major improvements in assessment systems must be part of a broader educational reform agenda that will be driven by—and constrained by—political, economic, and social considerations (Govinda, 1998; Little, 1992).

The importance of assessment for policy stems, in part, from the widespread recognition that educational indicator systems must include not only

1. Similar patterns of disparity are also common among developed countries, especially as they relate to language and race.

inputs but also outputs. A "results agenda" has become increasingly prominent in the planning of international donor agencies (Lockheed, private communication, May 21, 2004). This is complicated, however, by the fact that the assessment data used for policy may be incomplete, of poor quality, or even unreliable or invalid. Furthermore, lack of appropriate sensitivity to contextual issues can make data interpretation and subsequent action problematic. For example, in some regions of a country it may be common for large proportions of learners sitting for an examination to participate in "out-of-school tuition" (N. Postlethwaite, private communication, 15 October 2002) or shadow education (Baker et al., 2002). These practices may raise test scores, but they also distort inferences about the comparative effectiveness of different schools or different regions.

The increased salience of assessment to policy naturally leads to demands that it meet higher standards of quality and validity; this places still greater strains on the assessment capacity of many nations, especially developing countries. Indeed, these will likely have to look beyond available test instruments and consider anew the entire design and development process, in which local and national values and goals play a critical, if often not well articulated, role. At the same time, from a global perspective, there is a wealth of assessment related materials and expertise that developing nations should be able to tap into and adapt to their own needs.

The expectation that assessments aligned to national goals ought to be central to education and thus exert a beneficial influence on the economic and social conditions of the people is not a new one. In 1955, Godfrey Thomson said of the test movement in India (Bhatia, 1955: Foreword):

> It is of the greatest importance to India, and to the world, that her rising generation should be well educated, each in the way best fitted to his or her talents, and that her manpower, in adulthood, should be helped into those occupations most needed by the nation, most likely to profit by the individual's special abilities, and most likely therefore to make him happy and self-respecting. The object of the test movement…is exactly to forward such aims, not by dictatorial direction but by careful assessment of abilities, general and special, and helpful recommendations based on such assessment.

This paper focuses on assessment as a tool to improve learning, to monitor and credential students, and to evaluate some aspects of the education system itself. Certainly, assessment data, when appropriately aggregated, can be an important component of a broader educational indicator system. This paper, however, does not treat the use of assessment for such administrative purposes as the evaluation of teachers, principals, or schools. Lievesley (2001) presents a brief but insightful account of the potential and the pitfalls associated with development indicators.

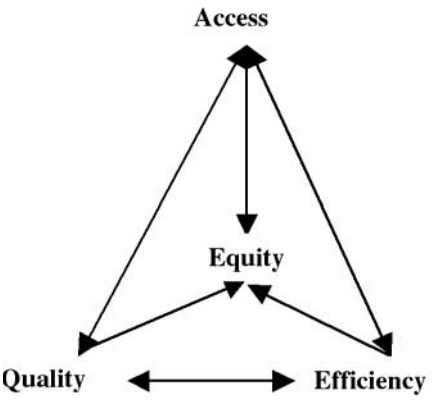
In this paper, we first propose a framework through which we conceptualize the role of assessment for improving access, quality, efficiency, and equity within the education system. Next we define assessment, outline the different

types of assessment and indicate the various uses to which assessment can be put. We then focus on the considerations particular to the secondary education sector in developing nations and highlight the various factors affecting assessment practices. This is followed by a discussion of ongoing concerns and constraints on improving assessment systems and practices. In addition, we address the role of technology in enhancing the practice of assessment, as well as improving the nature and quality of data collected. Case studies from several developing countries are presented to highlight current assessment systems and practices while the roles of current and recent regional/international assessment initiatives in developing countries are also noted. The paper concludes with a summary and a presentation of several strategies for moving forward.

CONCEPTUAL FRAMEWORK

In establishing a framework for discussing the role of assessment, we have identified four essential attributes of an education system: Access, Quality, Efficiency, and Equity, which we will refer to by the acronym AQEE (pronounced "a key").[2] Figure 1 illustrates the interdependence among these various attributes. While recognizing that these attributes are intimately linked, we provide a separate working definition for each. It is important to note that many different meanings and interpretations of the AQEE concepts have been proposed (Ndoye, 2002; Obanya, 2002; UNICEF, 2003). The intent of this paper is not to provide universally acceptable definitions. Instead, we offer these attributes as a starting point for systematically examining the uses of assessment in an education system.

Figure 1: Interdependence of AQEE concepts



Access

The concept of access generally refers to entry into the formal school system and comprises three aspects:

- Getting to school – how learners travel to school, how far they need to travel, and how long it takes
- Getting into school – obstacles to attending schools (e.g., disability, child labor, safety) and admissions policies (e.g., age/grade limits, fees, restriction to specific catchment areas, admissions tests, and availability of places)

2. In this framework, the concept of effectiveness has been excluded as it refers more to micro-level factors within any education system.

- Getting through school – promotion policies and practices, both influenced by the quality of education provided

Quality

The concept of "education quality" has as many different meanings as it has writers, and generally includes the following:

- What learners should know – the goals of the education system as reflected in missions/value statements and elaborated in the curriculum and performance standards
- Where learning occurs – the context in which learning occurs (e.g., class size, level of health and safety of the learning environment, availability of resources and facilities to support learning such as classrooms, books, learning materials, etc.)
- How learning takes place – the characteristics of learner-teacher interactions (e.g., the roles learners play in their learning, teacher and learner attitudes towards learning, other teacher practices, etc.)
- What is actually learned – the outcomes of education (e.g., the knowledge, skills, competencies, attitudes, and values that learners acquire)

Efficiency

Efficiency refers to the optimal use of educational resources and facilities to improve access to schooling and the quality of education provided. Efficiency generally comprises the following:

- The functioning of the current structures and systems at different levels (e.g., provinces, regions, districts, and schools) – how these are staffed and managed (e.g., district managers, school governing bodies) regarding the formulation, implementation, and monitoring of policy and practice within the system
- The availability, allocation, and use of human and financial resources – how available resources within a system are managed and employed at different levels within the system
- Throughput and repetition rates – the number of learners that enter and leave a system as well as the number of learners that repeat any grades

Equity

The concept of equity is based on the principle that essentially all children can learn and should be provided with an equal opportunity to do so, irrespective of their background. Equity within any education system is generally based on the following principles:

- Inclusivity – the capacity of the education system to address the specific needs of all children irrespective of their language, gender, religion, sexual orientation, (dis)ability, etc.
- Absence of unfair discrimination – the capacity of the education system to actively address unfair discriminatory practices or situations and their consequences for a specific subgroup. (In our view, the use of practices target-

ed at specific groups to address inequity within the system is both acceptable and necessary; for example, the introduction of additional math and sciences programs specifically for female learners.)

Evidently, there exists a complex interdependence among these attributes. For example, lack of efficiency in the context of limited resources will typically adversely affect access, quality and equity. Similarly, lack of quality, real or perceived, may well reduce access and equity as those families with fewest resources find the returns inadequate to justify the investments in school-related expenses and the opportunity costs incurred.

Systemic Validity

In evaluating the contributions of measurement to an education system, the principle that seems most appropriate is that of "systemic validity" (Frederiksen and Collins, 1989: 28). A systemically valid test is "…one that induces in the education system curricular and instructional changes that foster the development of the cognitive skills that the test is designed to measure." Their notion initially stemmed from a concern that high stakes tests can, and do, cause learners and teachers to focus their efforts on maximizing test scores. Such an effort may not be accompanied by achievement of the intended learning goals if there is a disjuncture between immediate outputs (test scores) and desired outcomes (student learning).

We propose to extend Frederiksen and Collins' definition of systemic validity in the following way:

> Assessment practices and systems are systemically valid if they generate useful information that supports the (continuous) improvement in one or more aspects of AQEE within the education system, without causing undue deterioration in other aspects or at other levels.

We recognize that to make any evaluation of systemic validity requires a judgment about both the nature of any improvement with respect to AQEE and whether particular changes in assessment practices would result in such an improvement.

Our rationale for this revised definition is that, in many instances, assessments can be systemically valid according to Frederiksen and Collins and yet not support educational improvement more broadly. For example, the academic content tested in a school-leaving examination may be suitable for those learners intending to continue their schooling, but not entirely appropriate for those leaving school, who would benefit from a curriculum and a preparation that covered a wider range of skills. There is a conceptual and practical distinction between certifying that a learner has met the standards associated with a given stage of education and determining whether he or she merits advancement to the next stage, and there are few tests that can serve both functions well.

The basic notion of systemic validity is not a new one, even in international education. Heyneman (1987) suggests that national examinations

could be used to improve classroom pedagogy while Heyneman and Ransom (1990) suggest that well-designed national examinations could lead to improvements in educational quality. They argue that because these examinations play such an important role in the allocation of life's chances, they have powerful "backwash effects" which can be harnessed to positive ends.[3]

However, existing testing practices often exert deleterious effects on the education system. For example, in a study of public examinations in Sub-Saharan Africa, Kellaghan and Greaney (1992) point out, inter alia, that these practices often result in unwanted narrowing of the curriculum, an unproductive focus on esoteric material, as well as a warping of the teacher's role and, often enough, compromised test results. (It is somewhat ironic that these difficulties are almost identical to the concerns expressed by many observers in the United States, as American states increasingly adopt end-of-grade assessments as a critical element of their own reform efforts.) Likewise, Govinda (1998), in his overview of testing in India over the past half century, offers conclusions on the consequences of mandated testing policies that are consistent with the findings of Kellaghan and Greaney.

The obvious lesson is that assessment is a double-edged sword, with significant departures from systemic validity likely signaling substantial inefficiencies—inefficiencies that the economies of developing nations can ill afford. Unfortunately, achieving systemic validity is not easy. Alignment of the different components of a multi-level education system is a daunting goal. The U.S. assessment initiative alluded to above is intended to be one aspect of accomplishing such a task, but the problems that the United States has encountered and the apparent lack of significant progress are testimony to the difficulties of such an undertaking.

The challenge, then, is how to nurture and develop, even under the manifold constraints characteristic of developing nations, assessment practices (and systems) that are systemically valid. The constraints range from lack of political will, of human or financial capital and insufficient infrastructure capacity, to the inertia attached to current practice. While these problems may be similar across countries, the specific national (or sub-national) contexts are different enough that it is unlikely that one can formulate meaningful general policy recommendations that will be operationally useful in more than a handful of settings. Accordingly, we subscribe to the aphorism: "Common challenges, local solutions."

Challenges in Implementing Systemically Valid Assessment Practices

The assessment system within the education sector comprises all policies and practices related to conducting assessments and evaluations, as well as the structures and organizations established to ensure effective implementation. Assessment and examination policies, examination structures and practices, national assessments, national standards, classroom assessments, certification

3. The "backwash effect" refers to the impact of assessment, particularly the uses of assessment results, on learning, teaching, curriculum, learning materials, and education programs.

bodies, and qualifications frameworks are all components of an assessment system. In practice, the assessment systems of countries vary significantly from each other, both in terms of policies, practices, and structures, as well as the capacity for effective implementation. Thus, it is possible for two seemingly identical assessment systems to have very different outcomes.

The effective functioning of an assessment system is determined not only by how this system (or subsystem) articulates with other facets of education, such as curriculum and instruction, but also by how well the various sectors (primary, secondary, higher) and structures within the education system articulate with one another. In an ideal context, all components of an assessment system would articulate perfectly and function effectively to produce the desired outcomes. However, this is difficult to attain in practice, and it is more likely that one of the three scenarios outlined below exist.

First, the assessment (sub)system, or components thereof, does not function effectively. For example, the national examination results are not regarded as reliable, or the information generated is not particularly relevant due either to the poor quality of the test instruments or the limited dissemination of the results, or both.

Second, the education system does not function effectively. In this instance, any assessment system will have little, if any, impact. For example, information from assessments conducted at the end of primary/secondary schooling will have little impact on issues of access if there are not enough places in the next level to accommodate all graduating and qualified learners.

Third, both the assessment and education systems function effectively. In this instance, assessment systems that seem to be functioning effectively can still result in unintended and educationally sub-optimal consequences. In India for example, an effort to implement minimum levels of learning was appropriately accompanied by large-scale teacher training programs. However, within a few years, researchers found that teachers were teaching to the test (Govinda, 1998). Govinda (1998) notes that there were additional negative consequences since the net effect of the program reinforced rote learning and "transmissionist" teaching methods, and it helped generate a major after-school test preparation industry that served to increase the bias against learners from poorer backgrounds.

Clearly, the roles and impact of an assessment system are substantially determined by the availability, and appropriate allocation, of both human and financial resources. However, decisions pertaining to the allocation of resources must account for the following: 1) the stage of development of the education system; 2) the form and function of the different assessments, which change from feedback to monitoring and evaluation as one moves up from the classroom to the school, district, and beyond; and 3) the frequency of assessments, which typically tends to decrease as one moves to higher levels of the education system. In general, one can argue that for those education systems that are at an early developmental stage, less frequent assessments, following a baseline assessment, should be sufficient because many of the issues that need to be addressed are known and a number of years are

required for substantial improvement. In this case, scarce resources are better devoted to assessments directed at improving learning and teaching, where the returns on investments are likely to be higher (Black and Wiliam, 1998).

In developing nations, the allocation of resources for assessment systems should also account for the specific needs of "vulnerable populations," e.g., the girl child, the out-of-school youth, and the illiterate adult learner. The assessment system should facilitate the collection of data from beyond schools, e.g., household surveys, and allow for the recognition of relevant experiences and skills that many adult learners acquire out of school, especially those who have little or no formal schooling. In this respect, UNICEF has focused on improving the education opportunities of the "girl child" in rural and poor communities as a means of improving the lives of both girls and their communities (UNICEF, 2002). In Brazil, the government has embarked on a national campaign to improve attendance of poor children at school by using financial incentives for poor families as a means of combating the social conditions that force their children to work (Dugger, 2004). In an effort to improve the literacy and numeracy skills of adults in the United States, the National Institute for Literacy has established a set of comprehensive standards to address the question of "what adults need to know and be able to do in the 21st century" (Stein, 2000). These standards are noteworthy in that they cover a broad range of competencies that extend well beyond the usual compendium of academic skills to include the following categories: communication skills, decision-making skills, interpersonal skills, and lifelong learning skills.

WHAT IS ASSESSMENT?

We begin by distinguishing among four related terms (Keeves, 1997; UNESCO, 2000b): measurement, testing, evaluation, and assessment. *Measurement* refers to the process by which a value, usually numerical, is assigned to the attributes or dimensions of some concept or physical object. For example, a thermometer is used to measure temperature while a test is used to measure ability or aptitude. *Testing* refers to the process of administering a test to measure one or more concepts, usually under standardized conditions. For example, tests are used to measure how much a student has learned in a course of mathematics. *Evaluation* refers to the process of arriving at judgments about abstract entities such as programs, curricula, organizations, and institutions. For example, systemic evaluations (e.g., national assessments) are conducted to ascertain how well an education system is functioning. In most education contexts, assessments are a vital component of any evaluation. *Assessment* is defined as "the process of obtaining information that is used to make educational decisions about students, to give feedback to the student about his or her progress, strengths and weaknesses, to judge instructional effectiveness and curricular adequacy and to inform policy" (AFT, NCME, NEA, 1990: 1). This process usually involves a range of different qualitative and quantitative techniques. For example, the language ability

of learners can be assessed using standardized tests, oral exams, portfolios, and practical exercises.

Assessment plays many roles in education and a single assessment can serve multiple, but quite distinct, roles. For example, results from a selection test can sometimes be used to guide instruction, while a portfolio of learner work culled from assessments conducted during a course of study can inform a decision about whether the learner should obtain a certificate of completion or a degree.[4] Simplifying somewhat, we can posit that from a learner's perspective, there are three main roles for assessments: Choose, Learn, and Qualify. The data from an assessment can be used to choose a program of study or a particular course within a program. Other assessments provide information that can be used by the learner, teacher, or parents to track learner progress or diagnose strengths and weaknesses. Finally, assessments can determine whether learners obtain certificates or other qualifications that enable them to attain their goals. Assessment in the service of individual learning is sometimes referred to as "formative assessment," in contrast to "summative assessment," which is intended to guide decision-making (see Black and Wiliam, 1998).

From the perspective of the authorities, the three critical functions of assessment are: Select, Monitor, and Hold Accountable. One of the most important functions is to determine which learners are allowed to proceed to the next level of schooling. Assessment results, along with other measurement data (such as those obtained through periodic surveys), are also used to track the functioning of different components of the system (generally referred to as national assessments), and sometimes are used to hold accountable the individuals responsible for those components.

Types of Assessments

To complement our categorization of the different roles of assessment, we present a brief overview of the different types of assessments that are typically employed by most nations. These are described more extensively in a recent report issued by UNESCO (2000).

The most common type of assessment is school-based. These assessments are usually devised and administered by class teachers, although some are the work of the school principal or other instructional staff. Typically, they are aligned with the delivered curriculum and may employ a broader array of media (e.g., oral presentations) and address a greater range of topics than is the case with centralized standardized assessments. They have a decided advantage over centralized assessments in that the results are immediately available to the teacher (and, presumably, the learners) and can influence the course of instruction. While these assessments can play an important role in

4. There are hundreds of books and articles on educational assessment (or measurement) with varying levels of comprehensiveness and degrees of association with the different schools of thought. A particularly readable and non-technical introduction to the subject can be found in the UNESCO publication "Status and Trends 2000: Assessing Learning Achievement" and in Black (1998). For a more technical treatment refer to Cronbach (1990).

promotion to the next grade, they are rarely used for high-stakes decisions such as admission to the next level of the education system (e.g., university). Black and Wiliam (1998) make a strong case for the potential of school-based assessment to accelerate learning for all students. The key to effective assessment at this level is to devise questions or probes that can elicit learner responses relevant to the learning goals, while ensuring that teachers are capable of interpreting the results in ways that are pedagogically useful and have sufficient resources to guide learners appropriately. They distinguish between perfunctory assessments, the results of which are simply entered into a grade book, and truly formative assessments, meant to guide instruction and focus learner effort.[5]

The second type of assessment, public examinations, can fulfill one or more of the following roles: selecting learners for admission to secondary or tertiary education, credentialing learners for the world of work, and/or providing data for holding school staff accountable for their performance. While such examinations are an important component of every nation's education system, they are particularly critical in developing countries, where the number of candidates for advancement is usually many times greater than the number of places available. In many countries, these are standardized multiple choice examinations, while in others they comprise various forms of performance assessment (sometimes in conjunction with multiple choice components). Typically, they are designed, developed, and administered centrally with an almost exclusive focus on academic subjects. There is meager feedback to the school except the scores and/or pass rate, and, as a result, they offer little utility for school improvement programs beyond an exhortation to do better next time. Moreover, as we have already noted, public examination systems often have negative consequences for the general quality of education.

National assessments are studies focused on generating specific information that policymakers need to evaluate various aspects of the educational system. The results can be used for accountability purposes, to make resource allocation decisions, and even to heighten public awareness of education issues. These assessments may be administered to an entire cohort (census testing) or to a statistically chosen group (sample testing) and may also include background questionnaires for different participants (learners, teachers, administrators) to provide a meaningful context for interpreting test results. The utility of the data generated depends on the quality and relevance of the assessment, the thoroughness of the associated fieldwork, as well as the expertise of those charged with the analysis, interpretation, reporting, and dissemination of results.

International assessments assess learners in multiple countries, with the principal aim of providing cross-national comparisons that can illuminate a variety of educational policy issues. As with national assessments, they may also include background questions for different participants (learners, teachers, administrators) to provide a meaningful context for interpreting test

5. For a specific example of the effective use of formative assessment in a secondary school setting, see Wiliam, Lee, Harrison, and Black (2004).

results. Such studies are planned and implemented by various organizations, including the IEA (International Association for the Evaluation of Educational Achievement) that conducts TIMSS (Trends in International Mathematics and Sciences Study) and PIRLS (Progress in International Reading Literacy Study), the OECD (Organization for Economic Co-operation and Development) that is responsible for PISA (Program for International Student Achievement) studies, UNESCO/UNICEF that conducts the MLA (Monitoring Learning Achievement) studies and coordinates regional groupings such as the Latin American Laboratory for Assessment of Quality in Education (Laboratorio),[6] the Southern African Consortium for the Monitoring of Education Quality (SACMEQ), and Program for the Analysis of Educational Systems of the CONFEMEN (Franco-phone Africa) countries (PASEC).[7]

Studies such as TIMSS, PIRLS, and PISA are characterized by high quality assessment instruments, rigorous fieldwork, and sophisticated analyses of results. At their best, they can also provide useful information to those who seek to improve classroom practice. For example, TIMSS included comprehensive surveys of educators and compiled an extensive video library of classes in participating countries. Both have proven to be rich sources of research on cross-national classroom practices and of putative explanations of the differences in results.

As the preceding exposition should make clear, assessment has the potential to contribute to one or more aspects of AQEE, depending on the type of assessment and the context in which it is employed. School-based assessments can enhance efficiency by helping to target the efforts of both learners and teachers. To the extent that they are able to use the information appropriately, the quality of the learning is improved.

Obviously, public examinations for selection or credentialing directly affect access. In principle, they should also enhance equity by providing a "level playing field" for all candidates. In reality, however, differences in "opportunity to learn" mean that not all learners are equally prepared, and this inequality is usually reflected in the outcomes. If these differential rates of success are associated with geographic and/or demographic groupings, there can be political consequences. Despite these failings, public examinations may be the best alternative at a particular time and efforts should be directed at improving education quality for more disadvantaged learners. In the long term, public examinations intended for accountability can lead to improvements in quality and efficiency, provided the results are incorporated into an indicator system that influences policy and resource allocation. The same is true of national and international assessments—again, provided that the data are relevant and useful to countries and are used productively in planning system improvement.

6. Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE).

7. Programme d'Analyse des Systèmes Educatifs des Pays de la CONFEMEN.

Foundations of Test Design

The term "test," as we have defined it, denotes an instrument of assessment that is conducted under some set of formal conditions. Tests developed outside the school (often referred to as external examinations) are produced through a process of design and development that varies widely across time and space. In the past, professional test developers have tended to regard their work as an art, not susceptible to rigorous analysis. One consequence has been the implication that expertise can only be developed through many years of apprenticeship. While there is some truth to that notion, significant progress over the last decade in automated item generation (Irvine, 2002; Bejar, 2002), test design (Mislevy, 2003), automated test assembly (Swanson and Stocking, 1993; van der Linden, 1998), and even automated scoring of complex learner-constructed responses (Braun, Bejar, and Williamson, 2006), has undermined the "strong interpretation" of the test developers' position.

Test design is governed by three general factors, which we denote as "constructs," "communication," and "constraints"(Braun, 2000). Constructs refer to the ultimate targets of measurement. They are usually derived from theoretical considerations and are organized into a framework that is later translated into a set of test specifications. Communication refers to the kinds of information to be provided; that is, the claims and inferences to be made on the basis of the test, while constraints refer to the relatively unchanging features of the context(s) in which the test is to be designed, developed, and delivered. Together, these three factors delineate the set of feasible designs from which a particular design must emerge.

Unfortunately, test designers often fail to explicitly invoke these factors or to generate alternative designs that represent different tradeoffs among various goals and constraints. On the other hand, they are not insensitive to the constraints under which they must operate and the conditions under which the test must be administered. The design of a high-stakes selection test, for example, involves constraints (such as the need for security) that are not as salient for low-stakes national assessments. Similarly, time constraints will influence the number and type of items that can be incorporated in a particular test. Designers must also cope with the technical demands of reliability and validity (discussed below). Moreover, the high level of interest in large-scale assessments, both nationally and internationally, has resulted in the development of innovative and highly technical methods for the design of assessment instruments, the collection and analysis of data, as well as the reporting of results. Refer to the National Research Council (2002) publication for a comprehensive review of methodological advances.

An important contribution to the practice of assessment is the work of Mislevy and his colleagues (2003), who have developed a conceptual approach to, and methodology for, test design, termed "evidence-centered design" (ECD). The approach directly links test design to both evidentiary reasoning and general design science. The basic idea of ECD is that designers should "work backwards," by first determining the claims they would like

users to make about the assessment and the evidence needed to support those claims. They can then develop the exercises (items, probes, performance challenges, etc.) to elicit desired learner responses, the scoring rubrics used to transform those responses into relevant evidence, and the measurement models that cumulate or summarize that evidence. This effort typically involves a number of cycles of development, data collection and analysis, reflection, and revision. Finally, designers must determine the overall structure and format of the assessment, mode(s) of delivery and presentation, data management, and the myriad other details that constitute an operational assessment system. Although expert (and successful) designers have typically engaged in some or all of these activities, the formalization of the process, the creation of a software system to support implementation, and the development of cognitively-based psychometric models represent a major advance in assessment.

Standardized Tests

Standardization is a prerequisite for fairness when scores must be comparable. It demands, at a minimum, that the tests be administered under uniform conditions and graded according to a fixed set of rules or rubrics. The degree of standardization is particularly important for school-leaving and selection examinations—and varies considerably from country to country and among regions within a country. In the past, standardization has been best achieved for examinations set by testing agencies, such as the University of Cambridge Local Examination Syndicate that operate internationally. External examinations, set by regional or national authorities, are almost always standardized.

Standardized tests are ordinarily constructed according to one of two models: norm-referenced or criterion-referenced. For the former, a distribution of test scores is established for a reference population. New scores are typically presented in conjunction with the corresponding percentile with respect to that reference distribution. For the latter, two or more ordered categories are defined in terms of fixed thresholds on the score scale, and a new score is labeled in terms of the category into which it falls. The test design process differs according to which model is being used. When the principal interest is in ranking all learners, norm-referenced tests are preferred. When the issue is whether the learner has met a particular standard, criterion-referenced tests are more appropriate.

Typically countries use both norm- and criterion-referenced assessments. Most examinations, e.g., end of school exams, are norm-referenced, while most national assessments are criterion-referenced. For developing nations, criterion-referenced assessments are certainly more useful for obtaining information regarding learner performance against set standards and/or mastery of curriculum objectives. Some examinations are hybrids, with standards set in part by considering substantive criteria but also influenced by normative (often historical) data. One example is the battery of Advanced Placement assessments in the United States.

Test Format and Content

Test designers may use any combination of item types, including multiple choice items, learner constructed response items (solving problems, providing short answers, writing essays), and extended work samples or portfolios. The choice is based, in part, on the appropriateness of the format for the objective to be tested and, in part, on operational issues such as timing and cost. Multiple choice items can be scored much more cheaply (particularly if the scoring is done mechanically) than items that require some degree of human judgment. In addition, learners can usually respond to a larger number of multiple choice items in a given amount of time, so that tests incorporating more of these items tend to give more consistent results.

Test designers must also consider the kinds of cognitive demands elicited by different kinds of items. While it is easy to write items that require factual recall or rote application of procedures, it is more difficult to devise items that demand reasoning, argumentation, and integration. While multiple choice items can be used to test some "higher order skills," many other skills can only be probed by formats that require the learner to produce an uncued response.

Aspects of Technical Quality

The purpose of any assessment is to provide information, which is usually used to support a decision of some sort. By a currently accepted definition, "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989: 13).

This definition requires that both developers of assessment instruments and users of assessment results marshal theoretical and empirical arguments to justify a particular application. One does not validate an assessment per se, but, rather, the inferences and implications for action in a specific setting. Thus, the use of an assessment may be quite defensible in one situation but the use of the same assessment – even with the same population – may be problematic in another situation. As we indicated earlier, an assessment that is used to confer high school diplomas may not be suitable for deciding which learners should be admitted to a highly selective college.

From a theoretical perspective, the two main threats to test validity are construct underrepresentation and construct-irrelevant variance (Messick, 1989). As its name suggests, the former refers to a situation in which the assessment does not adequately or fully capture those measurable qualities generally associated by test users with the target construct. For example, if the target construct is general writing ability, then an assessment that requires the learner to write a single essay on a general topic and to complete a set of multiple choice items that focus on grammar would suffer from construct underrepresentation. Construct-irrelevant variance arises when one or more of the sources of systematic (i.e., non-random) individual differences in scores are not closely related to the target construct. To carry the example above a bit further, suppose the learners were required to type their essays

using a word processor. If some fraction of the learners had negligible experience with the technology while the rest were quite comfortable with it as a result of substantial use in their schools, the scores of the first group would likely be systematically (and inappropriately) lower than those of the second, even if their writing abilities were equal on average.

These threats to construct validity must be addressed primarily through the design process, although construct-irrelevant variance can also arise through poor control of testing conditions or scoring. Maintaining the integrity of the assessment process is critical. If, for example, some learners are coached by teachers with knowledge (exact or approximate) of the content of the test or, what is worse, the learners have obtained access to the test in advance of the administration, then the validity of the test scores is undermined. This is an extreme example of how fairness is an integral aspect of validity. On the other hand, if some learners have not been exposed to the content of the test, then this differential "opportunity to learn" is a failure of equity. Another, somewhat more technical problem arises when learners who have taken different forms of a test are judged by the same standard. If the test forms have not been carefully constructed to be psychometrically parallel, one group can be advantaged relative to another. In some cases, lack of comparability can be addressed through a process termed "test equating"[8] (Braun and Holland, 1982).

An important empirical characteristic of any assessment is its reliability. Reliability is usually defined as the correlation between sets of scores obtained from a sample of individuals on two occasions. High reliability implies that the ranking of learners would be very similar across different administrations. Reliability is influenced by many factors, including the homogeneity of the items comprising an assessment instrument, the length of the instrument, the nature of the candidate population, and the uniformity of the testing conditions. Reliability is treated in every introductory text on measurement (e.g., Cronbach, 1990). Various extensions, including generalizability theory, are presented in summary fashion in Feldt and Brennan (1989).

There are no absolute standards for acceptable levels of reliability. In high stakes settings, reliabilities above 0.85 are considered desirable. Reliabilities below 0.7 are usually unsatisfactory. Low reliability undermines validity because it implies that a large proportion of the observed variability among scores is due to random error and, consequently, inferences made and actions taken on the basis of those scores are likely to be problematic. On the other hand, for assessments that serve formative purposes, high reliability need not be a priority.

CONSIDERATIONS FOR SECONDARY EDUCATION

Secondary education is critical to improving the quality of life in developing nations. This education sector plays a pivotal role in promoting rapid eco-

8. Test equating refers to a statistical method for adjusting raw scores on different tests forms for slight differences in difficulty.

nomic growth by preparing learners to enter the world of work or to pursue further education and training (including teacher training), and by preparing young people and at-risk-youth to participate more fully in their own socio-development and the development of society (Bregman and Stallmeister, 2001; Bhuwanee, 2001). However, despite the key role of secondary education systems, minimal attention has been paid to this sector in the past few years; instead, greater emphases have been placed on the primary and higher education levels of the system (Lewin and Caillods, 2001).

By their very nature, secondary schools face greater challenges than primary schools, given the need for learners at the secondary level to move beyond standard academic content to the acquisition of relevant competencies and skills that would better prepare them to function in society. The real challenge is to incorporate relevant knowledge, skills, and experience into the learning and teaching process in a manner that will address the country's specific growth and development needs. This alone is a daunting task for any nation, one that many developed nations also struggle with. Fortunately, a great deal of thought and a fair amount of work has already been devoted to meeting this challenge.

For example, the OECD project on the Definition and Selection of Competencies (Rychen and Salganik, 2003), the Equipped for the Future content standards (Stein, 2000), and the publication on Linking School and Work (Jenkins, 1996; Resnick and Wirt, 1996) provide a number of possible frameworks and examples to address a range of competencies that can require high levels of skills and expertise. These efforts represent valuable sources of ideas, examples, and information that can, at the very least, serve as starting points for addressing the specific challenges facing many developing nations in providing learners with skills appropriate to their society's needs.

Another source of materials and expertise can be found in the national assessments or public examinations conducted by developed countries. For example, in the United States, the National Assessment of Educational Progress (NAEP) is administered to samples of learners in the fourth, eighth, and twelfth grades (Horkay, 1999). The U.S. Department of Education makes available subject frameworks, test items, scoring rubrics, and informational materials for teachers. Similar materials are available from the local examination syndicates in the United Kingdom and (presumably) the testing authorities in other nations as well. At the upper secondary level, the High Schools that Work initiative (Kaufman, Bradby, and Teitelbaum, 2000) focuses on improving the academic skills of U.S. students on the vocational track. Concerns about preparing learners for the world of work and the school-to-work transition are longstanding (Secretary's Commission on Achieving Necessary Skills, 1992; Resnick and Wirt, 1996). We have already mentioned the related activities subsumed under the Equipped for the Future effort (Stein, 2000). For a European perspective, see Jenkins (1996). The point is that, with some diligence, developing nations can harvest guides, frameworks, and materials that can help them to jumpstart their education reform planning and, especially, their assessment strategies. Of course, these

resources cannot be simply translated and put to use—a process of critical review, adaptation, and innovation is required. Such a process, especially if it involves diverse stakeholders, can be valuable in itself as an important element of a capacity-building program.

Success in expanding the primary education sector has led to a massive increase in the numbers of learners seeking enrollment at the next level. This section focuses on issues in the secondary education sector that can be addressed by assessment within the framework of AQEE. The information presented is based primarily on our review of developing African countries, although we cite examples from other developing nations. We also offer some suggestions for how assessment can play a stronger and more constructive role in achieving the goals of AQEE.

What is Being Assessed?

Across the secondary school systems of the developing nations that we surveyed, we found both differences and similarities in what was assessed and how the assessments were conducted. The configuration of assessment practices in different countries naturally depended on how the education systems were structured, as well as the nature and delivery of the curriculum. In the countries surveyed,[9] we found the following:

- Secondary education offers between five and six years of schooling, generally divided into lower secondary (grades seven to nine) and upper secondary (grades ten to twelve).
- In all countries, learners were offered the options of academic, technical, and/or vocational tracks.
- A core curriculum usually includes languages, mathematics, and science, and learners are generally allowed to select additional subjects.
- Some countries specify standards or levels (e.g., the minimum levels of learning in India), while others have no specifications regarding what learners should achieve.
- Criteria for entrance to secondary school vary substantially. In many countries (e.g., South Africa), assessment results from primary schools are used. In some (e.g., Senegal), the results of national examinations at the end of primary school are used, while in others (e.g., Columbia), secondary schools administer their own entrance exams.
- Exit exams are administered at the end of secondary school in all countries surveyed, generally leading to certification. These exams may be administered by the education ministry (e.g., Brazil, China, India, South Africa), by a regional examination board (e.g., the members of the West African Examinations Council—Ghana, Liberia, Nigeria, Sierra Leone, and The Gambia), or outsourced to an international examination board (e.g., Mauritius).

9. Countries surveyed are: Brazil, Chile, China, Columbia, Cuba, India, Indonesia, Jordan, Mauritius, Mexico, Morocco, Senegal, South Africa, Uganda, Vietnam. Refer to the section on country-level landscapes for more detail.

- In all countries, assessment is used primarily for selection of learners into the next grade, while in some countries (e.g., Cuba) assessment is also used for placement of learners into specific programs, such as academic or technical tracks.
- Entrance into higher education institutions also varies among countries. In some countries, school leaving certificates are used for entrance into higher education institutions (e.g., South Africa), while in other countries universities administer their own entrance examinations (e.g., Brazil) or use national entrance examinations (e.g., China). In many countries, additional requirements are often imposed by some universities, or even by faculties or departments within the universities.

Factors Influencing Assessment Practices at the Secondary Education Level

The secondary education sector of many developing nations can be characterized by inappropriate policies, an inexperienced teaching force, inadequate facilities and limited human and financial resources to effect change, relatively low enrollment rates, inappropriate and inadequate systems and structures to address current needs, and examination systems that have a significant impact on the career paths of learners (Bregman and Stallmeister, 2001; Bhuwanee, 2001; Holsinger and Cowell, 2000; Monyooe and Kanjee, 2001). We address each of these six factors below.

Inappropriate policies. In most developing countries, assessment policies (practices) focus primarily on examinations with little or no emphasis on classroom assessment or on monitoring and evaluation of the system (Kellaghan and Greaney, 2001). In instances where specific assessment policies do exist, inadequate attention has been accorded to the impact of assessment on the system. For example, in Chile, where the conduct of national assessments has been a consistent policy of the government for many decades, Schiefelbein (1993) notes that these assessments have not created any improvement in the education system. In South Africa, the implementation of outcomes-based education created greater obstacles for teachers, instead of improving the teaching and learning environment. Fortunately, however, this situation was rectified after the Ministry of Education enacted new policies based on the recommendation of a committee empowered to review the implementation of the new curriculum (DoE, 2000). As Obanya (personal communication, 7 May 2004) argues, we have to consider appropriate policy development and the quality of policy implementation in order to improve our education systems.

Inexperienced teaching force. The shortage of qualified and experienced teachers, as well as the low morale and motivation of the teaching force, has been cited as the key factor for the low performance of the education systems in many developing nations (Bregman and Stallmeister, 2001). The implementation of effective teacher development programs, regarded as vital for improvement in the provision of quality education, has been a characteristic of many systems in the last decade. For example, teacher development comprised a critical feature of the education reform initiatives enacted in 1996 in

Brazil (Guimaraes de Castro, 2001b). Similarly, in Indonesia, the government launched a national in-service training program for primary school teachers using the Open University (Moegiadi and Jiyono, 1994).

A key focus of these training programs should be the use of appropriate assessment practices in the classroom and for examination purposes, because most teachers are able neither to conduct adequate assessments in their daily interactions with learners nor to design tests for end-of-year examination or certification purposes. However, limited information is available regarding the content of many teacher development programs.

In South Africa, training programs on the use of Assessment Resource Banks (ARB) to improve teaching have yielded highly successful outcomes (Kanjee, 2003). The ARB comprised a series of assessment tasks, each of which included: 1) the relevant curriculum outcome and assessment standard, 2) assessment items to assess learner performance against specific standards, 3) scoring criteria along with information on interpretation of scores, and 4) a framework for recording scores. Teachers were trained to use the ARB to identify learner strengths and weaknesses, develop relevant intervention programs for specific learners, and record and monitor learner progress. An unintended result was that teachers also used the resource banks for developing lesson plans, writing their own items, and setting homework exercises, indications that the ARB can perhaps be used in teacher development programs.

The lack of proper appraisal systems also contributes to the poor state of teacher qualifications. Appraisal systems focus on evaluating the ability of the teacher to perform his/her job and should, in principle, include teacher competency tests. The application of appraisal systems is vital for recognizing the contributions of individual teachers, rewarding the better teachers while also identifying teachers in need of assistance. Appraisal systems, however, are a contentious issue and are extremely difficult to implement. If these systems are to work, there has to be a consensus on their use by all stakeholders. The system should be based on fair assessment principles and not used for punitive purposes. In Colombia, for example, strong opposition by teacher unions has stymied attempts to introduce such systems (Guzman, personal communication, September 2003).

Inadequate facilities, limited human and financial resources. The lack of adequate facilities and human resources in the education system has had deleterious consequences for many developing nations, and will continue to do so in the near future. For example, the need for more qualified teachers in a number of disciplines adds a burden to the secondary education sector beyond that found in the primary sector. In addition, the education systems of many developing nations are characterized by limited capacity to obtain relevant information for identifying areas in need of intervention, as well as limited financial resources to effect any required change.

In these instances, the use of assessment should be recognized as both a cost effective and efficient way to obtain relevant information pertaining to aspects of AQEE and to identify appropriate interventions for improving both policy and practice. For example, in the MLA Africa study, Chinapah et al.

(2000) focused on reporting disparities within countries regarding gender, location (urban/rural), and school type (private/public), rather than on ranking countries by their national mean scores, and on reporting the various factors influencing learner performance in each country. For many countries, this report highlighted areas in need of intervention, information that could be used by policy makers to effect change. Of course, the availability of relevant and useful information does not necessarily mean that the information will be employed in practice.

Technical expertise in conducting such studies is essential to obtaining relevant and reliable information. Assistance and, in some instances, funding to participate in regional or international studies (i.e., MLA, SACMEQ, TIMSS, PISA), is readily available. Most of these studies give priority to developing the capacity of participating countries, an issue we discuss in greater detail in our country-level landscapes.

Relatively low enrollment rates. In many developing nations, low secondary enrollment rates are caused by high dropout rates and limited availability of places. Although an obvious solution is to increase access, Bregman and Stallmeister (2001) note that access cannot be expanded rapidly without compromising quality, and caution that with increased access comes additional costs that many developing countries can ill afford. Among other strategies proposed to reduce expenditures, the authors argue for the improvement of internal efficiency by lowering high drop out and repetition rates in secondary schools. In this instance, assessment can be usefully applied. Teachers can be trained to use assessment practices to identify and address learner weaknesses and thus better prepare learners to progress to the next grade. On the assumption that the learners who find school interesting and relevant will not drop out, assessment can also be used to identify learner interests, which should then be incorporated in the daily interaction with learners. This is a good example of how assessment contributes to quality and efficiency, leading in part to improved access.

Inappropriate and inadequate systems and structures. The manner in which components of an education system are structured and articulated across different levels, as well as with the employment sector, affects the pathways by which learners are able to access higher and further education. These systems have to function efficiently in order to make any positive impact. However, in practice, this is difficult to attain. Bregman and Stallmeister (2001) note that support systems and education pathway links are weak or non-existent for many Sub-Saharan schools and advocate the establishment of national frameworks that would provide more rational choices of subject matter for both learners and parents. The authors also argue that the availability of national frameworks would enable learners to map their career pathways, thereby enhancing motivation and reducing dropout rates.

In South Africa, the old curriculum (under apartheid) was replaced by a new curriculum that was aligned with the new National Qualifications Framework (DoE, 1996). The new curriculum and the qualifications framework afforded greater flexibility in obtaining qualifications, allowed for the

recognition of prior learning, and encompassed both the formal and non-formal education sectors. Both initiatives required the attainment of specific standards before any qualifications could be obtained. Thus assessment practices were, and still are, critical to successful implementation, especially in regard to the recognition of prior learning and certification of adult learners. However, whether the framework will be able to address the concerns noted by Bregman and Stallmeister is yet to be determined.

At the 2001 Assessment of Mathematics and Science in Africa (AMASA) meeting, participants representing twelve African countries also advocated for the implementation of changes to the assessment system (Monyooe and Kanjee, 2001). Participants noted that the assessment systems in their countries were limited by their focus on selection for the next level of schooling, certification of learners, and the ranking of schools. The participants strongly recommended that assessment systems ought to facilitate effective teaching and learning, diagnose and evaluate the extent to which the countries' educational goals were being met, and direct learners into areas of further study for full self-development. This recommendation attests to the increasing recognition of the potential of assessment to play a stronger role in education reform and to support teachers in improving learning by providing timely and relevant information.

Examination systems. Public examinations play a critical role in determining the career paths of learners in most developing nations. These examinations are used primarily to select learners into the secondary or higher education sector and have a powerful effect on the education system (UNESCO, 2000b). Given the central and critical role of examinations, desirable improvements to the system can possibly be effected through the exam system. As noted by Noah and Eckstein (1992a), changes in examinations have been used as levers to promote change in education and society, to reform the curriculum, to shift effective control of the system away from—or toward—the center, and to achieve specific political goals. Examinations systems can also be used for accountability purposes and for improving the quality of education, especially if the exams replicate what is required in the classroom.

For example, in South Africa, a school-based assessment approach is being used to certify learners at the end of compulsory education (DoE, 2002). This system, known as the Common Tasks of Assessment (CTA) is administered to all ninth grade learners in all subject areas. The assessments are conducted over a number of days for each subject and include standardized as well as performance assessment tasks that encompass a range of appropriate and relevant assessment techniques and activities. The final grades of learners are determined by both their performances throughout the year, as summarized by end-of-year marks, and their performances on the final CTA examination. However, Kellaghan and Greaney (1992) note that public examinations intended to raise quality cannot be the same as those for selection, as the latter generally do not take account of the needs of the majority of learners who are not proceeding to the next level.

Prospects

The establishment of an effective education system, and the accompanying assessment system, to adequately address both the needs of different learners and long term societal requirements is extremely difficult to achieve and requires vast resources. This problem besets education systems in both developing and developed nations. It is especially acute at the secondary level because of the diverse needs and interests of learners. In attempting to address this challenge, Murnane and Levy (1996) argue for restructuring the education system for the teaching of the "new basic skills" to prepare learners to meet the needs of a changing economy. The authors suggest three sets of skills: 1) "hard skills" that include basic mathematics, problem solving, and high level reading; 2) "soft skills" that include the ability to work in groups and to make effective oral and written presentations; and 3) the ability to use personal computers to carry out simple tasks. However, the viability of these suggestions has yet to be demonstrated in practice.

ONGOING CONCERNS

The problematic characteristics of both the technical and substantive aspects of assessments (especially examinations) are a persistent problem in the developing world. The questionable quality of the data collected through the use of unreliable instruments or mediocre administration procedures leads to system inefficiencies and to cynicism among stakeholders. This cynicism is deepened when the integrity of the assessment system is compromised, a widespread phenomenon in many countries. In this section, we highlight a few aspects of assessment that are of particular concern.

One significant difficulty in assessment involves communicating assessment data so that all stakeholders—from education ministry officials to school staff to parents—can make effective use of the results. Although the primary technical goal in test construction is to design instruments that provide valid and reliable data, turning that data into understandable and useable information requires very different skills and is rarely done well, even in the developed world. Yet, without this last step, the potential for assessment to drive system improvements is seriously compromised. These difficulties are exacerbated in some contexts by the natural reluctance of officials at all levels to disseminate information that may reflect poorly on their performance.

Unfortunately, most assessments now in place, especially national assessments in language arts, math, and science, provide little or no information on whether learners have acquired the skills required to function effectively in society. An assessment system should have the capacity to yield information on a broad range of competencies that mark learners as contributing members of their communities. To this end, the relevant competencies should be specified in the curriculum, assessments frameworks, and test specifications. For example, the OECD specified the mathematical, reading, and scientific literacy competencies that young adults (fifteen year olds) should acquire, and

implemented a cross national study, PISA, to assess whether these young adults were prepared to meet the challenges of today's knowledge society (OECD, 2000). In the United States, the National Institute for Literacy has developed comprehensive standards for adult learners that focus on the following four categories: communication skills, decision-making skills, interpersonal skills, and lifelong learning skills (Stein, 2000). These studies and experiences could prove useful for developing nations to the extent that relevant information can be adapted to the specific context of their learners and their communities.

The cost of developing assessment systems is a critical factor for most decision makers in developing nations. Although there is little debate on the need and value of examinations, the decision to fund national assessments, especially in education systems that lack basic resources, e.g., textbooks or classroom furniture, is extremely difficult to make. A lack of information regarding cost and benefits complicates this decision (Kellaghan and Greaney, 2001; Levin, personal communication, 28 January 2004). In this context, Kellaghan and Greaney (1996) argue that unless decision makers are able to articulate how investments in national assessments will benefit the education system as a whole, resources might be better utilized for other activities.

Of course, there are other sources of unfairness. Inequities in opportunity to learn among different groups are reflected in corresponding disparities in performance (Chinapah, 2000). In some countries, especially those with many language groups, inequities can be exacerbated when the language of instruction and/or assessment is not the learner's native tongue (CAL, 2001). Learners who are more familiar with the language of instruction and assessment will be at considerable advantage. Education systems in countries comprising multiple language communities require greater resources to adequately address the needs of all learners. The technical difficulties in conducting an assessment (especially at the national level, although classroom assessment practices are also affected) increase in this context, as do the possibilities for unfairness. All instruments need to be translated into one or more languages without undue bias against any group, additional analyses are required, and reports must be published in multiple languages. In South Africa, the Grade 3 Systemic Evaluation (national assessment) study was administered in all eleven official languages, as instruction at that level is provided in all of the official languages (DoE, 2003). However, in some countries this may not be possible or feasible. For example, in Papua New Guinea, there are approximately 850 spoken languages of which about 500 are written. In practice, there may be no alternative to the use of a single language but there are ways to mitigate some of the difficulties associated with testing (see Heyneman and Ransom, 1990).

Finally, it bears repeating that high stakes tests can lead to unwanted consequences such as a narrowing of the curriculum and an undue emphasis on test preparation. This is particularly harmful when the learner cohort is heterogeneous with respect to goals. An earlier and more serious narrowing of the curriculum may have already occurred when, for example, schools chose

to focus on academic disciplines to the exclusion of more practical subjects, such as typing or woodwork, that are of interest and value to substantial numbers of learners. Ideally, separate examinations should be set for different purposes, but this is usually not practical for developing nations.

CONSTRAINTS ON IMPROVING ASSESSMENT

Any initiative undertaken to improve assessment practice must take account of the formal assessments that are currently in use. Although there is wide variation in both the quality of these assessments and how well they support broad educational improvement, each performs a useful function (at least to some degree), relies on existing capacity, and has some constituency that favors its continued employment. In general, any changes in an assessment system must take into account the broader education transformation agenda of the system and have the support of key constituencies, especially education department officials and teachers.

The proposed introduction of a new school leaving examination, for example, must consider not only how it might influence instruction (an important criterion for systemic validity) but also how it will perform the functions of the current examination. With respect to the latter point, the success of the proposal will depend on its impact "downstream" as well as the potential political repercussions. Heyneman and Ransom (1990) give an example of a failed attempt in Sri Lanka to eliminate the selection function of existing secondary school examinations in favor of the use of a new battery that was better aligned with a new curriculum. Opposition by some universities and the general public forced the government to reverse its decision.

In many countries, attempts at improvement are limited by lack of expertise and inadequate infrastructure. Lack of experience is often accompanied by underestimation of the complexity of the systems and the resources required to support testing functions such as design, development, administration, scoring, analysis, and reporting. Expertise needed in these areas can often be borrowed or bought. Internal capacity, on the other hand, is best built through collegial relationships with assessment professionals in other developing nations and with international experts. Fortunately, a number of successful programs are building needed capacity, even under less than ideal circumstances. Some operate under the auspices of the MLA program. For example, a handbook published by UNESCO (Chinapah, 1997) provides guidance on test and questionnaire design as well as various data analysis methodologies. Relevant activities associated with SACMEQ, PASEC, the Laboratorio, as well as the various IEA studies (TIMSS, PIRLS), have also contributed to the diffusion of expertise among member nations. These and similar initiatives contribute to enhancing the professionalism of testing agency staff.

However successful these efforts may be, corresponding improvements in classroom instruction and learning depend on conveying information back to schools and enhancing the capacity of teachers to make use of this information. Improving classroom-based assessment will probably prove more refrac-

tory, because it is so closely tied to the complexities of teacher education and in-service training. Cost is also a perennial problem. Budgets directed specifically at assessment are usually meager and focused on high stakes examinations. Building systemically valid assessments will require substantial additional expenditures. Although some of the required funds could be obtained through targeted grants, we believe the best strategy is to strengthen the connections between assessment and instruction. In the long run, this will permit instructional budgets to be used to support assessment improvement as well.

ROLE OF TECHNOLOGY

There is general agreement that the convergence of computers, multimedia, and broadband communication networks will have a substantial impact on education. In many developed countries, enormous sums have been expended on hardware and software, but evaluations of the consequences for learning have been mixed (Institute for Higher Education Policy, 1999; Angrist and Lavy, 2002). In speaking of the situation in the United States, Braun argues:

> [T]he core functions in most educational systems have not been much affected [by technology]. The reasons include the pace at which technology has been introduced into schools, the organization of technology resources (e.g., computer labs), poor technical support and lack of appropriate professional development for teachers. Accordingly, schools are more likely to introduce applications courses (e.g., word processing, spreadsheets) or "drill-and-kill" activities rather than finding imaginative ways of incorporating technology into classroom practice. While there are certainly many fine examples of using technology to enhance motivation and improve learning, very few have been scaled up to an appreciable degree (2003: 267).

The prospects for developing countries must certainly be dimmer, given the greater disadvantages under which they labor. In seeking ways to apply technology, it will generally be wise to resist the blandishments of complexity in favor of the charms of simplicity. At the same time, specific technological advances such as wireless Internet connections have the potential to greatly enhance access to content and expertise, with implications for both teacher professional development and student learning.

Undoubtedly, the priority for technology investments will be to support instruction. Technology can also be used to enhance the practice of assessment, through the training of teachers in formative assessment and the interpretation of the learner's work. Assessment specialists can also benefit from improved training and access to the latest software. Postlethwaite (private communication, 15 October 2002) cites an example of how staff at the International Institute of Educational Planning (IIEP) in Paris were able to train staff in Vietnam in sampling, data entry, and data cleaning through a series of video conferences. In Pakistan, UNICEF, in cooperation with a local NGO, set up a number of teacher training centers with access to the Internet

(UNICEF, 2001) in an effort to improve teacher skills. Teachers were assisted in using these centers to access information, as well as with translations and distributions of materials they found useful and relevant. However, the prospect of technology-based testing itself still seems rather remote, given the infrastructure requirements. This is especially true with respect to classroom practice. Focusing specifically on South America, Gehring (2004) notes that, at the system level, most countries have not been able to harness technology in ways that improve instruction. This is also true for those countries (e.g., Chile, South Africa) where concerted efforts and massive investments have been made to introduce appropriate technology for improving the education system (Gehring, 2004; Zehr, 2004).

In an analysis of technology's impact on assessment, it is important to distinguish between direct and indirect effects. Direct effects refer to the tools and affordances that change how assessment is actually practiced, while indirect effects refer, in part, to the ways in which technology helps to shape the political and economic environment in which decisions about priorities and resource allocation take place. The technical literature, naturally enough, tends to focus on direct effects; ultimately, funding decisions often have a determining influence on the evolution of assessment. As Braun (2003: 268) points out, "…one can argue that while science and technology give rise to an infinite variety of possible assessment futures, it is the forces at play in the larger environment that determine which of these futures is actually realized."

Significant improvements in the field of psychometrics have had a profound impact on the areas of instrument development, test analysis, and score reporting. For example, item response theory (IRT) is the underlying technology for most large-scale assessments. It allows test developers to explicitly describe the operating characteristics of individual test items and enables analysts to generate comparable scores for individuals who have possibly taken different sets of items. The latter property is crucial for large-scale assessments, such as cross-national surveys that seek to cover broad domains while keeping individual testing time to a minimum, or for computer adaptive tests that are intended to efficiently determine an individual's level of achievement (see Hambleton et al., 1991 for an overview of IRT, and Wainer et al., 2000 for an overview of Computerized Adaptive Testing).

The consequences have been significant improvements both in the quality and range of the information collected and in the methods of reporting information and comparing trends over time. However, these studies typically require high levels of expertise and considerable experience to be successful. In this regard, technology requirements can be a constraint, especially for those nations that have little or no access to the required expertise. In addition, assessment tools and information must be made available to teachers in order to ensure maximum benefit to learners. The pursuit of more complex technologies for assessment can limit use and often results in tools and data that are psychometrically immaculate but educationally bankrupt. For many developing nations, especially those in the early stages of transforming their education systems, the critical issue is striking a balance between the use of

sophisticated hard and soft assessment related technologies and the successful transformation of the system.

Bennett (2001) makes a strong case that rapidly evolving technology, and especially the near-ubiquity of the Internet, will have a substantial impact on testing. Indeed, he argues that it will lead to a reinvention of large-scale assessment just as it has in business practices (and other spheres of activity) around the world. The pervasiveness of technology in developing countries has already facilitated their ability to take advantage of advances in measurement and cognitive science to make testing, in all its roles, more useful and efficient. This argument is further advanced in Bennett (2001). However, he properly cautions, "The question is no longer whether assessment must incorporate technology. It is how to do it responsibly, not only to preserve the validity, fairness, utility and credibility of the measurement enterprise but, even more so, to enhance it" (Bennett, 2001: 15).

These considerations apply all the more to developing nations that can ill-afford major investments in technology that fail to yield commensurate returns. Accordingly, they have to be strategic and judicious in their planning, taking heed of the hard-won lessons learned by those nations that have gone before.

COUNTRY-LEVEL LANDSCAPES AND INTERNATIONAL INITIATIVES

This section is based on an extensive review of the education and assessment systems in a number of developing countries.[10] For each region, we selected five countries that reflect the variety of challenges confronting that region's education systems. We judged these countries as being fairly representative of the variation in that region, differing along such dimensions as size, population heterogeneity, and language diversity. However, we caution against making any generalizations, given the large differences in the local contexts that shape the education systems of all countries. Full reviews can be found at http://www.amacad.org/ubase.aspx.

For each country, we provide an introduction, listing relevant demographic information, details pertaining to the structure and management of the education sector in the country, as well as information on education expenditures. We then describe the assessment system, with a brief history, the current assessment capacity, and the uses of assessment within the different levels of the system. Information on those areas integral to the assessment system (i.e., teacher education and curriculum) is also provided, where available. In addition, descriptions of the various regional and international assessment initiatives that have recently been completed or are currently underway are also provided, because these initiatives comprise a critical component of the assessment system in many countries. For each initiative listed, we provide a brief overview, the objectives, areas of focus, and contact information.

10. The following countries were included in our review: (Africa) Mauritius, Morocco, Senegal, South Africa, Uganda; (Asia and Middle East) China, India, Indonesia, Jordan, Vietnam; (Latin America) Brazil, Chile, Colombia, Cuba, Mexico.

Country-level Landscapes

The review of country-level landscapes described above provides a snapshot of assessment practices and systems currently in use in developing nations and identifies best practices and highlights exemplars of initiatives that are having a positive impact. Of course, the enormous diversity in local conditions across the developing world makes it difficult to present any generic model for improving assessment systems or to present an overview of best practices. For this section, therefore, we have opted to identify four countries (presented alphabetically) that have made progress in developing their assessment systems and identify specific policies or strategies that contributed to this progress. Nonetheless, we still caution against generalizing the effectiveness of such practices to other countries.

Brazil. In Brazil, the current assessment system emerged from the government's decision to redefine the mission of the National Institute for Educational Studies and Research. The Institute is now charged with coordinating the development of educational assessment systems and organizing the information and statistical systems to assist policymakers at all levels in making appropriate decisions (Guimaraes de Castro, 2001b). As a result, "Today there is solid consensus among authorities, educators and specialists on the relevance of assessment systems as a guide to educational reforms and, above all, to the adoption of policies to improve the quality of education" (Guimaraes de Castro, 2001b: 5).

The assessment system in Brazil is a multi-level system based on voluntary exams administered at the end of secondary school (exams which are generally used for entry into the higher education sector); national assessments at fourth, eighth, and eleventh grade conducted every other year to monitor the quality, equality, and effectiveness of the education system; and in the higher education sector, mandatory examinations for final year undergraduate students to assess the quality of selected courses.

Highlights of the Brazilian system include: 1) the creation of a single federal agency, National Institute for Educational Studies and Research (INEP), to develop and coordinate the country's national assessment system, 2) the linking of the country's assessment and information systems for use in monitoring policy formulation and implementation, 3) the inclusion of the higher education sector (undergraduate) as a component of the national assessment system, 4) the use of the National Secondary Education Examinations, although voluntary, to identify career choices by learners and as an alternative entrance examination by universities, and 5) participation in international assessments—i.e., the Laboratorio and PISA.

Chile. Wolff (1998) regards the assessment system in Chile as one of the most comprehensive and best-managed assessment systems in Latin America. It has served as a strong tool for implementing required reforms and has led to increasing learning. The system was conceived in 1978 as the National Program to Measure the Quality of Chilean Basic Education (SIMCE), "to help the Ministry of Education and regional and provincial authorities supervise the

education system, evaluate individual schools, and assist in teacher in-service programs" (Wolff, 1998: 6). The SIMCE program tests all fourth and eighth grade learners (i.e., census testing) in Spanish and arithmetic and 10 percent of them in natural sciences, history, and geography. Information is also collected on learner personal development and attitudes, as well as on attitudes and background of teachers and parents, and on school efficiency. Assessment takes place in alternate years and costs approximately $5 per learner.

Highlights of the Chilean program include: 1) the use of census sampling with a total estimated expenditure per year of $2 million; 2) the gradual improvement of the assessments over time (e.g., school reports are now delivered in the first month of the school year, compared to earlier assessments when reports were delivered much later); 3) the increased use of assessment results, as noted by the development of intervention programs for low performing schools, and the allocation of financial rewards to schools with high numbers of low socioeconomic learners that show significant gains in scores (though the latter has resulted in schools inflating the number of low socioeconomic learners in order to increase the possibility of showing greater gain scores); 4) massive media campaigns directed at teachers, principals, and parents regarding the purpose and use of national assessments; 5) a comprehensive strategy for disseminating results and suggestions for teachers on how to improve leaner performance (including the distribution of relevant manuals and video tapes as well as the use of trained supervisors to explain results); 6) the assessment of the affective domain, i.e. self-concept, attitudes towards learning, peer and social relations, vocational orientation and values acquisition (Himmel (1996) argues that this aspect was not successful and recommends that it be dropped); 7) the categorization of schools by the typical socioeconomic level of the learner population and their administrative status (i.e., municipal, or public, subsidized private, or private), along with the reporting of schools results within the levels; and 8) participation in regional and international assessments Laboratorio (1997) and TIMSS (1999 and 2003).

Jordan. The current assessment system in Jordan was established as a result of the Examination Reform and Assessment Project (ERAP) initiated in 1995 and was supported by the British Department of International Development (DFID) in the form of technical assistance and training. It involved several different initiatives: 1) The General Secondary Education Certificate Examination, a certificate given at the end of twelfth grade, was revised to improve assessment procedures as well as the skills and expertise of teachers. The project aimed at assessing the full range of curriculum objectives, including the measurement of higher mental abilities such as analysis, evaluation, and problem-solving. 2) The new assessment system also introduced diagnostic assessments for improving the learning-teaching process, which entailed training teachers in the application and development of relevant materials. 3) The use of the "Investigation" approach by teachers was encouraged through teacher training and the preparation of relevant support materials, the development of learner skills in the areas of planning, collecting evidence, processing information, and presenting findings. 4) Achievement tests for the

tenth and eleventh grades were introduced to inform teachers and learners about the learners' performance regarding standards of achievement and to prepare learners for the examination of the general secondary education certificate, 5) National testing was implemented to obtain information on learners' performance in order to inform priorities regarding the curriculum, instructional materials and teaching methods. A 5 percent sample of tenth grade learners was assessed in six subjects. 6) The introduction of an assessment of practical training, which entailed the improvement of the assessment skills of vocation education teachers and the development of relevant materials.

Highlights of the ERAP included: 1) the use of external expertise for technical assistance and training in developing the assessment system, 2) the focus on classroom assessment and the learning and teaching process in the form of diagnostic testing and implementation of the "investigation" approach, 3) the emphasis on the vocational education sector as an integral part of the system, and 4) participation in regional and international assessments including TIMSS (1999 and 2003), and the International Assessment of Educational Progress (IAEP II) in 1991.

Mauritius. The major reforms impacting the current assessment system in Mauritius date back to the 1980s, with the introduction of the Certificate of Primary Education system (Dansingani, 2001). In 1993, Essential and Desired Learning Competencies were introduced in primary schools as part of the curriculum reform, laying the groundwork for a mechanism for setting minimum and "higher order" standards. Dansingani (2001) notes, however, that these reforms have led to a massive failure rate (one out of every three learners). Recent reforms introduced by the ministry (Ministry of Education and Scientific Research, 2004) in primary schools focus on the implementation of a National Literacy and Numeracy Strategy to improve the teaching of literacy and numeracy as well as learner performance. A key feature of this strategy is the provision of diagnostic tools (developed jointly by the Mauritius Examinations Syndicate (MES) and Mauritius Institute for Education) to assist teachers in the early identification of learner problems.

Currently the MES, the national examination body, administers all examinations conducted in the country. These include the following (International Bureau of Education, 2001): the Certificate of Primary Education, which is a national examination administered at the end of primary schooling, i.e., Standard (Grade), used by learners for entry into secondary schools; the High School Certificate, a national examination administered at the end of high school to certify completion of formal schooling that is also used for selection into the higher education sector; and the Cambridge School Certificate, an external examination administered in collaboration with the University of Cambridge Local Examination Syndicate for those learners intending to obtain an internationally recognized certificate and thereby gain entry into overseas universities. The MES is also responsible for the administration of various technical and vocational examinations as well as professional examinations for more than fifty bodies.

Highlights include: 1) the use of examination results (i.e., all the certification exams) by the MES to identify weaknesses in the performance of learners and the provision of information to all stakeholders in order to promote relevant reform, 2) the ranking of learners based on their performance on the Certificate of Primary Education, currently under review with the intention of eliminating the practice of ranking learners, 3) the use of both national and international examinations to certify learners' completion of formal schooling, intended to allow Mauritians easier access into foreign universities, 4) the identification of Essential and Desired Learning Competencies as a mechanism for improving the learning and teaching process, 5) the provision of "standardized" diagnostic instruments for numeracy and literacy to assist teachers in detecting learning difficulties, and 6) participation in SACMEQ (1998, 2002) and MLA (1994, 1999) studies.

Current Regional/International Initiatives

This section offers an overview of current and recent assessment initiatives that have been undertaken at a regional or international level, and provides a guide to additional resources to those persons interested in conducting similar studies. The regional/international initiatives we identified are those studies, consortia, and/or projects currently in progress (or recently completed) that advocated for the use of assessment as a means of promoting educational change. The initiatives we report on are neither a comprehensive catalog nor a representative sample of all such studies. Rather, we selected regional and international initiatives that highlight specific issues, including curriculum coverage, technical approach, capacity development, and reporting practices.

In our review of the literature pertaining to the studies, we noted that all international/regional studies sought to provide countries with relevant information for use by policy makers. However the underlying philosophies and/or approaches varied between and within the different projects depending on circumstances. These approaches can be categorized as follows:

- Emphasis on meeting the highest technical standards (which often means the use of the latest and most sophisticated techniques, methodologies, and software) versus using simpler and more cost-effective approaches to obtain relevant and useful data for the countries involved;
- Use of the curriculum as a basis for assessing learner performance versus assessment of general competencies; and
- Greater emphasis on local capacity development versus attaining project milestones and completion of reports.

In the last decade, an increasing number of countries have begun conducting their own national assessments as well as participating in international assessments (Beaton et al., 1999; Benveniste, 2002). There is limited information on the how these studies have influenced the various national education systems and the cost-benefit tradeoffs (Kellaghan and Greaney, 2001; Levin, personal communication, 28 January 2004). In his study on the reaction of participating countries to the TIMSS 1995 results, Macnab (2000:

12) concludes that while the results of TIMSS study provided participating countries with a valuable opportunity for instituting required reforms, "not all the countries made use of this opportunity; of those that did, not all were prepared to accept what was revealed; and that among those who did accept the verdict of TIMSS, there was not agreement as to the nature and depth of the changes required."

For many countries participating in an international assessment, the most significant benefit, besides the availability of additional information, is the access these studies provide to technical skills and expertise and the opportunity for capacity development. This is especially true for those initiatives where capacity building and sharing is noted as one of the primary objectives, e.g., MLA, SACMEQ, and PASEC (UNESCO, 2000b). Other studies (e.g., the Laboratorio, TIMSS, PIRLS) also provide significant opportunities for professional development, even though capacity building is not a specified objective.

Over the last twenty years, the number and scope of international assessments has increased dramatically, with greater participation by developing nations (Johnson, 1999; Martin et al., 2000). Along with this expansion, there has been significant improvement in the assessment design and methodologies employed, with greater attention paid to capacity development. These studies have also gained greater prominence in many countries and, accordingly, have generally improved the policy discourse. This is not surprising given that international assessments not only provide valuable comparative information but also allow participating countries to benefit from each others' experiences. Results from these studies provide additional insights that would be difficult to obtain from national surveys alone, and are generally viewed as more authoritative than within-country research by both the general public and policy makers (O'Leary, Madaus, and Beaton, 2001; Porter and Gamoran, 2002).

Although there are clear benefits to participation in international surveys, they also bring various challenges that countries should recognize. The most significant issue for policy makers is the degree to which information derived from these studies is relevant to the national context. The actual value of the study will depend not only on the quality of the data but also on the capacity of the country to effectively use the data (Johnson, 1999). Given the large variation in AQEE of participating countries, "the value of international studies may lie more in their potential for generating hypotheses about causal explanations than in their use for testing hypotheses" (Porter and Gamoran, 2002: 15). That is, findings from international studies may only highlight specific issues that would require further investigation by participating countries. Thus, policy makers should understand that participation is only the first step toward developing evidence-based education policy. For example, countries could leverage the experiences and expertise available internationally to mine existing data and develop new assessments in order to generate more detailed in-country information that could be useful to both policy makers and school personnel.

SUMMARY

In this paper we have identified four essential attributes of an education system—Access, Quality, Efficiency, and Equity—and a general criterion, systemic validity, which addresses the question of whether assessment strategies and instruments contribute constructively toward more fully realizing one or more of the four attributes. After describing different kinds of assessments, as well as the variety of roles assessment can play, we considered a number of relevant issues—from technical aspects of testing to obstacles to improving the quality and efficacy of assessment in developing nations. Specifically, we have recognized that those who would improve secondary education confront special challenges. These are related to the broader curriculum and the greater variety of desired outcomes at that level, both of which contribute to a substantial, and sometimes enormous, gap between needed and available capacity. In particular, current assessments typically take the form of public examinations in academic subjects for high school leaving and/or entrance to tertiary education, while many other possible functions are left unfulfilled.

Our presentation concluded with a number of case studies of assessment practices in selected countries. The general picture we drew shows that each country employs a range of assessments in various formats and settings but that, in many respects, the assessment systems do not function optimally. Our review indicates that there has been a global trend toward greater use of assessment. Increasingly, countries are conducting national assessments with the express purpose of obtaining information to improve the quality of education. Concurrently, the range and scope of public examinations are expanding and they continue to dominate the assessment landscape. The principal role of assessment is, still, to determine learner advancement and the awarding of qualifications.

Our case studies were accompanied by descriptions of a number of regional or international efforts. Over the last decade there has been a marked increase in the range and frequency of international assessments spearheaded by both regional initiatives (e.g., SACMEQ, Laboratorio) and international organizations (e.g., IEA, UNESCO/UNICEF, OECD). The results typically have confirmed worries about the low levels of achievement attained by learners in the developing nations. The UNESCO/UNICEF studies and regional initiatives have usually focused on developing nations, with capacity development as one of the primary objectives. In the IEA studies (TIMSS, PIRLS) and OECD studies (PISA) on the other hand, participants are mainly drawn from among the more developed nations, although a number of developing nations have also taken part. Although professional development is not specified as a primary objective, there is evidence that most developing nations have benefited from participation (Elley, 2002), which is often funded by third parties such as the World Bank. Elley argues that continued support is warranted in view of the quality of the data obtained, the concomitant increase in technical capacity, and the opportunity (seized by some countries) to introduce policy reforms grounded in evidence of the comparative weakness of their education systems.

The growing global prominence of assessment has led to considerable investments in establishing improved assessment systems. A number of developed nations or jurisdictions (e.g., United States, England, and Wales) have embarked on ambitious assessment programs to spearhead and reflect education reform strategies. They have recognized the need to build assessment capacity at all levels of the system and look to international assessments to provide external benchmarks to evaluate their progress. We can expect developing nations to follow suit, though at a decidedly slower pace.

At this stage, there is a generally favorable opinion on the value of participation in international assessments. A balanced analysis is provided by Rowan (2002), who indicates some of the issues that arise when developed nations attempt to use assessment results to inform policy. He and other commentators note the unfortunate tendency to focus on the "league tables" that present country-level rankings, when there is equal or greater value in the careful examination of within country differences and patterns. Johnson (1999) addresses similar issues from the perspective of developing nations. She is generally supportive of their participation, although she does indicate some of the technical and logistical obstacles they face and, more to the point, the difficulties they have in making good use of the information gleaned from the raw data.

In the literature on national and international assessments we have reviewed, there is only passing mention—and almost no serious discussion—of the use of classroom assessments and even less on the inclusion of assessment techniques in either teacher training curricula or teacher professional development programs. (In part, this may be due to the difficulty in obtaining such information from publicly available documents.) Undoubtedly, there has been insufficient attention to helping teachers to use assessment results effectively. For example, diagnostic feedback is not very common, so that test data are not often used to guide improvements in teaching and to enhance learning. This is unfortunate because there is an emerging consensus that formative assessment can be a powerful, cost-effective tool for improving teacher effectiveness (Black and Wiliam, 1998).

We argue, therefore, that there are good reasons for developing nations, and the organizations that assist them, to develop coordinated strategies that will enable these countries to more fully exploit the power of assessment to strengthen their education systems. The next and final section presents some thoughts on the matter.

STRATEGIES FOR MOVING FORWARD

We begin with the premise that essentially all nations seek to enhance their education systems and most consider assessment a legitimate and potentially useful tool in the improvement process. We readily admit that focusing on assessment alone can have only a modest positive impact: meaningful and substantial education reform requires sustained and coordinated changes in all system components. At the same time, risking the accusation of acting like

the person with a hammer who sees a world full of nails, we strongly believe that assessment, broadly conceived, should be more prominent in discussions of serious education reform.

A first step is to cultivate among all stakeholders (politicians, policy-makers, education bureaucrats, principals, teachers, and parents) a deeper appreciation of the power and cost-effectiveness of assessment. This requires a comprehensive framework to structure discussions about assessment and assessment capacity, as well as case studies that document the (favorable) returns on investment (ROI) yielded by well-planned investments in assessment, often as part of a broader education reform initiative. This is essential to generating needed support for strengthening and reforming assessment so that it can in fact play a more productive role in education improvement.

We believe that the four goals of Access, Quality, Efficiency, and Equity, together with the criterion of systemic validity, offer a starting point for a useful framework. In principle, they can be used to conduct meaningful prospective evaluations of assessment and other education-related reforms. Of course, documenting ROI is very difficult (Levin and McEwen, 2001) and the calculations are necessarily crude. Nonetheless, a start can be made, and as randomized control trials become more prevalent in education policy initiatives, the empirical base for such calculations will become firmer.

To the extent that rational policy development with respect to assessment is consistent with trying to increase systemic validity, nations face a multilevel design problem of great complexity, with goals and constraints at each level. The prerequisites for even a modicum of success are clarity, coherence, and consistency. By clarity we mean that the goals of education at each level, as well as the links between those goals and the relevant assessments, must be explicit, and that the results must be meaningful to all interested parties. By coherence we mean that the assessments at the different levels must articulate properly with one another. Finally, by consistency we mean that the development, implementation, and evolution of the assessment system must be carried out in a disciplined manner over a substantial period of time, at least five to ten years. These are difficult enough to realize for any nation, in the face of the usual bureaucratic inertia and the opposition of entrenched interests. For developing nations, such difficulties are often magnified by other social, economic, and political challenges—not to mention the problem of allocating adequate resources to the education sector.

It is critical to focus on generating more and higher quality data, and then turning those data into the information relevant to improving learning. This will require systematic planning leading to changes in assessment design, development, and reporting. Equally important, the capacity of the system to absorb and use those data effectively must also be enhanced. Among other things, this will involve extensive training of all professionals in the system, with special attention to teachers. As we have argued before, helping teachers to use classroom-based assessments and information from national assessments more effectively can contribute both to their subject matter knowledge and their pedagogy. A number of studies demonstrate that teacher profes-

sional development centered on evaluating the work of learners and reflecting on what is valued can be a powerful lever for change (Black and Wiliam, 1998; Wiliam, Lee, Harrison, and Black, 2004).

In addition to such training, the education system (at the school, district, and regional levels) must develop the communication channels, feedback mechanisms, and protocols that characterize data-driven organizations. These are well established in England (Olson, 2004). In the United States, there are a number of efforts focused on helping schools collect, organize, interpret, and use data effectively both for short- and long-term planning. Some, like the "Baldrige in Education Initiative," are modeled on similar undertakings in the corporate world. Their approach to school improvement relies heavily on information generated from assessments and is being used in many school districts across the United States. Other similar initiatives include the "School Information Partnership" and "Just for the Kids." Many related ideas and interesting examples can be found in Senge (2000).

However appealing in principle these data-driven ideas may be, there is no substitute for understanding the political, cultural, social, and commercial context of the hoped for changes at the various levels. Without sensitivity to context, no strategy is likely to be successful. For this reason alone, we expect that regional collaborations like the Laboratorio and SACMEQ, as well as international studies that account for local context, will have a continuing and critical role to play in this effort.

We have already mentioned the extensive participation of developing nations in the UNESCO initiatives and their increasing, but still sporadic, involvement in such international studies as TIMSS, PISA, and PIRLS. Both Johnson (1999) and Elley (2002) discuss the benefits and challenges that developing nations face. Rowan (2002) provides an excellent general treatment, although it is largely focused on the U.S. perspective. Although we acknowledge that many developing nations can indeed derive substantial value from full participation, we contend that for many others such participation may be of limited utility at this time, given the modest absorptive capacity of their systems. As an alternative, we suggest that the world community should encourage nations in the latter group to develop a strategic plan leading to full participation only after a number of years.

Participation in international studies is essentially a political decision and is not taken lightly, in part because of concern about the consequences of poor performance. One possibility is that nations could apply to join a study consortium as an "associate," with the opportunity to participate in the planning and test development, and then to administer the assessment with a primary focus on addressing specific national issues, as opposed to meeting international criteria. For example, experienced teachers could participate (informally) as part of a professional development program. The results of such a "toe-in-the-water" approach, along with the adaptation of ancillary materials generated by the consortium, could then be used to strengthen curriculum and instruction. Over time, participation could be expanded until the internationally stipulated criteria are reached.

Another possibility is to harness existing networks to take advantage of the valuable resources associated with these international studies. For example, in Africa, a group of nations under the aegis of the International Institute for Capacity-Building in Africa (the AMASA initiative) could organize itself to replicate some aspects of a study, again with a view to building capacity in an incremental but sustainable manner. It could draw on the relevant materials (ordinarily freely available) and invite experts to provide assistance. Each nation would be free to decide on the level of participation and how to employ the results.

Interested entities could also purchase services such as those provided by the Australian Council for Educational Research (ACER). ACER offers "International Benchmark Tests" in mathematics and sciences modeled on the TIMSS assessments. Learner results from these tests could be compared to the results of the different nations that participated in TIMSS. More important, carrying out such an assessment would provide a natural path to drawing on the pedagogical materials and secondary research related to TIMSS. As we indicated in our discussion of considerations for secondary education, there are many national and sub-national assessment programs that make valuable resources available, resources that can easily be adapted to the needs of developing nations.

UBASE means AQEE for all children and adolescents. In this regard, it is fully consistent with the goals that UNESCO has set for its member states (Chinapah, 2001). Under the right circumstances and properly employed, assessment can be a powerful tool for improving access, quality, and efficiency towards developing a more equitable system. But it can also be a crude tool, one that can lead to unintended and deleterious consequences if misapplied. These consequences can generally be avoided when assessment change is part of a comprehensive reform initiative that takes best advantage of what assessment can offer. In the final analysis, all education role players must acknowledge that "testing alone cannot improve learning, nor can it necessarily make education systems more responsive. But it does tune societies and governments alike to the possibilities of their schools and education systems. And, if the past is any guide to the future, well-designed and applied assessments can change the course of education reform and the menu inputs used to promote it" (Schiefelbein and Schiefelbein, 2003: 154).

References

American Federation of Teachers (AFT), National Council on Measurement in Education (NCME), and National Education Association (NEA). 1990. *Standards for Teacher Competence in Educational Assessment of Students*. Washington, DC: American Federation of Teachers.

Angrist, J., and V. Lavy. 2002. "New Evidence on Classroom Computers and Pupil Learning." *The Economic Journal* 112: 735–765.

Baker, D. P., M. Akiba, G. K. LeTendre, and A. W. Wiseman. 2002. "Worldwide Shadow Education: Outside School Learning, Institutional Quality of Schooling and Cross-national Mathematics Achievement." *Educational Evaluation and Policy Analysis* 23 (1): 1–17.

Beaton, A. E., T. N. Postlewaite, K. N. Ross, D. Spearitt, and R. M. Wolf. 1999. *The Benefits and Limitations of International Educational Achievement Studies*. Paris: UNESCO/International Institute for Educational Planning.

Bejar, I. I. 2002. "Generative Testing: From Conception to Implementation." Pp. 199–217 in *Item Generation for Test Development*, ed. S.H. Irvine and P. Kyllonen. Mahwah, NJ: Lawrence Erlbaum.

Bennett, R. E. 2002. "Inexorable and Inevitable: The Continuing Story of Technology and Assessment." *Journal of Technology, Learning, and Assessment* 1 (1). http://www.jtla.org.

———. 2001. "How the Internet Will Help Large-scale Assessment Reinvent Itself." *Education Policy Analysis Archives* 9 (5).

Benveniste, L. 2002. "The Political Structuration of Assessment: Negotiating State Power and Legitimacy." *Comparative Education Review* 46 (1): 89–115.

Bhatia, C. M. 1955. *Performance Tests of Intelligence under Indian Conditions*. Oxford, UK: Oxford University Press.

Bhuwanee, T. 2001. "Concept Paper, Regional Workshop on Secondary Education in Africa." Paper presented at the Regional Workshop on Secondary Education in Africa, Port Louis, Mauritius. December 3–6.

Black, P. 1998. *Testing: Friend or Foe*? London: Falmer Press.

Black, P., and D. Wiliam. 1998. "Inside the Black Box: Raising Standards through Classroom Assessment." Kings College London, School of Education. http://www.kcl.ac.uk/depsta/education/publications/blackbox.html.

Bloom, D. E., and J. E. Cohen. 2002. "Education for All: An Unfinished Revolution." *Daedalus* 131 (3): 84–95.

Bordia, A. 1995. "Indian Education." Pp. 430–439 in *International Encyclopedia of National Systems of Education*, ed. N. Postlethwaite. Oxford/New York/Tokyo: Elsevier Science.

Braun, H. I. 2000. "A Post-modern View of the Problem of Language Assessment." Pp. 263–272 in *Fairness and Validation in Language Assessmen*t, ed. A. J. Kunnan. Cambridge, UK: Cambridge University Press.

———. 2003. "Assessment and Technology." Pp. 267–288 in *Optimizing New Modes of Assessment*, ed. M. Segers, P. Dochy, and E. Cascallar. Dordrecht, Netherlands: Kluwer.

Braun, H. I., I. I. Bejar, and D. M. Williamson. 2006. "Rule-based Methods for Automated Scoring: Application in a Licensing Context." In *Automated Scoring of Complex Constructed Response Tasks in Computerized Testing*, ed. D. M. Williamson, R. J. Mislevy, and I. I. Bejar. New Jersey: Lawrence Erlbaum Associates.

Braun, H. I., and P. W. Holland. 1982. "Observed-score Test Equating: A Mathematical Analysis of Some ETS Equating Procedures." Pp. 9–49 in *Test Equating*, ed. P.W. Holland, and D.B. Rubin. New York: Academic Press.

Bregman, J., and S. Stallmeister. 2001. "Secondary Education in Africa (SEIA): A Regional Study of the Africa Region of the World Bank." Paper presented at the Regional Workshop on Secondary Education in Africa, Port Louis, Mauritius, December 3–6.

Center for Applied Linguistics (CAL). 2002. *Expanding Educational Opportunity in Linguistically Diverse Societies.* Washington, DC: CAL.

Chinapah, V. 2001. "Quality Education: UNESCO Position Paper." Paper presented at the Regional Workshop on Secondary Education in Africa, Port Louis, Mauritius. December 3–6.

———. 1997. *Handbook on Monitoring Learning Achievement: Towards Capacity Building*. Paris: UNESCO.

Cohen, J. E., and Bloom, D. E. 2005. "Cultivating Minds." *Finance and Development* 42 (2): 8–14.

Chinapah. V., M. H'ddigui, A. Kanjee., W. Falayojo, C. O. Fomba, O. Hamissou, A. Rafalimanana, and A. Byamugisha. 2000. *With Africa for Africa: Towards Quality Education for All*. Pretoria: Human Sciences Research Council.

Cronbach, L. J. 1990. *Essentials of Psychological Testing*, 5th ed. New York: Harper & Row.

Dansingani, H. 2001. "Mauritius." Pp. 52 in *Curriculum Development and Education for Living Together: Conceptual and Managerial Challenges in Africa*, ed. John Aglo and Mankolo Lethoko. Final report of the International Bureau of Education seminar held in Nairobi, Kenya, June 25–29.

Department of Education (DoE). 1996. "Lifelong Learning through a National Qualifications Framework." Report of the Ministerial Committee on the NQF. Pretoria: Department of Education, South Africa.

———. 2000. "A South African Curriculum For The Twenty First Century." Report of the Review Committee On Curriculum 2005. Pretoria: Department of Education, South Africa.

———. 2002. *Guidelines for the Assessment of Learners in Grade 9 in 2002*. Pretoria: Department of Education, South Africa.

——. 2003. *National Report on Systemic Evaluation: Mainstream Education – Foundation Phase*. Pretoria: Department of Education, South Africa.

Dugger, C. W. 2004. "To Help Poor be Pupils, not Wage Earners, Brazil Pays Parents." *The New York Times*, January 3, 2004: A1, A6.

Eckstein, M. A., and H. J. Noah. 1992. *Examinations: Comparative and International Studies.* Oxford: Pergamon Press.

Elley, W. B. 2002. "Evaluating the Impact of TIMSS-R in Low- and Middle-Income Countries: an Independent Report on the Value of World Bank Support for an International Survey of Achievement in Mathematics and Science." World Bank: Unpublished Report.

Feldt, L. S., and R. L. Brennan. 1989. "Reliability." Pp. 105–146 in *Educational Measurement,* 3rd edition, ed. R. L. Linn. New York: American Council on Education & Macmillan.

Feuer, M. J., and K. Fulton. 1994. "Educational Testing Abroad and Lessons for the United States." *Educational Measurement: Issues and Practices* XX (2): 31–39.

Foster, P. J. 1992. "Commentary." Pp. 121–126 in *Examinations: Comparative and International Studies*, ed. M. A. Eckstein, and H. J. Noah. Oxford: Pergamon Press.

Frederiksen, J. and A. Collins. 1989. "A Systems Approach to Educational Testing." *Educational Researcher* 18 (9): 27–32.

Fuller, B. 1987. "What School Factors Raise Achievement in the Third World?" *Review of Educational Research* 57 (3): 255–292.

Gaspirini, L. 2000. "The Cuban Education System: Lessons and Dilemmas." *Country Studies: Education Reform and Management Series* 1 (5): July 2000. Washington, DC: The World Bank.

Gehring, H. 2004. "South America. Technology Counts 2004. Global Links: Lessons from the World." *Education Week* 35: 50–54.

Govinda, R. 1998. *Using Testing as a Tool to Improve School Quality: Reflections on Indian Policies and Practices*. New Delhi: National Institute of Educational Planning and Administration.

Greaney, V., and T. Kellaghan. 1995. *Equity Issues in Public Examinations in Developing Countries*. Washington, DC: World Bank.

——. 1996. *Monitoring the Learning Outcomes of Educational Systems*. Washington, DC: World Bank.

Guimaraes de Castro, M. H. 2001a. "Education Content and Learning Strategies for Living Together in the 21st Century." A Report from the 46th Session of the International Conference on Education, Geneva, Switzerland, UNESCO.

——. 2001b. "Education Assessment and Information Systems in Brazil." Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. http://www.inep.gov.br/idiomas/ingles/.

Hallak, J. 2000. "Globalisation and its Impact on Education." Pp. 21—40 in *Globalisation, Educational Transformation and Societies in Transition*, ed. T. Mebrahtu, M. Crossley, and D. Johnson. Oxford, UK: Symposium Books.

Hambleton, R. K., H. Swaminathan, and H. J. Rogers. 1991. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publishers.

Hannum, E., and C. Buchmann. 2003. *The Consequences of Global Educational Expansion: Social Science Perspectives.* Cambridge, MA: American Academy of Arts and Sciences.

Heyneman, S. P. 1987. "Uses of Examinations in Developing Countries: Selection, Research, and Education Sector Management." *International Journal of Educational Development* 7 (4): 251–263.

Heyneman, S. P., and A. W. Ransom. 1990. "Using Examinations and Testing to Improve Educational Quality." *Educational Policy* 4 (3): 177–192.

Himmel, E. 1996. "National Assessment in Chile." Pp. 111–128 in *National Assessments: Testing the System*, ed. Paud Murphy, Vincent Greaney, Marlaine E. Lockheed, and Carlos Rojas. Washington, DC: World Bank.

Holsinger, D. B., and R. N. Cowell. 2000. *Positioning Secondary Education in Developing Nations*. Paris: UNESCO/IIEP.

Horkay, N, ed. 1999. *The NAEP Guide: A Description of the Content and Methods of the 1999 and 2000 Assessments* (NCES 2000-456). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Institute for Higher Education Policy. 1999. *A Review of Contemporary Research on the Effectiveness of Distance Learning in Higher Education*. Washington, DC: Institute for Higher Education Policy.

International Bureau of Education, United Nations Education, Scientific, and Cultural Organization (IBE-UNESCO). 2001. World Data on Education, IBE-UNESCO. http://www.ibe.unesco.org/International/Databanks/Wde/profilee.htm.

Irvine, S. H. 2002. "Item Generation for Test Development: An Introduction." Pp. xv–xxv in *Item Generation for Test Development*, ed. S. H. Irvine, and P. Kyllonen. Mahwah, NJ: Lawrence Erlbaum.

Jenkins, D. 1996. "The Role of Assessment in Educating for High Performance Work: Lessons from Denmark and Britain." Pp. 381–427 in *Linking School and Work: Roles for Standards and Assessment*, ed. L. B. Resnick, and J. G. Wirt. San Francisco, CA: Jossey-Bass Publishers.

Johnson, S. 1999. "International Association for the Evaluation of Educational Achievement Science Assessment in Developing Countries." *Assessment in Education* 6 (1): 57–73.

Kanjee, A. 2003. "Using Assessment Resource Banks to Improve the Teaching and Learning Process." Pp. 59–71 in *Improving the Quality of Primary Education: Good Practices and Emerging Models of District Development*. Pretoria: District Development Support Program/Research Triangle Institute.

Kaufman, P., D. Bradby, and P. Teitelbaum. 2000. "High Schools that Work and Whole School Reform: Raising Academic Achievement of Vocational Completers through the Reform of School Practice." National Center for Research in Vocational Education. http://www.sreb.org/programs/hstw/publications/special/ncrrpt.doc.

Keeves, J. P., ed. 1997. *Educational Research, Methodology and Measurement: An International Handbook*, 2nd ed. New York: Pergamon.

Kellaghan, T. 1992. "Exam Systems in Africa: Between Internationalization and Indigenization." Pp. 95–104 in *Examinations: Comparative and International Studies*, ed. M. A. Eckstein and H. J. Noah. Oxford: Pergamon Press.

———. 1996. "Can Public Examinations be Used to Provide Information for National Assessments." Pp. 33–48 in *National Assessments: Testing the System*, ed. P. Murphy, V. Greany, M. E. Lockheed, and C. Rojas. Washington, DC: World Bank.

Kellaghan, T., and V. Greaney. 1992. *Using Examinations to Improve Education: A Study in Fourteen African Countries*. Washington, DC: World Bank.

———. 2001. *Using Assessment to Improve the Quality of Education*. Paris: UNESCO.

———. 2003. "Monitoring Performance: Assessment and Examinations in Africa." Paper presented at the Association for the Development of Education in Africa (ADEA) Biennial Meeting: Grand Baie, Mauritius, December 3–6.

Kellaghan, T., and P. J. McEwen. 2001. *Cost-effectiveness Analysis*, 2nd ed. Thousand Oaks, CA: Sage.

Levin, H. M., and P. J. McEwan, eds. 2000. *Cost-effectiveness Analysis: Methods and Applications,* 2nd ed. Thousand Oaks, CA: Sage.

Lewin, K., and F. Caillods. 2001. *Financing Secondary Education in Developing Countries: Strategies for Sustainable Growth*. Paris: UNESCO/IIEP.

Lievesley, D. 2001. "Making a Difference: A Role for the Responsible International Statistician." *The Statistician* 50 (4): 367–406.

Little, A. 1990. "The Role of Assessment, Re-examined in International Context." Pp. 9–22 in *Changing Educational Assessment: International Perspectives and Trends*, ed. P. Broadfoot, R. Murphy, and H. Torrance. London: Routledge.

———. 1992. "Decontextualizing Assessment Policy: Does it Make Economic Sense?" Pp. 127–132 in *Examinations: Comparative and International Studies*. M. A. Eckstein, and H. J. Noah. Oxford: Pergamon Press.

Lockheed, M. E. 1995. "Educational Assessment in Developing Countries: The Role of the World Bank." Pp. 133–147 in *International Perspectives on Academic Assessment*, ed. T. Oakland, and R. Hambleton. Boston: Kluwer.

Lockheed, M.E., and H. M. Levin. 1993. "Creating Effective Schools." In *Effective Schools in Developing Countries*, The Stanford Series on Education & Public Policy. Stanford, CA: Stanford University.

Macnab, D. S. 2000. "Forces for Change in Mathematics Education: The Case of TIMSS." *Education Policy Analysis Archives* 8 (15).

Makgmatha, M. 1998. *Assessment Practices in Asia and Oceania: Review of Education System in Indonesia, Malaysia, India, China, Australia, and New Zealand*. Unpublished report. The Education and Training Assessment Studies Unit. Pretoria: Human Sciences Research Council.

Martin, M. O., I. V. S. Mullis, E. J. Gonzales, K. D. Gregory, T. A. Smith, S. J. Chrostowski, R. A. Garden, and K. M. O'Connor. 2000. *TIMSS 1999 International Science Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade*. Chestnut Hill, MA: Boston College.

Messick, S. J. 1989. "Validity." Pp. 13–103 in *Educational Measurement*, 3rd edition, ed. R. L. Linn. New York: American Council on Education & Macmillan.

Ministère de l'Education Nationale et de la Jeunesse. 2003. L'Education au Maroc. http://www.men.gov.ma/, accessed February 2004

Ministry of Education, India. 2003. Central Board of Secondary Education. http://www.education.nic.in, accessed February 2004.

Ministry of Education and Scientific Research. 2004. "National Literacy and Numeracy Strategy." http://www.gov.mu/portal/site/education (See Publications: Reports), accessed February 2004.

Ministry of Education – Jordan. 2003. General Directorate of Examinations and Tests. http://www.moe.gov.jo/ex/eng/IntroductionE.html, accessed February 2004.

Ministerio de Educación Nacional. 2001. *Informe Nacional sobre el Desarrollo de la Educación en Colombia*. Bogota: Ministerio de Educación Nacional.

Mislevy, R. J. 2003. "On the Structure of Educational Assessments." *Measurement: Interdisciplinary Research and Perspectives* 1: 3–62.

Moegiadi and Jiyono. 1994. "Indonesia: System of Education." Pp. 2784–2792 in *The International Encyclopedia of Education*, ed. T. Husen and T. Postlethwaite. Oxford: Pergamon.

Monyooe, L., and A. Kanjee. 2001. "Final Report." Regional workshop on Assessment of Mathematics and Sciences in Africa (AMASA) held in Johannesburg, South Africa, November 26–30.

Mullis, I. V. S., M. O. Martin, et al., eds. 2002. *PIRLS 2001 Encyclopedia*. Chestnut Hill, MA: International Study Center, Boston College.

Murnane, R. J., and F. Levy. 1996. *Teaching the New Basic Skills: Principles for Educating Children to Thrive in a Changing Economy*. New York: The Free Press.

National Research Council. 2002. *Methodological Advances in Cross-national Surveys of Education Achievement.* A.C. Porter and A.Gamoran, eds. Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Ndoye, M. 2002. "Reaching Schools: Where Quality Starts." *ADEA Newsletter* (14) 3. Paris: Association for the Development of Education in Africa.

Noah, H. J., and M. A. Eckstein. 1992a. "Introduction." Pp. 5–6 in *Examinations: Comparative and International Studies*, ed. M. A. Eckstein and H. J. Noah. Oxford: Pergamon.

——. 1992b. "The Two Faces of Examinations: A Comparative and International Perspective." Pp. 147–170 in *Examinations: Comparative and International Studies*, ed. M. A. Eckstein and H. J. Noah. Oxford: Pergamon.

O'Leary, M., G. F. Madaus, and A. E. Beaton. 2001. "Consistency of Findings Across International Surveys of Mathematics and Science Achievement: A Comparison of IAEP 2 and TIMSS." *Education Policy Analysis Archives* (8) 43.

Obanya, P. 2002. *Revitalizing Education in Africa.* Nigeria: Sterling-Horden Publishers.

Olson, L. 2004. "Value Lessons." *Education Week* 23 (May 5): 36–40.

Organisation for Economic Co-operation and Development (OECD). 2000. *Knowledge and Skills for Life: First results from PISA 2000*. http://www1.oecd.org/publications/e-book/9601141E.pdf.

Porter, A. C., and A. Gamoran. 2002. "Progress and Challenges for Large-Scale Studies." Pp. 3–23 in *Methodological Advances in Cross-national Surveys of Educational Achievement*, ed. A.C. Porter and A. Gamoran. Board of International Comparative Studies in Education. Washington, DC: National Academies Press.

Rajput, J. S. 2003. "Towards a Dynamic Evaluation System, National Council of Educational Research and Training." http://ncert.nic.in/icutodyev.htm.

República de Colombia. 1999. *Informe de Países*. Bogotá: Education For All.

Resnick, L. B., and J. G. Wirt, eds. 1996. *Linking School and Work: Roles for Standards and Assessment*. San Francisco, CA: Jossey-Bass Publishers.

Robitaille, D. F., ed. 1997. *National Contexts and Science Education: An Encyclopedia of the Education Systems Participating in TIMSS*. Vancouver: Pacific Educational Press.

Rojas, C., and J. M. Esquivel. 1998. "Los Sistemas de Medición del Logro Académico en Latinoamérica." *LCSHD paper series* 1(25): 7.

Rowan, B. 2002. "Large-scale Cross National Surveys of Educational Achievement: Pitfalls and Possibilities." Pp. 321–349 in *Methodological Advances in Cross-national Surveys of Educational Achievement*, ed. A.C. Porter and A. Gamoran. Board of International Comparative Studies in Education. Washington, DC: National Academies Press.

Rychen D. S., and L. H. Salganik, eds. 2003. *Key Competencies for a Successful Life and a Well-Functioning Society*. Göttingen: Hogrefe & Huber Publishers.

Schiefelbein, E. 1993. "The Use of National Assessments to Improve Primary Education in Chile." Pp. 117–146 in *From Data to Action: Information Systems in Educational Planning*, ed. D. W. Chapman and L. O. Mählck. Paris: UNESCO, International Institute for Educational Planning.

Schieflbein, E., and P. Schiefelbein. 2003. "From Screening to Improving Quality: the Case of Latin America." *Assessment in Education* (10) 2: 141–154.

Secretaría de Educación Publica. 2001. *Educational Development: National Report of Mexico*. Mexico: Secretaría de Educación Publica.

Secretary's Commission on Achieving Necessary Skills. 1992. *Learning A Living: A Blueprint for High Performance—A SCANS Report For America 2000*. Washington, DC: U.S. Department of Labor.

Senge, P. M. 2000. *Schools that Learn: A Fifth Discipline Fieldbook for Educators, Parents, and Everyone Who Cares about Education*. New York: Doubleday.

Stein, S. 2000. *Equipped for the Future Content Standards: What Adults Need to Know and be Able to do in the 21st Century*. Washington, DC: National Institute for Literacy.

Swanson, L., and M. L. Stocking. 1993. "A Model and Heuristic for Solving Very Large Item Selection Problems." *Applied Psychological Measurement* 17: 151–166.

Umar, J. 1996. "Grappling with Heterogeneity: Assessment in Indonesia." Pp. 233–247 in *Assessment in Transition: Learning, Monitoring and Selection in International Perspective*, ed. Little and A. Wolf. Oxford: Pergamon Publishers.

van der Linden, W. J., issue ed. 1998. *Applied Psychological Measurement* 22 (3), Special Issue: Optimal Test Assembly. Thousand Oaks: Sage Publications.

Walberg, H. J., and G. D. Haertel. 1990. *The International Encyclopedia of Educational Evaluation*. Oxford, UK: Pergamon Press.

UNESCO. 1998. "Wasted Opportunities: When Schools Fail." In *Education for All: Status and Trends 1998*. Paris: UNESCO.

———. 2000a. *The Dakar Framework for Action. Education for All: Meeting our Collective Commitments*. Paris: UNESCO.

———. 2000b. *Status and Trends 2000: Assessing learning achievement*. Paris: UNESCO.

UNDP/UNESCO/UNICEF/World Bank. 1990. *World Conference on Education for All, Meeting Basic Education Needs*. Final Report. Paris: UNESCO.

UNICEF. 2001. "Education Technology." *Education Update* 4 (2). New York: Education section, UNICEF.

———. 2002. "Learning Achievement." *Education Update* 5 (3). New York: Education section, UNICEF.

———. 2003. "Girl's Education." http://www.unicef.org/girlseducation/index.html, accessed September 2003.

Wainer, H., ed. 2000. *Computerized Adaptive Testing: A Primer*. Mahwah, NJ: LEA.

Walpole, M., and R. J. Noeth. 2002. *The Promise of Baldrige for K-12 Education*. Iowa City, IA: ACT.

Watkins, K. 2000. *The Oxfam Education Report*. Oxford, UK: Oxfam.

Wiliam, D., C. Lee, C. Harrison, and P. Black. 2004. "Teachers Developing Assessment for Learning: Impact on Student Achievement." *Assessment in Education: Principles, Policy, and Practice* 11 (1): 49–65.

Wolff, L. 1998. "Educational Assessment in Latin America: Current Progress and Future Challenges." PREAL. http://www.iadialog.org/publications/preal/preal11e.html.

Zehr, M. 2004. "Africa. Technology Counts 2004. Global Links: Lessons from the World." *Education Week* 35: 56–59.

CHAPTER 2

Evaluating Educational Interventions in Developing Countries

ERIC BETTINGER

Randomized experiments are an increasingly popular means of evaluating educational reforms throughout the world. Innovative researchers, policymakers, and foundations have implemented randomized evaluations of programs ranging from educational vouchers in Colombia, to teacher supply in India, to textbook provision and deworming in Kenya. The use of experimental approaches in the provision of social services is not a new practice; policymakers and researchers have long recognized that experimental approaches can produce reliable evidence of the efficacy (or inefficacy) of social-service provision.

In recent years, the use of randomized implementation and evaluation in the provision of public services has garnered increased support from policy organizations throughout the world. The World Bank, for example, advocates that countries introduce new social programs using random assignment (e.g., Newman, Rawlings, and Gertler, 1994). They argue that "randomized designs are generally the most robust of the evaluation methodologies" (World Bank, 2003). Evidence from randomized experiments is often the most persuasive to policymakers. For example, the "No Child Left Behind" legislation in the United States ties local school funding to "scientifically based research." The act defines scientifically based research, in part, as research that

> …is evaluated using *experimental or quasiexperimental designs* in which individuals, entities, programs, or activities are assigned to different conditions and with appropriate controls to evaluate the effects of the condition of interest, *with a preference for random-assignment experiments* (PL 107–110: 1965, emphasis added).

Although randomized evaluation can produce persuasive and perhaps conclusive evidence, it also has significant drawbacks. First, implementation of a randomized evaluation may present an ethical dilemma. Inherent in a randomized evaluation is the condition that while one group experiences an innovation (the treatment group) another does not (the control group). Withholding treatment from an individual is difficult to justify, especially when it is believed to be beneficial,[1] and withholding treatment from some

1. See Cook and Panye (2002) for further discussion on the ethicality of random assignment for educational interventions.

groups may be politically unpopular. Furthermore, politicians often have incentives to overstate the effectiveness of their programs or to publish the effects of only successful programs.

Even if researchers and policymakers take care of ethical and political considerations, other problems may remain. For example, randomized evaluation can be expensive and time-consuming, especially in the field of education. Besides the cost of the innovation itself, randomization also entails substantial administrative costs. Researchers and policymakers must administer the randomization and track both treatment and control students over multiple years, as it often takes multiple years of treatment before researchers can gauge the effects of an educational innovation. Likewise, deviations from randomization may limit the validity of subsequent evaluation.

Randomization also presents issues of internal and external validity. If an evaluation has internal validity, then comparisons between treatment and control groups provide meaningful results. In theory, randomization provides an unbiased group through which to gauge the effects of the program on the application pool. Internal validity can be threatened for a number of reasons. For example, if researchers are unable to gather follow-up data for even a portion of the control or treatment groups, then comparisons of the control and treatment groups may not provide accurate estimates of the program's effect. External validity may also be problematic for randomized studies. If an evaluation has external validity, the measured effects of the treatment on the applicant pool will be similar in other populations. Oftentimes, however, researchers apply experimental designs on samples that differ substantially from the overall population. Additionally, many experiments are small, and although they may provide an accurate estimate of the partial equilibrium effect, they may not be able to estimate the general equilibrium effect.[2] The impact of the treatment on a small group, as shown in the partial equilibrium effect, may not hold if substantial changes are required for the expansion of the social service to a more general population. For example, a small voucher program may not affect the supply of schools in a city; however, a large voucher program may provide incentives for the creation of new schools. These schools may be better or worse than the existing schools, but they provide a different quality of education. The voucher may not only affect students directly by allowing them to attend different schools, but it may also affect students by changing the supply of available schools.

Since the early 1970s, economists, education researchers, and education practitioners have studied a range of educational phenomena. Only a few of these educational studies have exploited randomization. This paper discusses how the few studies with randomization have augmented the larger body of evidence on educational innovations in developing countries.[3] It highlights some of the strengths and weaknesses of randomized evaluations. The paper

2. Partial equilibrium refers to the effect of a policy holding the institutions providing the public service and their surrounding infrastructure constant. The general equilibrium effect allows the institutions and the infrastructure to change in response to the program.

3. Kremer (2003) and Pritchett (2002) also review recent randomized experiments.

speaks less to the cost-effectiveness of such programs (see Kremer, 2004) and instead focuses on the knowledge gained from randomized experiments.

This paper's focus on randomization is not intended to devalue the contributions that other types of studies have made to the field of education research. Randomization is one approach that has gained popularity and, under some conditions, can be more persuasive and compelling than other types of research. However, as this paper argues, other types of research have also provided significant insights into educational knowledge. For example, rarely does randomized research comment on ways to improve the quality of implementation. Oftentimes, alternative approaches can complement and provide important synergies to our understanding of the implications of randomized trials. However, in its focus on randomization, this paper is narrower in its scope.

The paper starts by presenting a brief overview of "selection bias" and demonstrates how randomization may help researchers avoid such bias. It then discusses why randomized research is not relied upon more heavily and presents a simple model to demonstrate why policymakers may be reluctant to undertake educational projects that rely on randomized evaluation. The second section discusses four types of educational innovations: school infrastructure development, student health, the provision of textbooks and other learning aids, and incentives to schools, teachers, and students. For each type of educational innovation, the paper reviews both non-experimental and experimental evidence. It also attempts to identify specific cases in which evidence from randomized policies has improved policymakers' and researchers' understanding of educational phenomena.

THE PROMISE OF RANDOMIZATION

Randomized experiments have the potential to produce unbiased estimates of a program.[4] Studies without randomization are susceptible to selection bias, the potential bias arising when participants in a given intervention systematically differ from non-participants. For example, economists have long been interested in knowing how private schooling affects student outcomes relative to public schooling.[5] Studies of private schooling often compare students in public schools to students in private schools. Unfortunately, such comparisons may not be entirely valid. Students who attend private schools typically come from more affluent homes where education may be more highly valued than it is in the homes of students in public schools. Even if private school students were to attend public schools, they might still perform better than other public school students because of a difference in home and family environment.

4. There are a number of studies that characterize randomized experiments and their technical strengths and weaknesses (e.g., Meyer, 1995; Angrist and Krueger, 2001).

5. For studies in developing countries, see: Bashir, 1997; Bedi and Garg, 2000; Alderman, Orazem, and Paterno, 2001. In developed countries: Neal, 1997; Evans and Schwab, 1995.

Typically, researchers using non-experimental research designs must include ample controls for any characteristic that may distinguish the students who participate in the program from the students who do not; yet, even if substantial controls are included in the model, selection bias may not be eradicated. If there is an unobservable characteristic that affects both the likelihood that students participate and their performance in the activity (e.g., ability), then estimates based on non-experimental research designs may still be biased.

Randomized experiments offer a solution to this problem. Students randomly selected to participate should not differ from students who are randomly not included in the program. Mathematically, it is possible to demonstrate how randomness may eliminate selection bias. For simplicity, I return to the example of private schooling. Suppose that there are two types of students, high (H) and low (L) ability. These students attend both private and public schools. The following notation can simplify the following discussion of a hypothetical research program based on test scores:

α_H = Test score effect of being high ability

α_L = Test score effect of being low ability

δ_{pub} = Test score effect of public school

δ_{priv} = Test score effect of private school

These quantities are assumed to exist, although researchers may observe only a test score. For the sake of the example, I will assume that student achievement is a combination of both student ability and a school effect. Hence, if a high-ability student attends private school, her test score would be equal to $\alpha_H + \delta_{priv}$. Similarly, the test score of a low-ability student in public school would be equal to $\alpha_L + \delta_{pub}$.

To demonstrate selection bias, we can further suppose that students perfectly sort by high and low ability into private and public schools respectively. If this is the case, then:

$$\text{Average Achievement in Public Schools} = \alpha_L + \delta_{pub}$$

$$\text{Average Achievement in Private Schools} = \alpha_H + \delta_{priv}$$

A naïve research design would attempt to identify the effects of private schools by looking at the differences in these quantities and attribute this difference to be the "effect of private schooling." However, this comparison is flawed. We cannot distinguish between the effects of private schooling and the effects of ability differences.

To identify the effects of attending public school versus private school, we would ideally like to observe the difference between δ_{priv} and δ_{pub}. We could observe this by comparing the achievement of a student of the same ability in both private and public schools:

$$\begin{aligned}&\text{Achievement of High Ability in Private School} - \\ &\text{Achievement of High Ability in Public School} \\ &= (\alpha_H + \delta_{priv}) - (\alpha_H + \delta_{pub}) \\ &= \delta_{priv} - \delta_{pub}\end{aligned}$$

The problem with perfect sorting is that we do not observe the test score of a high-ability student in public schools. Similarly, we do not observe the test score of a low-ability student in public schools. These hypothetical outcomes are never observed given perfect sorting. Because we cannot observe these outcomes, we cannot deduce the difference in the effect between private and public schools.

However, with randomization, we can produce an unbiased estimate of the difference between private and pubic schools. Suppose that private schooling was randomly assigned, and that all students applied to private schools. Because of randomization, the number of students attending private or public school should be just equal to the proportion of high-ability students in the population (for simplicity, we will assume that half of students are high ability). With randomization, the average achievement level for private schools will be equal to:

$$\tfrac{1}{2}(\alpha_H + \delta_{priv}) + \tfrac{1}{2}(\alpha_L + \delta_{priv}) = \tfrac{1}{2}(\alpha_H + \alpha_L) + \delta_{priv} \quad (1)$$

The average achievement level for public schools will be equal to:

$$\tfrac{1}{2}(\alpha_H + \delta_{pub}) + \tfrac{1}{2}(\alpha_L + \delta_{pub}) = \tfrac{1}{2}(\alpha_H + \alpha_L) + \delta_{pub} \quad (2)$$

To compare private-school and public-school test scores, we would subtract equation (2) from equation (1). The difference is equal to the effect of private schools relative to the effect of public schools. Hence, randomization can provide unbiased estimates of the effects of private schooling.

Similarly, randomization can provide unbiased estimates of the effects of other interventions, such as class size. For example, in a series of review articles, Hanushek has found no consistently measured effect of class size using research not based on randomization. In evaluating the random assignment of class size in Tennessee, Krueger writes, "A more general point raised by the reanalysis of Hanushek's literature summary is that not all estimates are created equal." He goes on to say that "one good study can be more informative that the rest of the literature." Krueger, quoting Galileo's description of one "Barbary steed" as faster than hundreds of packhorses, concludes that as a result of the sample size and use of randomization "Tennessee's project STAR is the single Barbary steed in the class size literature" (Krueger, 2000: 4).[6]

Under-utilization of Randomization

Besides private schooling, there are endless examples of where entry into an educational program is correlated with some unobserved characteristic. All too frequently, students who would have succeeded in the absence of a program are the same people who choose to participate in the program. Randomization can overcome the bias that such participation patterns may create. One might then ask why randomization remains under-utilized.

Eric Hanushek explains, "Although educators are dedicated to teaching students, they are reluctant to submit to the often-painful process of evalua-

6. I thank an anonymous referee for providing this quote.

tion and learning. Therefore, new ideas are seldom subjected to thorough evaluation, nor are active decisions often made about their success or failure" (1995: 241). He further explains that groups conducting educational experiments produce evaluations that "seldom involve any detailed analytical work that would permit dissemination of new techniques or new organizational forms."

There are a number of reasons that institutions and organizations may be reluctant to engage in randomized experiments in their evaluations. First, there are substantial costs associated with implementing a randomized experiment versus implementing some other intervention. These costs can be financial, political, or technical.

The financial costs can be substantial. Retrospective, nonrandom studies often rely on secondary sources, such as population surveys, to understand the impact on a set of people with certain characteristics in a given region Randomized studies cannot rely on data collected by others. Certain people participate in the randomization, and those are the people for whom data must be collected. The researcher must track specific individuals over time, which can be extremely costly (e.g., see Ludwig et al., 2001). Some studies have gone so far as to hire expensive private investigators to find study participants (Katz et al., 2000).

There are also significant political costs to implementing randomized evaluations. For example, Colombia recently instituted a educational program similar to Mexico's PROGRESA program. Families were to receive cash stipends if their children attended school and received regular health checkups. The World Bank strongly encouraged Colombia to implement the program using randomization across cities, but Colombia's central government chose not to do so. At the time, the government was negotiating a cease-fire with guerillas. The educational program provided some political leverage, and the central government implemented the program in selected cities to appease the guerillas.

There are also political costs to evaluation in terms of the long-run viability of a project. For example, Pritchett (2002) provides a model to explain why randomized evaluations are infrequent and when evaluations can be expected. In the model, those who conduct the program face the uncertainty that the program is not generating any useful results. These individuals will experience substantial costs if the program is proved ineffective. In the case of a large-scale project or one that is well funded, knowledge of its efficacy (or lack thereof) could be harmful to its organizers. As Pritchett summarizes, "No advocate would want to engage in research that potentially undermines support for his/her program. Endless, but less than compelling, controversy is preferred to knowing for sure" (Pritchett, 2002: 268).

Finally, the technical costs can also be substantial. Oftentimes, it is difficult to find someone who can accurately manage either the innovation or its implementation. For example, many program managers do not understand what researchers mean by randomization. In one voucher program in the United States, vouchers were initially randomized, but after all voucher win-

ners accepted or rejected the award, there were still awards left over. Rather than randomize again, the voucher organizers "arbitrarily" chose additional students to win. Unfortunately, their "arbitrary" choices were students who had applied to a high-quality private school. There was nothing random about this selection of students (Bettinger, 2001). Other voucher programs, including selected cities involved in Colombia's PACES program, have shown some non-randomness (Angrist et al., 2002).

In other cases, records are not kept or are kept inaccurately. For example, in the Colombian voucher program, many jurisdictions kept lists of lottery winners but not lottery losers. As a result, research could not be conducted because there was not a control group to which the treatment group could be compared. Other jurisdictions deleted records for students who won the voucher lottery but declined to use the voucher. The remaining voucher lottery winners are likely not comparable to the voucher lottery losers, as students using the voucher may differ systematically from students declining the voucher. In yet other jurisdictions, contact information was kept and updated for voucher winners but not for lottery losers. As a result, voucher winners were easier to find and interview than voucher lottery losers. The groups interviewed subsequently differed systematically. In each of these cases, the person maintaining the data did not understand the role and importance of randomization in the assessment, and consequently, expensive innovations could not be accurately evaluated. In each case, the internal validity of the evaluation (discussed below) was threatened.

Many researchers are reluctant to engage in research that identifies effects through randomization because of the difficulty of maintaining internal validity or because of the limitations of external validity. Internal validity refers to the ability of the evaluation to produce accurate estimates of the effects of the innovation while focusing on the people who actually participated in the program. If the randomization is compromised, for example, "treated" students may differ significantly from "control" students. In this situation the true effect of the program cannot be estimated. Besides the examples of administrative errors in randomization cited above, internal validity is often threatened in the data-collection process. Researchers need to be able to identify a specific group of students—the students who participated in the intervention—but sometimes these students are difficult to find or unwilling to participate in the research. If the reluctance to participate is correlated with the treatment (e.g., students who lost the lottery are angry and do not want to participate), then those students receiving the treatment for whom post-experiment/intervention data are available may differ systematically from those not receiving treatment for whom data are available. In sum, not only may the survey effort be costly, but failure to monitor response rates may lead to biased and internally invalid inferences and attrition from the sample may undermine randomization.

One way to avoid the problems that response rates may create is to use administrative data. Recent work by Angrist et al. (forthcoming) uses administrative records to demonstrate the differences between voucher lottery win-

ners and losers. By matching national identification numbers to other databases, the authors are able to follow up with all students who applied for the voucher lottery. This strategy has a lower cost than other means of getting data, although it may not be internally valid if record keeping is not equivalent across winners and losers.[7]

External validity is another worry of researchers. External validity refers to whether or not the results can be replicated and are relevant for populations besides those participating in the educational intervention. For example, the students who applied to Mexico's PROGRESA program are among the poorest in the nation. As discussed below, the PROGRESA program has improved student enrollment rates amongst these students. The external validity of this result hinges on whether or not this same result would be observed in other populations of students. If the subsidies to families were targeted at middle class or upper class families in the country, would those students also have seen the bump in enrollment rates? The results from PROGRESA shed no light on this question.

Additionally, it is not clear that the PROGRESA results will be externally valid even among students with comparable socioeconomic characteristics. Families in the PROGRESA program had to maintain records and regularly visit doctors, and highly motivated families may be more willing to do these tasks. If these families are more likely to apply to the program, then the observed results may be the effects for "highly motivated" families rather than the effects for all families. In this case, there is an unobservable characteristic (internal motivation) that prompts some families to apply to the program. Comparisons outside of that group may not be valid.

The ability of one country to replicate the programs of another country may threaten external validity. Oftentimes, there are cultural or political barriers preventing reasonable replication. Mexico's PROGRESA program has been implemented in Brazil, Colombia, and other countries. In Colombia, the program was not implemented in a way similar to PROGRESA due to the political environment, and therefore the results may differ significantly. Threats to external validity may occur for regions within a country as well. In the United States, evidence from the Tennessee STAR program on the effectiveness of smaller class sizes led other states (e.g., California) to adopt similar programs. Much like the replication of PROGRESA in other countries, it is not clear that programs in other states experienced the gains from reduced class size that the Tennessee program led to.

Concerns about external validity may be a short-run weakness of randomized evaluation. A single randomized experiment allows researchers to identify a set of conditions under which an intervention may affect student welfare. The question of whether the results from that experiment generalize to a different set of conditions is empirical. By changing the conditions and imple-

7. In one jurisdiction, as mentioned above, program managers maintained up-to-date records for voucher lottery winners but not for losers. As a result, voucher winners were much more likely to have a valid national identification number and therefore much more likely to match to post-intervention/experiment data.

menting a similar randomized experiment, researchers can identify a specific intervention's effects in other settings. The duplication of studies under varying conditions allows researchers to draw greater conclusions about the external validity of a particular intervention. In the United States, for example, the implementation of housing subsidies (i.e., Section 8 vouchers) was conducted through a sequential series of randomized experiments. The series of randomizations allowed the U.S. government to identify various conditions affecting the success of housing subsidies.

Finally, most randomized experiments measure a partial equilibrium effect of a program and not the general equilibrium effect. The partial equilibrium effect is the effect of the program on a select group of people under a specific set of circumstances. If the program was deemed successful and expanded to the general population, the set of circumstances that attended the randomized trial may change. For example, studies of voucher programs in the United States frequently focus on the effects for a small group of students attending established schools, typically parochial schools. Where voucher programs have been expanded to larger populations (e.g., Cleveland), numerous new private schools entered the market. These schools have less experience and potentially a different effect on students than the schools participating in the randomized trial. Similar expansions of private school have take place in Colombia and Chile (for Colombia, see Angrist et al., 2002; for Chile, see Hsieh and Urquiola, 2003).

Treatment Intensity and the Interpretation of Effects

One of the inherent difficulties of randomized evaluation is the accurate identification of the effects of an intervention on the people who have actually participated in it. Randomization can almost always identify the effects of an intervention on those to whom policymakers offered the intervention. This is often called the "intention to treat" parameter.[8] However, many who are offered an intervention never enroll. For example, in the Colombian voucher program, not everyone accepted the voucher. In the PROGRESA program, not every family in a participating village chose to participate. Because people who accept the offer to participate in a program may differ systematically from people who decline the offer, and because we cannot observe intent to participate among people who are randomly excluded from participation in the program, it may be difficult to measure accurately the effect on only those who participate in the intervention. Randomization does not facilitate the estimation of such an effect unless everyone offered the intervention utilizes the intervention. Some have argued that the "intention to treat" is the parameter of central interest to policymakers (e.g., Rouse, 1998), because it meas-

8. Researchers have also attempted to estimate the effect of the "treatment on the treated." This parameter measures the effect on people actually participating in the program as opposed to all people who were invited (or randomly selected) to participate. Under certain assumptions, randomization may provide a suitable instrumental variable for identifying this parameter. Rouse (1998) and Angrist et al. (2002) discuss this possibility and the assumptions necessary to calculate the effect.

ures the net effect of the program across people offered the intervention—as opposed to just those people who participate. However, others may want to know the specific effect of the treatment on the people who actually participate in the program.

Finally, the treatment may oftentimes affect the control group as well as the treatment group, making it difficult to compare the two. The most obvious example of this can be found in the results of randomized deworming projects in Kenya. Within schools, students were randomly chosen to receive oral treatments for hookworm, roundworm, and schistosomiasis (Miguel and Kremer 2001); however, because these treatments led to a decrease in contamination in the general population, all students at the school benefited from reduced rates of transmission, even those who had not received these treatments. This made it difficult for researchers to identify the effect of treatment. Similarly, if the control or treatment groups alter behavior as a result of the program, it may bias estimates of the effect of the program. For example, Hoxby (2000) criticizes the Tennessee STAR class-size experiments because teachers participated whether they had a large or small class. Teachers generally prefer small classes, and so participating teachers may have adjusted their behavior to improve the attractiveness of small classes to policymakers.

Political Economy Model of Project Implementation

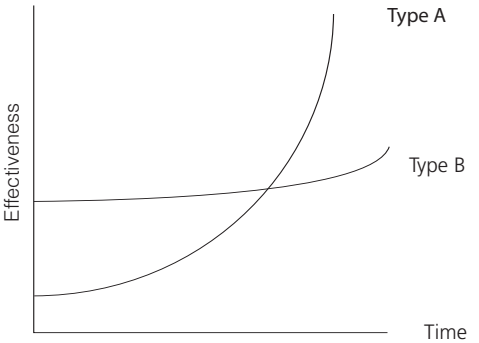
The appendix shows a simple political economy model that evaluates which types of projects policymakers are likely to implement. The key assumption in the model is that success varies according to the size of a project and the length of time it will take to complete. Some small projects yield immediate short-run benefits, but their long-run impact does not differ substantially from their short-run impact. Other projects may yield few short-run benefits, but have a long-run impact that is much higher than the projects that show short-term benefits. The second type of project is riskier in that it may have higher rewards but may also fail, leading to a large loss of investment. The intuition behind these projects is two-fold: successful programs may take time to produce, and given time, a program can identify and improve upon its weaknesses.

For simplicity, we can assume that there are only these two types of projects. Type A are those that are often large-scale and slow to develop, with small short-run effect and greater long-term effects. Type B projects are easy to implement and can succeed in the short run but may not be as effective in the long run.[9] Figure 1 plots this relationship.

Time creates an obvious tension in this model. Because individuals are impatient, they will be reluctant to embark on Type A projects. Moreover, not only do Type A projects appear to take too long, but they may not suc-

9. The designation between Type A and Type B projects is fuzzy at best. Oftentimes, we do not know the true importance of a project until we know its effects. Interventions that appear small may actually provide cost-effective solutions. I use the A/B designation to conceptualize the trade-offs associated with a large-scale intervention like PROGRESA in comparison to a number of smaller projects that may have less overall impact.

Figure 1: Effectiveness Over Time by Project Type



ceed. Because the variance of project success increases in size as a project increases in scale, risk-averse individuals are less likely to embark on large projects. In practice, impatience and risk aversion manifest themselves in both political and academic spheres. For example, politicians may be reluctant to enact important interventions because they will be up for reelection before the results of a long-term intervention will be available. Likewise, researchers—especially those with short tenure clocks—may pursue projects of lesser importance because results from research on topics of higher importance may take "too long" to obtain.

As discussed above, randomized evaluation can be more costly and more time consuming than other types of evaluation. If you factor these possibilities into the model described in the Appendix, policymakers and researchers may be even more reluctant to implement a program, particularly if randomization increases the time that one must wait for success. However, as Kremer (2004) points out, higher cost need not accompany randomized evaluation. There are numerous examples of inexpensive randomized studies of interventions that have yielded large, long-lasting effects on students. Data on the costs of interventions and evaluations are not commonplace, but when available can provide more detail on the trade-offs associated with different interventions. In addition, the opportunity costs of not undertaking any intervention or not generating new knowledge may be substantial. As Derek Bok is credited with saying, "If you think education is expensive, try ignorance."

The model also suggests that certain conditions (e.g., a longer time horizon or government stability) will improve the likelihood that an organization or government will embark on a randomized experiment. The greater the likelihood that the central leadership will remain in power for a long period of time, the more likely that it may be able to enact a policy. This may be one reason that Mexico was able to implement the PROGRESA program. Also, the more incentives that researchers have to produce more long-term projects, the more likely that they will undertake Type A projects.

CHANGES IN KNOWLEDGE BASE FROM RANDOMIZED EXPERIMENTS

This paper shows how randomized experiments have augmented the body of academic knowledge in regard to four types of educational innovations: school infrastructure development, student health, the provision of textbooks

and other learning aids, and teacher and student incentives. This paper attempts to identify the value added by randomized studies in these fields.

Ideally, one would create a complete census of randomized and nonrandomized interventions in education; however, this is impractical for two reasons. First, as the quote from Hanushek (1995) points out above, although many informal experiments are taking place across the world, most are not disseminating their findings. As a result, many experimental studies are lost in dusty cabinets or in library basements. To the extent that it is possible to track down such evaluations, I have attempted to do so. Nonetheless, the set of randomized experiments upon which I focus will be more representative of recent randomized experiments than of the entire body of work in this field.

The second problem is the sheer volume of research regarding these educational topics. Assessing the state of knowledge in the absence of randomized experiments is difficult, especially because there are large discrepancies in the quality of evaluations. For example, there are many non-experimental studies that acknowledge but do not resolve selection bias. There are yet other non-experimental studies that take advantage of "natural experiments" (e.g., changes in a country's policies) and identify the effects of the program using quasi-random variation. There are still other studies that make distributional assumptions and use observable characteristics to provide clues to the nature of the unobservable variables that may cause selection bias.[10] I rely heavily on pre-existing literature reviews to assess academic understanding prior to the recent waves of experimental evidence.

School Infrastructure Development

School construction is an area where all of the promise and problems of randomized experiments are evident. School construction is costly, and it takes a substantial amount of time to actually gauge its effects on student enrollment and attainment. Moreover, building a school in one location may exclude access to students in another, presenting difficult ethical issues. In recent years, two randomized experiments involving school building have taken place.

Before describing these studies, it may be useful to show what economists and researchers actually knew about the effects of school infrastructure prior to the experiments. In the mid-1960s, economists and policy makers were unsure as to the best way to improve access to school. Most agreed that increasing the supply of schools would improve access (Lockheed and Verspoor, 1991). However, Kindleberger summed up the knowledge as follows, "Work to improve [educational] data and the analysis goes forward to clarify the choices to be made. The need for more education is clear in underdeveloped countries at least, even though the amounts to be sought cannot be quantified" (1965: 115). In a survey of the (non-experimental) literature on

10. There are a number of studies which analyze randomized experiments as if they are not randomized. For example, Dehejia and Wahba (1998) find that propensity score matching rather than matching based on randomization can yield similar results to randomized studies.

school access, Lockheed and Verspoor argued that while there had been an observed increase in enrollment, access remained limited and groups of children were still completely excluded in very low-income countries.

Efforts to improve the supply of schools led to a number of projects. For example, in the late 1960s, Taiwan embarked on an aggressive school-building project, and in the mid-1970s, windfalls from oil revenue in Indonesia led to massive school-building projects. Taiwan and Indonesia's school-building projects were not implemented using randomization. Recent empirical work suggests that these projects could have affected enrollment and attainment. For example, Clark and Hsieh (2000) find that students received almost 0.5 years additional education as a result of Taiwan's school-building project. From 1967 to 1968, Taiwan almost doubled the number of junior high schools in the country. Using national survey data, Clark and Hsieh compare the educational attainment of young men who had already passed junior-high age to the educational attainment of men who had yet to enter junior high in 1968. The males that received the most exposure to the program received 0.42 additional years of schooling compared to the control group.

Although the result suggests that these increases in schooling were an effect of school construction, the results may be confounded by other factors. As Spohr (2003) shows, around this same time, Taiwan introduced mandatory enrollment requirements. The students who were exposed to the junior-high construction project were also exposed to mandatory enrollment requirements. Because of this confounding factor, it is unclear what the true effect of the Taiwan school-building project was on enrollment rates.

Duflo (2000) evaluated the rapid increase in school construction in Indonesia. As in Taiwan, the program was not randomized; however, Duflo shows that the allocation of schools differed by region over time, creating "quasi-randomness" that can be used to evaluate the effects of school construction. She argues that differences in education should be higher not only for younger individuals who were in school when the program began but also for children in regions with more schools built. Duflo shows that the increase in primary-school construction led to both an increase in educational attainment in Indonesia and an increase in wages. Given the regional and temporal variation of school allocation and the corresponding changes in student access, the evidence for the positive effects of school construction is compelling.

Two randomized experiments in school construction, one in Pakistan and the other in Kenya, shed further light on the effects of construction on education. The first study focuses on the creation of private girls' schools under the Urban Girls' Fellowship program, which began in 1995 in Quetta, Pakistan (the capital city of Balochistan). Kim et al. (1998) examine a pilot project that created additional schools. To appease political constituencies, the government of Balochistan guaranteed that each of the ten major urban slum areas would have one school; however, they randomized within neighborhoods as to the location of the school. Kim et al. show that girls' enrollment increased by 33 percentage points and boys' enrollments increased by 27.5 percentage points.

The authors suggest this occurred in part because boys were also allowed to attend the new schools. Many parents would not send their daughters to school if they could not also send their sons. One interesting finding was that although the success of the program differed across neighborhoods, it was not necessarily related to the relative wealth of the neighborhood or the education levels of parents. Thus, the authors conclude that this program offers promise for increased enrollments in other poor urban areas.

A similar randomized experiment took place in Kenya. Kremer et al. (2003) evaluated a program that offered funding to seven schools randomly selected from a group of fourteen schools with poor performance. The funding provided for new uniforms and textbooks, as well as for the construction of new classrooms. Not only did the dropout rates fall at the schools where funding increased, but the schools also received an influx of new students. Although class sizes grew, parents still supported the program and were willing to trade off larger classes for the additional funding.

Oftentimes, the key problem for developing countries is not the availability of school buildings but rather the availability of teachers. Student-teacher ratios can be used to illustrate the extent to which teachers are not available. There have been a number of studies that look at the effect of student-teacher ratios on student access and achievement, although few use randomization.

Fuller (1985) examines the effect of student-teacher ratios on achievement and finds little evidence of any effect. Lockheed (1987) looks at the effects of school characteristics on students, particularly the effects of student-teacher ratios on student achievement in Thailand. In her review of the literature on student-teacher ratios, she claims that high student-teacher ratios negatively affect student outcomes. Deaton and Case (1998) look at the effect of student-teacher ratios on educational attainment in South Africa. They find that student-teacher ratios have little effect on student outcomes except in the case of black students. Black students in classes with lower student-teacher ratios advance more quickly and are more likely to be in school than those in classes with higher student-teacher ratios.

There are few studies of student-teacher ratios that rely on randomization. In the most publicized study of student-teacher ratios, Krueger and Whitmore (2002) examine how student-teacher ratios affect student attainment through a randomized experiment in Tennessee. They find that small class size improves student outcomes. In developing countries, there is less evidence to support this finding. Banerjee, Jacob, and Kremer (2002) examine the effect of adding a second teacher in rural schools in Udaipur, India. This second teacher, who was female whenever possible, was randomly assigned to 21 of 42 one-teacher non-formal schools operated by an NGO. The effect of the program was significant and positive on the fraction of girls attending school; however, the authors did not find evidence that the additional teacher affected test scores.

Both random and nonrandom studies conclude that student-teacher ratios and school buildings matter. The studies on school construction provide some evidence as to the precise effect of new schools, although these

results may not be externally valid to other scenarios. Interestingly, randomized studies of teacher supply find an enrollment effect but not a corresponding effect on attainment. The results suggest that the effects of teacher supply may also differ by context.

Student Health and Education

A number of randomized experiments in developing countries have focused on improvements in student health. The rationale for many of these programs has been that improving student health may have indirect effects on student education.

Economists have long postulated that there is a direct link between health and education. Enke writes, "Health and education are often joint goods. If children have poor health, they will lose many days from school, so that better health may result in better education… Health and education are alike in that their benefits accrue partly to the individual and partly to others. When a person is cured of tuberculosis there is also a gain for the people whom he might otherwise infect" (1963: 404).

Alderman et al. (1997) investigate the effect of children's health and nutrition on school enrollments in rural Pakistan. They do not rely on randomization. Instead, they use longitudinal data and examine how price shocks to food affected health and nutrition. They identify price shocks that occurred when children were preschool age to determine their health and nutrition stock. They find health and nutrition are three times more important to school-enrollment decisions than suggested by earlier estimates that considered child health and nutrition to be predetermined rather than determined by household decisions. Especially relevant to this paper, they find that improvements in nutrition were more pronounced for girls and contributed to reduced gender differences in enrollment. They conclude that improvements in the health and nutrition of preschool children are likely to have long-run productivity effects that result in more schooling and possibly reduce the gender gaps in schooling.

A number of experimental studies have evaluated the effects of health innovations on student outcomes. Miguel and Kremer (2001), for example, study the effects of a deworming project in 75 primary schools in Kenya. These schools were phased into the program in a randomized order. Miguel and Kremer's research differs from other studies of deworming by randomizing across schools rather than randomizing across children in the same school. Studies that focus on students within the same school fail to find significant impacts of deworming (Dickson et al., 2000). Miguel and Kremer find that deworming programs at the school level led to significantly higher primary-school attendance after two years of medical treatment and that absenteeism fell by almost 25 percent. Miguel and Kremer (2001) also find that deworming creates large externality benefits by reducing the local incidence of infection within the population not participating in the program. Their study suggests that curbing tropical diseases, especially in Sub-Saharan Africa, can improve school participation.

Bobonis et al. (2003) evaluate efforts to deworm preschool-age children in the urban slums of Delhi, India. Preliminary findings suggest that preschool participation rates increased by 6.3 percentage points for participants and school absenteeism fell by one fifth. Based these initial findings, the authors advocate the program as a cost-effective way of improving enrollment for children in poor urban areas where the spread of intestinal worms is a problem.

There are a number of interesting lessons that emerge in the comparison of randomized and nonrandomized studies of student health and educational access. First, constructing an appropriate control group for a nonrandomized study is difficult. Researchers must use regional and temporal variation in treatment to construct their studies, and as before, these types of variation can mask confounding factors that may also affect health and/or educational access. In comparison, randomized experiments can provide an accurate estimate of the effect of health interventions on student outcomes. Second, within randomized experiments, the level at which the randomization occurs makes a difference. This is particularly true in health innovations. Treating students within a locale may have external effects on non-treated students (e.g., less incidence of infection). Randomized experiments can be difficult to evaluate if the treatment and control groups are both affected by the intervention. Experimental studies at the individual level within schools could not measure the effects of deworming because of a decreased incidence of infection within the school, but experimental studies across schools measured a significant effect because they studied populations more isolated from one another.

Student Vouchers and Incentives

Many policymakers have attempted recently to improve educational access by changing the incentives to students. Two policy reforms in particular have been implemented in multiple countries. The first policy reform is a large-scale educational voucher program, such as the programs implemented in Chile and in Colombia. The second policy reform is a set of student subsidies that pay students and families for school attendance, regular health check-ups, and in some cases, student achievement.

Chile established educational vouchers in 1981. The voucher program is a universal program that allows any student to transport the voucher to any participating school. Although many researchers have attempted to measure the effect of the Chilean voucher system on educational access and attainment, there is still no definitive evidence that the voucher system had positive effects on students. A number of studies find positive effects of the voucher (e.g., Carnoy and McEwan, 2001). Other studies find no significant improvement in educational attainment as a result of the voucher program (e.g., Hsieh and Uruquiola, 2003). In their evaluations of the Chilean program, these researchers face the difficulty of constructing a credible control group. As in the studies that investigate school construction, some researchers have compared students who entered the schools prior to the voucher program to

those who entered afterward. As before, if there are other systematic changes (e.g., Chile increased teacher pay dramatically after the voucher program started), then it may be unclear whether the effects are due to the voucher program or to other innovations.

Although the voucher program in Colombia was not as large and widely recognized, studies of this program have provided more definitive evidence of the effect of educational vouchers on student outcomes. The key difference between the Chilean and Columbian programs (and as a result, in the research evaluations) is the use of randomization. In the Colombian voucher program, there was a small supply of vouchers, and demand far exceeded supply. Policymakers used randomization to distribute educational vouchers fairly across applicants. Angrist et al. (2002) use this randomized distribution to identify the effects of the educational voucher. They compare students who won the voucher lottery to students who did not. Although they do not find differences in dropout rates after three years, they find that students receiving the voucher were less likely to repeat grades. Subsequent work by Angrist et al. (forthcoming) finds that students who won the voucher lottery were more likely to take and score higher on college entrance exams. Because the randomization only occurs at the level of students applying for the voucher, it is difficult to identify whether the observed effects are the result of private schooling or changes in student incentives. The voucher led to a large increase in the number of students attending private school; however, the voucher changed the students' incentives because students lost their vouchers if they did not maintain academic progress.

Although the randomization in the Colombian voucher program enabled researchers to identify effects of the voucher program, it was only possible to do so in selected cities. Angrist et al. (2002) measured the effect of the educational voucher in Bogotá and Jamundi only. There are a number of other cities for which complete voucher records exist, but in almost every case there appears to be some evidence of nonrandomness. For example, in multiple cities, students with phones and students who were older won the voucher lottery more frequently, suggesting either nonrandomness or differences in record keeping for voucher lottery winners and losers. In one city, there was additional evidence of nonrandomness. Because students could only apply for the voucher after being admitted to a private school, the lottery was random not only among students but among schools as well. In one city, 100 percent of applicants from one school won the lottery, while no other school had above a 20 percent rate of winning the lottery. Even more disconcerting, local citizens claimed that this was the most politically connected school in the city.

Another intervention that affects student incentives is the use of cash payments to reward students for attendance, regular health check-ups, and in some places, achievement. The most widely cited program is the PROGRESA program in Mexico, which was implemented in 1998. In its initial phases, the Mexican government randomly chose 320 out of 506 potental rural villages to take part in the program. Families received a cash payment if their children

both attended school and had regular health check ups. There are a number of papers that evaluate PROGRESA (e.g., Schultz, 2002; Bobonis and Finan, 2002). I focus on Behrman et al. (2001). The authors of this study use the randomization to measure the impact of PROGRESA on initial age at entering school, dropout rates, grade repetition rates, and school reentry rates. They find that the program was effective in reducing dropout rates and facilitated "progression through the grades," especially the transition from primary to secondary school. The program also induced a significant number of children who had dropped out prior to the implementation of PROGRESA to re-enter school. Unlike the health experiments, Behrman et al. do not find evidence of strong spillover effects on children who lived in the communities where PROGRESA was implemented, but did not receive subsidies. Behrman et al. (2001) project that PROGRESA might improve secondary school enrollments by as much as 19 percent.

Programs in other countries have also suggested that cash payments may influence educational decisions. In Israel, Angrist and Lavy (2002) found that providing cash-incentives for low-income students could increase *Begrut* (the Israeli high-school matriculation exam) completion, even though the value of the cash-incentive was much less than the actual returns to education. In the United States, Keane and Wolpin (2000) evaluated a 1998 proposal by Robert Reich that would offer cash bonuses to students from low-income families to graduate from high school. Keane and Wolpin found that such an incentive would reduce dropout rates for black students by two-thirds and dropout rates for white students by one-half.

The results of PROGRESA and the voucher programs provide convincing evidence that changing student incentives can alter enrollment and achievement. Still, randomization is not without drawbacks. There were a number of places where additional evidence could have been gathered from the Colombian voucher project, but administrative misunderstanding of the role of randomization and other sources of nonrandomness made it difficult to evaluate these a number of settings. This reflects the challenges of internal validity mentioned in the previous section. The Colombian voucher program ended in 1997. The evaluation of its efficacy was published in 2002. This time delay accentuates the length of time it takes to produce accurate research on the effects of educational innovations.

Innovations that Improve Quality of Education

There are a number of educational innovations that focus on better preparing students for future education or the workforce by improving the quality of instruction. Many initiatives have attempted to train teachers to teach more effectively. Other programs have focused on improving classroom instruction through audio-visual materials, particularly computers, and learning aids such as chalkboards, flip charts, and textbooks.

Recent programs have aimed to change incentives to teachers. In many cases, teachers receive substantial bonuses based on their students' performance. Advocates of such programs argue that these programs can increase the

incentive for teachers to provide effective instruction, but opponents feel these programs promote "teaching to the test."

Glewwe et al. (2003) report evidence from a randomized evaluation of a program in Kenya that provided primary-school teachers with incentives based on their students' test scores. They find that students in schools where teachers received monetary bonuses for student improvement were more likely to take exams and had higher test scores than students in other schools without the teacher incentive program. However, they do not find support for the hypothesis that these teacher incentives could reduce dropout rates or increase long-run learning. Teachers' attendance rates and the amount of homework assigned were similar across treatment and control schools. The key difference was the increase in test preparation and encouragement for students to take the test in the treatment schools. When the program ended, test score differences across treated and untreated schools disappeared.

The outcome of teacher incentives in this program was consistent with teachers using strategies that increased test scores in the short-run but did not promote long-run learning. Also, given that Kenya's centralized educational system does not provide incentives for teachers to promote long-run learning, alternative programs such as decentralizing control of teachers to the local level or allowing parents to choose schools might prove more effective in improving long-term learning.

Evaluation of the effectiveness of introducing textbooks and other learning aids is an area where randomized experiments have significantly changed our understanding of the educational process. Studies from nonrandomized programs suggest positive effects, but recent evidence based on randomized experiments in Kenya suggests that the true effect, if any, may be extremely small.

In his summary of the research on textbook implementation, Fuller (1985) reports that most studies (14 of 22) found that textbooks significantly improved student achievement. Of the 22 studies that Fuller considers, only a few used experimental research designs, and Fuller concludes that these provide the clearest evidence. Lockheed and Hanushek (1987) review evidence from 15 empirical studies concerning the cost-effectiveness of several educational inputs, and find that textbooks are among the more cost-effective inputs.[11] Lockheed and Hanushek discuss the possibility of some heterogeneity in the studies assessing the effectiveness of textbooks. They show that different studies could have reached different conclusions depending on how well matched the level, language and teacher preparation was for that book in a particular classroom. Like Fuller, they push for more experimental evaluation. They summarize, "Although the firmest conclusions about effectiveness come from experiments, very few true educational experiments have been undertaken in developing countries, particularly on a large scale. Many of what are described as 'experiments' are actually changes in national policy,

11. The most cost effective inputs were textbooks, interactive radio, peer tutoring and cooperative learning. The least cost-effective interventions included teacher training and technical-vocational schools.

which, by being implemented uniformly, lack variation. The impact of these 'experiments' on student learning, moreover, are seldom evaluated" (1987: 18).

Glewwe et al. (2004) and Kremer et al. (2000) evaluate a series of efforts to bring flip charts and textbooks to students in Kenya. Glewwe et al. examine the effectiveness of flip charts as a teaching tool. Their sample included 178 schools, of which half received flip charts from a non-government charity. They analyze the data in two comparable ways. First, they do not exploit the randomization and instead compare flip chart schools to all other schools. Using this method, they find that flip charts raised test scores by up to 20 percent of a standard deviation. This number did not change once controls for other educational inputs were added. When they exploit the randomized implementation to evaluate the program's effectiveness, however, they find no evidence that test scores increased with the use of flip charts. They conclude that many retrospective studies of nonrandomized trials would have greatly overestimated the effect of this type of program, and they stress the fact that the results would have been misleading because of omitted-variable bias.

Glewwe et al. (2000) evaluate the effects of textbook provision on student outcomes. A Dutch non-profit organization (Internationaal Christelijk Steunfonds) began a program in 1996 to help 100 primary schools in rural Kenya. In the first year, they randomly selected 25 schools to receive textbooks. In contrast to previous studies of textbook provision, the randomized evaluation of this program produced no evidence that the textbooks had a large positive effect on average test scores nor that the program affected other outcomes such as attendance and dropout rates. Using a variety of methods, the authors compare test scores across students in both treatment and control schools. For all three estimates, after one year, the impact of the textbooks across all subjects and grades is close to zero, and depending upon the estimator used, this estimate is sufficiently precise to reject the hypothesis that the textbooks increased test scores by as little as 0.07 standard deviations. They find the results of the estimates after two and three years to be similar. Because these findings differ from previous studies on textbooks, the importance of other components of textbook use—such as the teacher training for the use of the books and the particular textbook used, as well as family background—come into question. In addition, the findings further illuminate the results from the earlier paper by Glewwe et al. (2000).

CONCLUSION

The use of randomized experiments in the implementation of educational interventions continues to become more prevalent across developed and developing countries. Although randomization has the potential to provide key answers to the types of educational programs that should be implemented and the method of implementation, this paper provides some caution on the ways in which randomized experiments should be applied. Although randomization can greatly improve global knowledge about education and its provision, it does not guarantee clear conclusions.

Randomization has many potential pitfalls. It may be costly to conduct, it may require substantial and detailed oversight to ensure the integrity of the randomization, and it may require substantial time to yield meaningful conclusions. In the case of educational vouchers, hundreds of cities in Colombia embarked on ambitious educational voucher initiatives; however, most of these did not yield lessons as to the efficacy of the voucher programs. Poor record-keeping and compromises to the randomization prevented evaluation of a number of sites. Still, although many years elapsed before researchers were able to evaluate the other sites where the randomized evaluation was appropriately conducted, these programs provided conclusive evidence on the efficacy of vouchers in specific settings.

There are other trade-offs that researchers and policymakers must consider in using randomization. Because of the substantial time required to implement and evaluate some of the most important interventions, researchers and policymakers must balance their desire for quick results with their desire for comprehensive and important solutions. Researchers must also consider the costs and benefits of both the intervention and the randomized evaluation.

Has randomization improved global understanding of education and its provision? Undoubtedly. For example, although economists have long suspected relationships between health interventions and educational interventions, randomized experiments in Kenya and other places have demonstrated conclusive evidence that drug treatments can have significant effects on school attendance not only for students receiving them but also on other students (Miguel and Kremer, 2001). The randomized experiments have given us an idea of the magnitude of the effects of such interventions.

There are some educational interventions that may still be evaluated in the absence of randomization. For example, researchers exploiting quasi-randomness in school building projects provide convincing evidence of their effects on student enrollment and attainment (Duflo, 2000). These other approaches both complement randomized evaluations and provide important insights to educational knowledge and the implications of randomized trials. However, in other cases, retrospective evaluations may give misleading results. As Glewwe et al. (2000) illustrates, in some settings, failure to exploit randomization could lead to spurious findings. In the case of flip-chart provision, non-randomized evaluations suggest that flip charts had large effects on student achievement; however, evaluations that took advantage of randomization did not show an effect.

Finally, internal and external validity must also be scrutinized in randomized evaluations. Administrative data may provide comprehensive follow-up and hence improve the internal validity of estimated effects. Systematic ways to improve the external validity of a particular educational reform must be considered when structuring educational reforms. External validity may be an important weakness of random experiments, although it may be overcome through a careful determination of the location and nature of the educational intervention.

Appendix: Political Economy Model of Project Implementation

In this model, consumers implement educational projects. Consumers attempt to maximize their lifetime utility subject to the cost of the project, as equation (4) shows:

$$\max E\left[\sum_{t=1}^{\infty}\beta^{t} u(\tilde{S}(I,t))\right] \text{ subject to } \sum_{t=1}^{\infty} R^{t} C(I,t) \leq C \quad \text{(A1)}$$

where $E[\,]$ denotes the expectation operator and _ is the rate of time preference. Individual's utility is a function of the success of a program (), which is drawn from a distribution whose variance increases with the size of a project (I). Success varies by the size of a project and over time. The budget constraint is such that the present discounted value of the cost of the project must be lower than the capacity of the economy to support it.

We can assume that the people's utility increases with the success of a program but at a decreasing rate. This assumption implies that people will be risk averse. We can also make some reasonable assumptions about the nature of the success function. In particular, we can assume that success increases with the time that the program operates, and that the rate of increase in success also increases. Another way to phrase this assumption is that while it takes time to produce a successful program, over time a program can improve upon its weaknesses. We can write these assumptions as follows:

$$u'(S)>0,\ u''(S)<0,\ \frac{\delta E(S)}{dt}>0,\ \frac{\delta^{2} E(S)}{dt^{2}}>0 \quad \text{(A2)}$$

For simplicity, we can assume that there are two types of projects—Type A and Type B. Type A projects are often large-scale and slow to develop. They may not have large short-run effects, but their long-run effects may be greater than other projects. Type B projects are easy to implement and can succeed in the short run but may not be as effective in the long-run. Figure 1 in the text plots this relationship.

Time creates an obvious tension in this model. Because individuals are impatient, they will be reluctant to embark on Type A projects. Moreover, because individuals are risk averse and because the variance of project success increases over time, risk-averse individuals are less likely to embark on large projects: not only do they take too long, but there is a greater variance in the likelihood that they will succeed.

References

Alderman, Harold, Jere Behrman, Victor Lavy, and Rekha Menon. 1997. "Child Nutrition, Child Health, and School Enrollment: A Longitudinal Analysis." World Bank Policy Research Working Paper 1700.

Alderman, Harold, Peter Orazem, and Elizabeth Paterno. 2001. "School Quality, School Cost and the Public/Private School Choices of Low-Income Households in Pakistan." *Journal of Human Resources* 36: 304–326.

Angrist, Joshua, Eric Bettinger, Erik Bloom, Michael Kremer, and Elizabeth King. 2002. "The Effects of School Vouchers on Students: Evidence from Colombia." *American Economic Review* 92(5): 1535–1558.

Angrist, Joshua, Eric Bettinger, and Michael Kremer. Forthcoming. "Evidence from a Randomized Experiment: The Effect of Educational Vouchers on Long-run Educational Outcomes." *American Economic Review*.

Angrist, Joshua, and Alan Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives* 15: 69–85.

Angrist, Joshua, and Victor Lavy. 2002. "The Effect of High School Matriculation Awards: Evidence from Randomized Trials." NBER Working Paper 9389.

Banerjee, Abhijit, Suraj Jacob, and Michael Kremer, with Jenny Lanjouw, and Peter Lanjouw. 2002 " Promoting School Participation in Rural Rajasthan: Results from Some Prospective Trials." Mimeo.

Bashir, Sajitha. 1997. *The Cost Effectiveness of Public and Private Schools: Knowledge Gaps, New Research Methodologies, and an Application in India*. New York: Clarendon Press.

Bedi, Arjun, and Ashish Garg. 2000. "The Effectiveness of Private versus Public Schools: The Case of Indonesia." *Journal of Development Economics* 61: 463–494.

Behrman, Jere, Piyali Sengupta, and Petra Todd. 2001. "Progressing Through PROGRESA: An Impact Assessment of a School Subsidy Experiment."

Bettinger, Eric. 2000. "The Effect of Vouchers on Educational Achievement: Evidence from Michigan." Mimeo. Case Western Reserve.

Bobonis, Gustavo, and Frederica Finan. 2002. "Transfers to the Poor Increase the Schooling of the Non-Poor: The Case of Mexico's PROGRESA Program." Mimeo. University of California, Berkley.

Bobonis, Gustavo, Edward Miguel, and Charu Sharma. 2003. "Child Nutrition and Education: A Randomized Evaluation in India." Mimeo. University of California, Berkeley.

Carnoy, Martin, and Patrick McEwan. 2001. "Privatization Through Vouchers in Developing Countries: The Cases of Chile and Colombia." Pp.151–177 in *Privatizing Education: Can the Marketplace Deliver Choice, Efficiency, Equity, and Social Cohesion?* Boulder: Westview Press.

Clark, Diana, and Chang-Tai Hsieh. 2000. "Schooling and Labor Market Impact of the 1968 Nine-Year Education Program in Taiwan" Mimeo. Princeton University.

Deaton, Angus, and Anne Case. 1998. "School Quality and Educational Outcomes in South Africa." *Papers* 184. Princeton: Woodrow Wilson School-Development Studies.

Dehejia, Rajeev, and Sadek Wahba. 1998. "Propensity Score Matching Methods for Non-experimental Causal Studies." NBER Working Paper Number 6829.

Dickson, Rumona, Shally Awasthi, Paula Williamson, Colin Demellweek, and Paul Garner. 2000. "Effect of Treatment for Intestinal Helminth Infection on Growth and Cognitive Performance in Children: Systematic Review of Randomized Trials." *British Medical Journal* 320 (June 24): 1967–1701.

Duflo, Esther. 2000. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review* 91(4): 795-813.

Enke, Stephen. 1963. *Economics for Development*. New Jersey: Prentice-Hall.

Evans, William, and Robert Schwab. 1995. "Finishing High School and Starting College: Do Catholic Schools Make a Difference?" *Quarterly Journal of Economics* 10: 941–974.

Fuller, Bruce. 1985. "Raising School Quality in Developing Countries: What Investments Boost Learning?" World Bank Discussion Paper Series. Washington, DC: The World Bank.

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2003. "Teacher Incentives." NBER Working Paper 9671.

Glewwe, Paul, Maichael Kremer, and Sylvie Moulin. 2000. "Textbooks and Test Scores: Evidence from a Prospective Evaluation in Kenya." Mimeo. Harvard University.

Glewwe, Paul, Michael Kremer, Sylvie Moulin, and Eric Zitzewitz. 2004 "Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya." *Journal of Development Economics* 74(1): 251-268.

Hanushek, Eric A. 1995. "Interpreting Recent Research on Schooling in Developing Countries." *World Bank Research Observer* 10 (August): 227–246.

Hoxby, Caroline M. 2000. "The Effect of Class Size on Student Achievement: New Evidence from Population Variation." *Quarterly Journal of Economics* 115(4): 1239-1285.

Hsieh, Chang-Tai, and Miguel Urquiola. 2003. "When Schools Compete, How Do They Compete? An Assessment of Chile's Nationwide School Voucher Program." NBER Working Paper 10008.

Katz, Lawrence, Jeffrey Kling, and Jeffrey Liebman. 2000. "Moving to Opportunity in Boston: Early Results of Randomized Mobility Experiment." NBER Working Paper 7973.

Keane, Michael, and Kenneth Wolpin. 2000. "Eliminating Race Differences in School Attainment and Labor Market Success." *Journal of Labor Economics* 18(4): 614–653.

Kim, Jooseop, Harold Alderman, and Peter Orazem. 1998. "Can Private Schools Subsidies Increase Schooling for the Poor? The Quetta Urban Fellowship Program." Working Paper Series on the Impact Evaluation of Education Reforms, No 11. Washington, DC: The World Bank.

Kindleberger, Charles. 1965. *Economic Development*. New York: McGraw-Hill.

Kremer, Michael. 2003. "Randomized Evaluations in Developing Countries: Some Lessons." *American Economic Review* 93(2): 102–106.

———. 2006. "Expanding Educational Opportunity on a Budget: Lessons from Randomized Evaluations." In *Improving Education through Assessment, Innovation, and Evaluation*, H. Braun, A Kanjee, E. Bettinger, and M. Kremer. Cambridge, MA: American Academy of Arts and Sciences.

Kremer, Michael, Sylvie Moulin, David Myatt, and Robert Namunyu. 1997. "The Quality-Quantity Tradeoff in Education: Evidence from a Prospective Evaluation in Kenya." Working paper.

Kremer, Michael, Sylvie Moulin and Robert Namunyu. 2003. "Decentralization: A Cautionary Tale." Mimeo. Harvard University.

Krueger, Alan. 2000. "Economic Considerations and Class Size." Mimeo. Princeton University.

Krueger, Alan, and Diane Whitmore. 2002. "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 Paper W7656.

Lockheed, Marlaine. 1987. "School and Classroom Effects on Student Learning Gain: The Case of Thailand." World Bank Discussion Paper Education and Training Series, Report No. EDT 98.

Lockheed, Marlaine, and Eric Hanushek. 1987. "Improving the Efficiency of Education in Developing Countries: Review of the Evidence." World Bank Discussion Paper Series. Washington, DC: The World Bank.

Lockheed, Marlaine, Adriaan Verspoor, and associates. 1991. *Improving Primary Education in Developing Countries*. World Bank Publication. Oxford: Oxford University Press.

Ludwig, Jens, Greg Duncan, and Paul Hirschfield. 2001. "Urban Poverty and Juvenile Crime: Evidence from a Randomized Housing-Mobility Experiment." *Quarterly Journal of Economics* 116: 655–679.

Meyer, Bruce. 1995. "Natural and Quasi-Experiments in Economics." *Journal of Business and Economic Statistics* 13: 151–161.

Miguel, Edward, and Michael Kremer. 2001. "Worms: Education and Health Externalities in Kenya." NBER Working Paper no. 8481.

Neal, Derek. 1997. "The Effects of Catholic Secondary Schooling on Educational Achievement." *Journal of Labor Economics* 15: 98–115.

Newman, John, Laura Rawlings, and Paul Gertler. 1994. "Using Randomized Control Designs in Evaluating Social Sector Programs in Developing Countries." *World Bank Research Observer* 9: 181–201.

No Child Left Behind Act. 2001. http://www.ed.gov/policy/elsec/leg/esea02/107-110.pdf.

Pritchett, Lant. 2002. "It Pays to be Ignorant: A Simple Political Economy of Rigorous Program Evaluation." *Policy Reform* 5(4): 251–269.

Rouse, Cecilia. 1998. "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program." *Quarterly Journal of Economics* 113: 553–602.

Schultz, T. Paul. 2002. "School Subsidies for the Poor: Evaluating a Mexican Strategy For Reducing Poverty." Yale Economic Growth Center Paper 384.

Spohr, Christopher. 2003. "Formal Schooling and Workforce Participation in a Rapidly Developing Economy: Evidence from 'Compulsory' Junior High School in Taiwan." *Journal of Development Economics* 70: 291–327.

World Bank. 2003. "Evaluation Designs." http://www.worldbank.org/poverty/impact/methods/designs.htm.

CHAPTER 3

Expanding Educational Opportunity on a Budget: Lessons from Randomized Evaluations

MICHAEL KREMER

Although there has been tremendous progress in expanding school enrollments and increasing years of schooling in recent decades, 113 million children of primary school age are still not enrolled in school (UNDP, 2003).[1]

This paper reviews what has been learned from randomized evaluations of educational programs about how best to increase school participation. I first outline the intuition behind the important role of randomized evaluations in obtaining credible estimates of the impact of educational interventions, review other non-experimental methods of evaluation, and present some evidence on the biases that can arise with such non-experimental methods. I then discuss two types of programs that have been found to be extremely cost-effective and that could be implemented even within a very limited budget: school-based health programs and remedial education programs (that take advantage of available inexpensive sources of labor). I then outline a series of programs aimed at lowering the costs of school, or even paying students for attending school, that could be implemented if a higher level of financial support is available, and discuss the possibility of educational reform through school choice, which could be implemented given sufficient political will within a country. The paper concludes by drawing some general lessons about the contribution of randomized evaluations to understanding the cost-effectiveness of various interventions.

Given the widespread consensus on the importance of education and several existing reviews of the impact of education on income and other outcomes, this paper focuses not on the effects of education but on issues internal to the education system. The scope of this paper is also limited in that it does not address interventions intended to improve the quality of education, such as computer-aided instruction, unless these interventions cut costs and thus free resources that can be used to expand education.

1. For information on the "Education for All" initiative (which involves numerous organizations including UNESCO, UNICEF, the European Commission, and the World Bank), see UNESCO (2000, 2002).

WHY THE RESULTS OF RANDOMIZED EVALUATIONS ARE VALUABLE

There are many difficulties inherent in evaluating the impact of educational programs, as discussed below. By avoiding many of the potential biases associated with other evaluation methods, randomized evaluations are able to offer clearer estimates of program impact.

The Program Evaluation Problem

Evaluations of educational interventions—if and when they do occur—are most often conducted as afterthoughts, and not as a planned part of the program. Thus, when an estimate of the impact of an educational intervention is available, the estimate is most often based on retrospective data that are generated by everyday (non-experimental) variation across schools and households. In retrospective studies, it is very difficult to address the essentially counterfactual questions: How would the individuals who participated in the program have fared in the absence of the program? How would those individuals who did not participate in the program have fared in the presence of the program?

The difficulties inherent in examining these questions are obvious. Consider a simple illustration: a program is implemented in Uganda that seeks to improve school enrollment rates by offering free school meals, the motivation being to create additional incentives for students to attend school, as well as to possibly impact other outcomes such as nutritional status. In theory, we would like to observe a given group of students in both the state of participating in the school meals program as well as the state of not participating, and keep all other factors (rainfall, economic shocks, etc.) equal. If the group of students could be observed in both states, the evaluation would be simple; we could compare the outcomes in each scenario and know exactly what the effects of the program were because all other factors were kept constant.

Given the impossibility of observing any group of students in both states, in practice we compare data on some set of outcomes (such as school enrollment rates) for program participants to data on the same set of outcomes for some similar group of individuals who were not exposed to the program. Obtaining credible estimates hinges critically on the establishment of this second group of individuals that is "similar" to the program participants. The idea is that this "comparison" group gives us an idea of what would have happened to the program participants (the "treatment" group) had they not been exposed to the program.

In practice it can be quite difficult to construct a credible comparison group retrospectively, because individuals who did not participate in a program are most often not a good comparison group for those who did; for example, participation may be voluntary or programs may be specifically placed in poor areas. Any differences between the treatment group and the comparison group can be attributed to two separate factors: pre-existing dif-

ferences (the "bias" term) and the actual impact of the program. Because we have no reliable way to estimate the size of this bias, we typically cannot decompose the overall difference between the treatment and comparison groups into a treatment effect and a bias term.

Bias could potentially occur in either direction: for example, estimates may be biased upward if programs are placed in areas that are more politically influential, or biased downward if programs are placed in areas that have particular problems attracting children to school. Bertrand, Duflo, and Mullainathan (2004) provide evidence that even controlling for pre-existing levels of outcome variables may not be sufficient to address such concerns.

This problem of bias in program evaluations can be addressed by carefully planning the evaluation in advance in order to construct a credible comparison group. In particular, the bias disappears if the treatment and comparison groups are selected randomly from a potential population of participants (such as individuals, communities, schools, or classrooms). In randomized evaluations we can be assured that, on average, individuals who are exposed to the program are not different, by more than chance variation, from those who are not, and thus that a statistically significant difference between the groups in the outcomes affected by the program can be confidently attributed to the program.

Other Techniques to Control for Bias

By construction, randomized evaluations address the bias problem discussed above. In part because randomized evaluations are not always possible to conduct, researchers (most notably labor economists) have made significant progress in developing alternative techniques that control for bias as well as possible, such as regression-discontinuity design, difference-in-difference techniques, and propensity score matching (see Angrist and Krueger, 1999; Card, 1999; and Meyer, 1995). However, each of these non-experimental methods rests on assumptions that cannot be tested, and in practice these techniques may contain large and unknown biases as a result of specification errors. LaLonde (1986) finds that many of the econometric procedures and comparison groups used in program evaluations did not yield accurate or precise estimates; econometric estimates often differed significantly from experimental results. Although Heckman and Smith (1995) argue that there have been important developments in non-experimental evaluation methods since LaLonde's 1986 review, there is nonetheless strong evidence that in practice the results of randomized evaluations can be quite different from the estimates offered by other evaluation methods.

One strategy to control for bias is to attempt to find a control group that is as "comparable" as possible to the treatment group, at least along observable dimensions. This can be done by collecting as many covariates as possible and then adjusting the computed differences through a regression, or by "matching" the program and the comparison group through the formation of a comparison group that is as similar as possible to the program group. One such method, "propensity score matching," first predicts the probability that a given

individual is in the comparison or the treatment group on the basis of all available observable characteristics, then forms a comparison group of people who have the same probability of being treated as those who were actually treated. The weakness of this method, as with regression controls, is that it hinges on the identification of all potentially relevant differences between treatment and control groups. In cases where the treatment is assigned on the basis of a variable that is not observed by the researcher (demand for the service, for example), this technique can lead to misleading inferences.

A second strategy is often called the "difference-in-difference" technique. When a sound argument can be made that, in absence of the program, trends in educational outcomes in regions receiving the program would not have differed from trends in regions not receiving the program, it is possible to compare the change in the variables of interest between program and non-program regions. However, this assumption cannot be tested. To ascertain its plausibility, one needs time series data from before the program was implemented to compare trends over a long period. One also needs to be sure that no other program was implemented at the same time, which is often not the case. Finally, when drawing inferences, one must take into account that regions are often affected by time-persistent shocks that may look like "program effects." Bertrand, Duflo, and Mullainathan (2004) find that difference-in-difference estimations, as commonly performed, can severely bias standard errors: the researchers randomly generated placebo laws and found that with about twenty years of data, difference-in-difference estimates found an "effect" significant at the 5 percent level, for as many as 45 percent of the placebo laws.

In one example of where difference-in-difference estimates can be used, Duflo (2001) takes advantage of a rapid school-expansion program in Indonesia in the 1970s to estimate the impact of building schools on years of schooling attained and subsequent wages. Identification is possible because the allocation rule for the school is known (more schools were built in places with low initial enrollment rates), and because the cohorts participating in the program are easily identified (children twelve years or older when the program started did not participate in the program). The increased growth in years of schooling attained and wages across cohorts in regions that received more schools suggests that access to schools contributed to increased education. The trends, quite parallel before the program, shifted clearly for the first cohort exposed to the program, thus reinforcing confidence in the identification assumption. However, this identification strategy is not usually valid; often when the timing of a policy change is used to identify the effect of a particular policy, the policy change is itself endogenous to the outcomes it was meant to affect, thus making identification impossible (see Besley and Case, 2000).

Finally, a third strategy, called "regression-discontinuity design" (see Campbell, 1969) uses discontinuities that are generated by program rules in some cases to identify the effect of the program through a comparison of those who made it and those who "almost made it." That is, if resources are allocated on the basis of a certain threshold, it is possible to compare those just above the threshold to those just below.

Angrist and Lavy (1999) use this technique to evaluate the impact of class size in Israel, where a second teacher is allocated every time the class size grows above 40. This policy generates discontinuities in class size when the enrollment in a grade grows from 40 to 41 (as class size changes from one class of 40 to one class each of size 20 and 21), 80 to 81, etc. Angrist and Lavy compare test scores in classes just above and just below this threshold, and find that those just above the threshold had significantly higher test scores than those just below. This difference can confidently be attributed to the class size because it is difficult to imagine that schools on both sides of the threshold have any other systematic differences. Discontinuities in program rules, when enforced, are thus sources of identification. However, such discontinuities are not often enforced strictly enough to generate discontinuities that can be used for identification purposes, especially in developing countries. For example, researchers attempted to use as a source of identification the discontinuity in a policy of the Grameen bank (the flagship microcredit organization in Bangladesh), which restricts lending to include only people who own less than one acre of land (Pitt and Khandker, 1998). In practice, Grameen bank lends to many people who own more than one acre of land, and there is no discontinuity in the probability for borrowing at the threshold (Morduch, 1998).

Identification problems with non-randomized evaluation methods must be tackled with extreme care because they are less transparent and more subject to divergence of opinion than are problems with randomized evaluations. Moreover, the differences between good and bad non-randomized evaluations are difficult to communicate, especially to policy makers, because of the many caveats that must accompany the results. These caveats may never be provided to policy makers, and even if they are provided they may be ignored. In either case, policy makers are likely to be radically misled. Although non-randomized evaluations will continue to be necessary, there should be a commitment to conduct randomized evaluations where possible.

Evidence That the Results of Randomized Evaluations May Differ From Other Estimates

Several studies from Kenya offer evidence that estimates from prospective randomized evaluations can substantially differ from estimated effects in a retrospective framework, which suggests that omitted variable bias is a serious concern. For example, a Kenyan study (Glewwe et al., 2004) examines the potential educational impacts of providing schools with flip charts, which are poster-sized charts with instructional material that can be mounted on walls or placed on easels. This intervention covered 178 primary schools, half of which were randomly selected to receive flip charts on topics in science, mathematics, geography, and health. Despite a large sample size and two years of follow-up data from a randomized evaluation, the estimated impact of flip charts on students' test scores is very close to zero and completely statistically insignificant. This implies that the provision of flip charts had no effect on educational outcomes. In contrast, several conventional retrospective ordinary-least-squares (OLS) estimates, which presumably suffer from the

omitted variable biases as discussed, show impacts as large as 0.2 standard deviations, or 5–10 times larger than the estimates based on randomized trials.

As discussed below, similar disparities between retrospective and prospective randomized estimates have been found in studies of the impact of deworming in Kenya (Miguel and Kremer, 2003; 2004). These results are consistent with the findings of Glazerman, Levy, and Meyers (2002), who assess both prospective (experimental) and retrospective (non-experimental) methods in studies of welfare, job training, and employment-service programs in the United States, synthesizing the results of twelve design-replication studies. Their analysis finds that retrospective estimators often produce results dramatically different from the results of randomized evaluations and that the estimated bias is often large. They are unable to identify any strategy that could consistently remove bias and still answer a well-defined question.[2] I am not aware of any systematic review of similar studies in developing countries.

EFFECTIVE PROGRAMS FOR A LIMITED BUDGET

In this section, I outline two categories of programs that have been found to be extremely cost-effective means of making progress towards universal basic education. First, I discuss evidence, gathered from randomized evaluations of school-based health programs in Kenya and India, that suggests that simple and inexpensive treatments for basic health problems such as anemia and intestinal worms can have dramatic impacts on increasing the quantity of schooling that students attain. Second, I discuss the results of the randomized evaluation of a remedial education program in India that has been found to be an extremely cost-effective means of delivering education, particularly for weak students.

It is worth briefly defining the terminology of "school participation" as used in this paper. In developing countries, many pupils attend school erratically and the distinction between a frequently absent pupil and a dropout is often unclear. Attendance rates can vary dramatically among individuals, and thus large differences in the quantity of schooling would be overlooked by considering only years of schooling. One attractive way to incorporate wide variation in attendance when measuring the quantity of schooling is to focus on a more comprehensive measure of schooling, often called "participation." For any child, participation is defined as the proportion of days that he or she is present in school to the number of days that the school is open, over a given period (e.g., Miguel and Kremer, 2004). This can be applied over one or more years, or just for a few days for which reliable data are available.

2. Two recent studies not included in the analysis of Glazerman, Levy, and Meyers (2002) are those of Buddlemeyer and Skoufias (2003) and Diaz and Handa (2003). Both studies use randomized evaluation results as a benchmark to examine the performance of non-experimental methods (regression-discontinuity design and propensity score matching, respectively) for evaluating the impact of the PROGRESA program, discussed below. They find that appropriate methods provide an accurate analysis in these cases, but it is not clear that appropriate methods could have been selected *ex ante*.

Participation differs from attendance because attendance is usually defined only for children officially enrolled in school, whereas participation includes all children in the appropriate age range.

School-Based Health Programs

Poor health may limit school participation, especially in developing countries. Intestinal helminths (such as hookworm, roundworm, whipworm, and schistosomiasis) affect a quarter of the world's population, and are particularly prevalent among school-age children. Moderate-to-severe worm infections can also lead to iron deficiency anemia, protein energy malnutrition, and undernutrition. Below, I review evidence from the evaluations of school-based health programs in Kenya and India.

Available low-cost, single-dose oral therapies can reduce hookworm, roundworm, and schistosomiasis infections by 99 percent (Butterworth et al., 1991; Nokes et al., 1992; Bennett and Guyatt, 2000), and the World Health Organization (WHO) has endorsed mass school-based de-worming programs in areas with high helminth infections. Miguel and Kremer (2004) examine the impact of a twice-yearly primary school de-worming program in western Kenya, where the prevalence of intestinal worms among children is very high (with an estimated 92 percent of pupils having at least one helminth infection). The de-worming program was implemented by a Dutch non-governmental organization (NGO), Internationaal Chistelijk Steunfonds (ICS) Africa, in cooperation with a local District Ministry of Health office. Due to administrative and financial constraints, the health intervention was randomly phased in over several years.

The authors find that child health and school participation improved not only for treated students but also for untreated students at treatment schools (measurable because 22 percent of pupils in treatment schools did not receive de-worming medicine) and untreated students at nearby non-treatment schools due to reduced disease transmission. Previous studies of the impact of de-worming fail to account for potential externalities, and Miguel and Kremer use two approaches to address identification issues that arise in the presence of these disease-reduction externalities. First, randomization at the level of schools allows them to estimate the overall effect of de-worming on a school even if there are treatment externalities among pupils within treatment schools. Second, cross-school externalities—the impact of de-worming for pupils in schools located near treatment schools—are identified using exogenous variation in the local density of treatment-school pupils generated by the school-level randomization.

Using this methodology, the authors find the direct effect of the de-worming program, including within-school health externalities, led to a 7.5 percent average gain in primary school participation in treatment schools, a reduction in absenteeism of at least 25 percent. Including the cross-school externalities, the authors find that de-worming increased schooling by 0.15 years per pupil treated; decomposed into an effect of the treatment on the students treated and a spillover effect, school participation on average

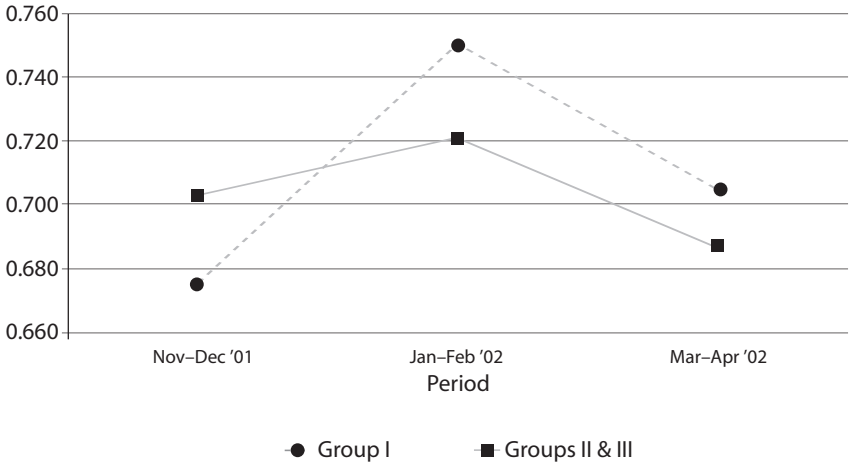
increased by 7.5 percent among pupils in treatment schools and by 2 percent among pupils in comparison schools. Including these externality benefits, the cost per additional year of school participation gained is only $3.50, making de-worming an extremely cost-effective method of increasing school participation.

Bobonis, Miguel, and Sharma (forthcoming) also find evidence that school-based health programs can have substantial impacts on school participation. Iron deficiency anemia is another of the world's most widespread health problems, affecting approximately 40 percent of children in African and Asian settings (Hall et al., 2001). Bobonis et al. evaluate the impact of an NGO project in the slums of Delhi, India that delivers iron supplementation, de-worming medicine, and vitamin A supplements to 2–6 year old preschool students (an intervention that costs only $1.70 per child per year). Before the start of the project, 69 percent of children in the sample were anemic and 30 percent suffered from worm infections. Similar to the Kenyan de-worming program, the Delhi program was phased in randomly—in this case reaching 200 preschools over a period of two years. The authors found a sharp increase of 5.8 percent in preschool participation rates, a reduction in preschool absenteeism of roughly one-fifth. Effects were most pronounced for girls and children in areas of low socioeconomic status. The study also found large weight gains (roughly 1.1 pounds on average) within the first five months of the project. In combination with the findings from the Kenyan de-worming program, these results provide compelling evidence that school-based health programs can very cost-effectively increase school participation in low-income countries.

These findings raise an important question: if school health programs can increase the quantity of schooling, how can such programs best be implemented in developing countries? Some contend that reliance on external financing of medicine is not sustainable and instead advocate health education, water and sanitation improvements, or financing the provision of medicine through local cost sharing. Kremer and Miguel (2003) analyze several de-worming interventions, including numerous "sustainable" approaches, such as cost sharing, health education, verbal commitments (a mobilization technique), and improvements in sanitation. They examine all except the sanitation efforts using randomized evaluations. Overall, their results suggest that there may be no alternative to continued subsidies for de-worming. The "sustainable" public health strategies of health education, community mobilization, and cost recovery were ineffective. For example, a small increase in the cost of the de-worming drugs led to an 80 percent reduction in take-up, relative to free treatment. On the other hand, provision of free de-worming drugs led to high drug take-up and large reductions in serious worm-infection rates. Miguel and Kremer find that the benefits of the health externalities alone are sufficient to justify not only fully subsidizing de-worming treatment, but also paying people to receive treatment.

In light of the preceding discussion of problems that arise with retrospective evaluation methods, I note that Miguel and Kremer (2004) find significant disparities between retrospective and prospective estimates of the de-

Figure 1: **Iron Supplementation and De-worming Program in India: Pre-school Participation Rates Through Time.** This table illustrates mean preschool participation rates over time for students in the treatment and comparison groups, respectively. Group I received treatments from January–April 2002 and, as this graph illustrates, experienced a sharp increase in participation rates that remained greater than comparison rates through the end of year one.



Source: Bobonis, Miguel, and Sharma (forthcoming).

worming project. For example, Miguel and Kremer estimate that students who were moderately or heavily infected in early 1999 had 2.8 percent lower school participation from May 1998 to March 1999. In contrast, an instrumental-variable specification (which imposes the condition that all gains in school participation result from changes in measured worm-infection status) suggests that each moderate-to-heavy infection leads to 20.3 percent lower school participation on average. The authors note several reasons why the instrumental-variable estimates are substantially larger, including issues with recurring infection, accounting for complementarities in school participation, and measurement error.

Remedial Education Programs

Many developing countries have substantial numbers of educated, unemployed young people who could be cheaply hired to provide supplemental or alternative instruction in schools. Pratham, an Indian NGO, implemented a remedial education program in 1994 that now reaches over 161,000 children in twenty cities. Motivated by the belief that children often drop out of school because they fall behind and feel lost in class, the program hires young women from the communities to provide remedial education in government schools. These women, the "Balsakhis," have the equivalent of a high school degree and are from the slum communities in which the schools are located. The Balsakhis teach children who have reached grade 2, 3, or 4 but have not mastered the basic grade 1 competencies. Children identified as lagging behind are pulled out of the regular classroom for two hours a day to receive this instruction.

Pratham wanted to evaluate the impact of this program, one of the NGO's flagship interventions, as they looked simultaneously to expand it. Expansion into a new city, Vadodara, provided an opportunity to conduct a randomized evaluation (Banerjee, Cole, Duflo and Linden, 2005). In the first year (1999–2000), the program expanded to forty-nine (randomly selected) of the 123 Vadodara government schools. In the following school year, the program expanded to all schools, but half received a remedial teacher for grade 3, and half received a teacher for grade 4. Grade 3 students in schools that received teachers for grade 4 served as the comparison group for grade 3 students who were directly exposed to the program. A similar intervention was conducted simultaneously in a district of Mumbai, where half the schools received the remedial teachers in grade 2, and half received teachers in grade 3. The program continued for an additional year, with each school switching the grade level to which the teacher was assigned. The program was thus conducted in several grades, in two cities, and with no school feeling that they were deprived of resources relative to others, because all schools participated in the program. After two years the program increased student test scores by 0.39 standard deviations, on average. Moreover, the gains were largest for children at the bottom of the distribution: those in the bottom third gained 0.6 standard deviations after two years. The impact of Pratham's program is increasing over time, and is very similar across cities and regardless of gender. The educational impact of the program, combined with data on the costs of teachers, suggests that hiring remedial education teachers from the community (at a cost of one or two dollars per child per year) appears to be twelve to sixteen times more cost-effective than hiring new teachers.

The positive effects of the program on children's academic achievement is remarkably stable across years and across cities, especially when the instability of the environment is considered—namely, that there was a major riot and catastrophic earthquake while the evaluation was being implemented. In their analysis, the authors carefully take into account these events, as well as their impacts on measures such as attrition, and treat that year of the program as a pilot program.

Table 1: Balsakhi Remedial Education Program in India: Estimated Cost Comparison of Balsakhis and Primary School Teachers in Mumbai. The costs of hiring Balsakhis is notably lower than the costs of hiring primary school teachers, both in terms of cost in rupees per month and in terms of rupees per student.

		Rupees per month	Rupees per student
Balsakhi	Year 1	500	54
	Year 2	750	62
Primary school teachers	Years 1 & 2	7500	1318

Source: Banerjee, Cole, Duflo, and Linden (2005).

PROMISING INVESTMENTS IF ADDITIONAL RESOURCES ARE AVAILABLE

Several sources of evidence suggest that reducing the costs of education—or taking the further step of paying students to attend school—may significantly improve participation rates. In many developing countries, school fees and required inputs such as uniforms create significant private costs of education for parents. For example, in Kenya parents have historically been required to purchase uniforms that cost about $6, a substantial expense in a country with a per capita income of around $340.

One might assume that a simple way to increase the quantity of schooling would be to reduce the cost of school or to pay students for school attendance. However, there is significant debate over the desirability of school fees. Proponents argue that fees are necessary to finance inputs, that they increase parental participation in school governance, and that the price elasticity of the demand for schooling is low (Jimenez and Lockheed, 1995). Opponents argue that school fees prevent many students from attending school and cite dramatic estimates from sub-Saharan Africa. When free schooling was introduced in Uganda in 1997, primary school enrollment reportedly doubled from 2.6 to 5.2 million children (Lokshin, Glinskaya, and Garcia, 2000); when primary school fees were eliminated in Tanzania in 2002, initial estimates were that 1.5 million students (primarily girls) reportedly began attending primary school almost immediately (Coalition for Health and Education Rights, 2002); and when Kenyan President Mwai Kibaki eliminated primary school fees in late 2002, the initial response was reportedly a massive influx of new students (Lacey, 2003). Although the elimination of school fees undoubtedly generated large increases in enrollments, the magnitude of the numbers cited in these journalistic accounts should be taken with a grain of salt for a number of reasons: the underlying data on which they are based are often unclear; free schooling is sometimes announced simultaneous to other policy initiatives; and free schooling is often accompanied by a program that replaces school fees with per-pupil grants from the central government, which creates incentives for schools to overreport enrollments.

Evidence from several recent randomized evaluations suggests that programs designed to increase participation rates through a reduction in the costs of schooling, or even payments to students to attend school, can be effective. Below, I review evidence from the Mexican PROGRESA program as well as from a series of educational interventions in Kenya, including a school meals program, a program that provided school uniforms (among other inputs), and a girls' scholarship program.

Mexico's PROGRESA Program

The PROGRESA program in Mexico[3] distributed cash grants to women, conditional on their children's school attendance and participation in preventa-

3. For more information on the PROGRESA program, see http://www.ifpri.org/themes/progresa.htm.

tive health measures (nutrition supplementation, health care visits, and health education programs). When the program was launched in 1998, officials in the Mexican government took advantage of the fact that budgetary constraints limited their ability to reach the 50,000 potential participant communities of PROGRESA immediately. They instead started with 506 communities, half of which were randomly selected to receive the program while baseline and subsequent data were collected in the remaining communities (Gertler and Boyce, 2003). Another reason for this system of implementation was that it increased the probability of the program's continuation through shifts in political power, as proponents of PROGRESA understood that it would require continuous political support to be scaled up successfully.

The task of evaluating the program was given to academic researchers through the International Food Policy Research Institute (IFPRI), who made the data accessible to numerous researchers. A number of papers have been written on PROGRESA's impact, most of which are accessible on the IFPRI web site. The evaluations show that the program was effective in improving both health and education; in a comparison of PROGRESA participants and non-participants, Gertler and Boyce (2003) find that children on average had a 23 percent reduction in the incidence of illness, a 1–4 percent increase in height, and an 18 percent reduction in anemia. Adults experienced a reduction of 19 percent in the number of days lost due to illness. Schultz (2004) finds an average 3.4 percent increase in enrollment for all students in grades 1 through 8; the increase was largest among girls who had completed grade 6, at 14.8 percent.

School Meals Programs

In some contexts, the success of conditional transfers such as those awarded through the PROGRESA program may be undermined if the people administering the program do not enforce the conditionality in practice (Sen, 2002). In these circumstances, school meals may provide a stronger incentive to attend school because children must come to school to participate.

Government-subsidized school meals have been provided in India, Bangladesh, Brazil, Swaziland, and Jamaica in order to increase both enrollment and attendance (World Food Programme, 2002). Proponents of school meals also claim that school meals can increase both the quantity of schooling and academic performance by improving child nutrition. Critics argue that families may reduce resource allocation to children who receive school meals; however, if this were the case, school meals would still serve as an incentive for families to send children to school. Moreover, a retrospective study (Jacoby, 2002) from the Philippines suggests that parents do not reduce food provided at home in response to school feeding programs (see also Long, 1991, and Powell et al., 1983).

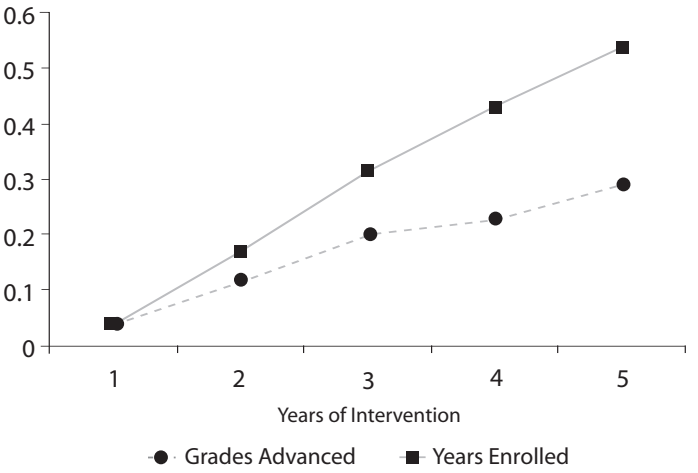
Vermeersch and Kremer conducted a randomized evaluation of the impact of school meals on participation in Kenyan preschools, and Vermeersch and Kremer (2004) find that school participation was 30 percent greater in the 25 Kenyan preschools where a free breakfast was introduced than in the 25 comparison schools. There was some evidence that the provi-

sion of meals cut into instruction time. In schools where the teacher was relatively well trained prior to the program, the meals program led to higher test scores (0.4 of a standard deviation) on academic tests. There were no effects on tests of general cognitive skills, which implies that the school meals program did not improve children's nutritional status and that the academic test-score increases were likely due to the increase in time spent in school.

Provision of School Uniforms

Kremer et al. (2002) conducted a randomized evaluation of a program in rural Kenya in which ICS Africa provided uniforms, textbooks, and classroom construction to seven schools that were randomly selected from a pool of fourteen poorly performing schools. Dropout rates fell considerably in the seven schools that were randomly selected to participate in the program, and after five years pupils in those schools had completed about 15 percent more years of schooling. In addition, many students from nearby schools transferred into program schools, raising class size by 50 percent. This outcome suggests that students and parents were willing to trade substantially larger class sizes for the benefit of free uniforms, textbooks, and improved classrooms. The authors argue that the main reason for the increase in years of schooling was most likely the financial benefit of free uniforms. A separate randomized evaluation of a program which provided textbooks in Kenya (Glewwe et al., 2003) shows that textbooks had almost no impact on the quantity of schooling, and although new classroom construction may have had an impact, the first new classrooms were not built until the second year of the program, whereas dropout rates fell dramatically in the first year. It is

Figure 2: Kenyan School Uniform, Textbook, and Classroom Construction Program: Program Effect on Grades Advanced and Years Enrolled. Given that the schools receiving the program were randomly selected, this graph illustrates the program effect by reporting the differences between the treatment and comparison groups over time.



Source: Kremer, Moulin, and Namunyu (2002).

possible in theory that anticipation of later classroom construction affected participation, but the authors note that the presence of effects for students in the upper grades, who would have finished school by the time the classrooms were built, casts doubt on this argument.

Girls' Scholarship Programs

In many countries there are significant gender disparities in access to education. It is estimated that about 56 percent of the 113 million school-age children not in school are girls, and in low-income countries there is a 9 percent gender gap in primary gross enrollment rates and a 13 percent gender gap at the secondary level (UNESCO, 2002). In sub-Saharan Africa, some studies estimate that barely 50 percent of girls complete primary school (Carceles, Fredriksen, and Watt, 2001). The question of how to increase enrollment rates of girls in primary and secondary schools in developing countries is often especially important.

There is some evidence that the elasticity of demand for schooling may be higher for girls than for boys, so policies and programs that do not specifically target girls may still result in greater increases in school participation for girls than for boys. Both Schultz (2004) and Morley and Coady (2003) find this trend in the evaluations of PROGRESA.

The alternative is to implement programs that specifically target girls. Kremer, Miguel, and Thornton (2004) conducted a randomized evaluation of the Girls' Scholarship Program, which was introduced in rural Kenya in late 2001 to enhance girls' education. From a set of 128 schools, half were ran-

Table 2: Girls' Scholarship Program: Impact on School Attendance, Busia

		Dependent variable: Attendance in 2001, 2002 (boys and girls)	
Program school		0.05**	
		(0.02)	
		Dependent variable: Attendance in 2001, 2002	
		Girls	**Boys**
Program impact	*Cohort 1 (2001)*	0.06	0.08*
		(0.04)	(0.05)
	Cohort 2 (2002)	0.01	-0.03
		(0.02)	(0.02)
Post-program impact	*Cohort 1 (2002)*	0.02	0.02
		(0.02)	(0.03)
Pre-program impact	*Cohort 2 (2001)*	0.10**	0.10*
		(0.05)	(0.06)
		Dependent variable: Teacher attendance in 2002	
Program school		0.05***	
		(0.02)	

Source: Kremer, Miguel, and Thornton (2004).

Notes: All estimates are ordinary least squares (OLS) estimates, marked as significantly different than zero at 90 percent (*), 95 percent (**), and 99 percent (***) confidence. Huber robust standard errors are in parentheses.

domly chosen to be eligible for the program. The program consisted of a merit-based scholarship—one portion, intended for school fees, paid directly to the school and a second portion, intended for school supplies and uniforms, paid to the family—that rewarded girls in two districts of Western Kenya who scored in the top 15 percent on tests administered by the Kenyan government.

In the Busia district, the scholarship reduced student absenteeism by approximately 40 percent. Across all districts participating, the program increased the average probability of school attendance by 6 percent among girls in the first cohort of the program. It had a pre-program effect of 10 percent among girls in the second cohort in the year prior to their eligibility for the scholarships, possibly due to anticipation of the future scholarship opportunities or through peer effects. In addition, the test scores of girls eligible for the scholarship increased, by 0.2 standard deviations, as a result of the program. Moreover, schools offering the scholarship had significantly higher teacher attendance after the program was introduced, and scholarship winners were 7 percent more likely to rate themselves as a "good student" than girls who did not win scholarships.

OTHER EDUCATIONAL REFORMS: SCHOOL CHOICE

Given sufficient political will within a country, another possible educational reform aimed towards increasing enrollment is that of school choice. Angrist et al. (2002) examine the effects of Colombia's voucher program on education outcomes. The program offered vouchers to attend private secondary schools to over 125,000 students from poor, urban neighborhoods. In most communities the demand for vouchers exceeded the supply, so voucher eligibility was determined by a lottery, generating a natural experiment. Data were collected from 1600 applicants for the vouchers (primarily from Bogota) three years after they had started high school. The sample was stratified so that half those sampled were lottery winners and half were lottery losers. Angrist and his co-authors find that lottery winners were 15–20 percent more likely to be in private schools, 10 percent more likely to complete grade 8, and that they scored 0.2 standard deviations higher on standardized tests than non-winners, equivalent to a full grade level.

A number of channels could account for the impact of the vouchers. First, lottery winners were more likely to have attended participating private schools, and these schools may be better than public schools. Second, vouchers allowed some pupils who would have attended private schools in the absence of vouchers to attend more expensive schools. Finally, voucher recipients who failed a grade risked losing their voucher, which increased the incentive to these students to devote more effort to school. The authors also find that vouchers affected noneducational outcomes: lottery winners worked less than lottery losers and were less likely to marry or cohabit as teenagers. Analysis of the economic returns to the additional schooling

Table 3: Colombia School Vouchers Program: Effects of the Bogota 1995 Voucher Lottery

	Coefficient on ever having used a private school scholarship (Bogota 1995 voucher lottery)		
Dependent variable	Non-lottery winner's mean	Ordinary least squares (OLS)	Two-stage least squares (2SLS)
Highest grade completed	7.5 (0.965)	0.167** (0.053)	0.196** (0.078)
In school	0.831 (.375)	0.021 (0.021)	0.010 (0.031)
Total repetitions since lottery	0.254 (0.508)	-0.077** (0.029)	-0.100** (0.042)
Finished 8th grade	0.632 (0.483)	0.114** (0.028)	0.151** (0.041)
Test scores (total points)	-0.099 (1.00)	0.379** (0.111)	0.291* (0.153)
Married or living with companion	0.016 (0.126)	-0.009 (0.006)	-0.013 (0.009)

Source: Angrist et al. (2002).

Notes: Results are from models which control for city, year of application, whether applicant had access to a phone, age, type of survey and instrument, strata of residence, and month of interview. Standard deviations are reported in parentheses for the non-lottery winner's means, and robust standard errors are reported in parentheses for the OLS and two-stage least squares (2SLS) columns. As relevant, estimates are marked as significantly different than zero at 90 percent (*), 95 percent (**), and 99 percent (***) confidence.

attained by winners after three years of participation suggests that the benefits likely greatly exceeded the $24 per winner additional cost to the government of supplying vouchers instead of public school places.

Angrist, Bettinger, and Kremer (forthcoming) suggest that the vouchers not only had significant effects on the short-run outcomes of recipients, but that their impact also persisted over time. Using administrative records of registration and test scores for a centralized college-entrance examination, the authors find that lottery winners were 7–8 percent more likely to take the university entrance exam (a good predictor of high school graduation given that 90 percent of all high school graduates take the exam), an increase of 15–20 percent in the probability of taking the exam. The authors also find an increase of 0.33 standard deviations in language test scores. Overall, these results point to a substantial gain in both high school graduation rates and achievement as a result of the voucher program. The size and persistence of these impacts suggest the voucher program was cost-effective.

One important concern about school vouchers is the effect of such programs on non-participants. On one hand, pupils left behind in public schools may be hurt by the departure of motivated classmates for private schools. On the other hand, voucher programs may enhance the education of non-partici-

pants if public schools may respond positively to increased competition. The available evidence from retrospective evaluations suggests the second effect, namely that public schools may indeed respond positively to increased competition (for evidence from retrospective studies in the United States, see Hoxby, 2000 and Bettinger, 2001). Two recent studies analyze this issue in the context of Chile's nationwide school-choice program. The first study, Hseih and Urquiola (forthcoming), finds that private enrollment rates negatively affect the relative test scores, repetition rates, and socioeconomic status of students in public schools; however, the authors' retrospective fixed-effects estimation strategy is likely problematic given that private schools entered exactly where public schools were weak. The second study, Gallego (2005), analyzes the same Chilean school-choice program using a more credible instrumental variables estimation strategy, and finds that the entry of voucher schools has positive and statistically significant effects on test scores of both public and voucher school students. Such general equilibrium effects cannot be assessed by comparing lottery winners and non-winners, but both authors note that any negative external effects on non-participants would have to be extraordinarily large to outweigh program benefits.

LESSONS

Several broad lessons can be drawn about the role of randomized evaluations in education policy, which I detail below. In addition, I briefly address some critiques of randomized evaluations that are frequently raised.

Costs

As is clear from the examples discussed in this paper, randomized evaluations are feasible and have been conducted successfully—they are labor intensive and costly, but no more so than other data-collection activities. The randomized evaluations discussed in this paper were conducted in concert with programs implemented by NGOs, and the cost-benefit estimates discussed include the costs to NGOs of program implementation.

Conducting evaluations in conjunction with NGOs has a variety of benefits. Once an evaluation staff is trained, they can work on multiple projects. Because data collection is the most costly element of these evaluations, cross-cutting the sample can also dramatically reduce costs. For example, many of the programs seeking to increase school participation and learning were implemented in the same area, by the same organization. Of course, this approach must consider potential interactions between programs, which can be estimated if the sample is large enough, and may be inappropriate if one program makes the schools atypical. Another advantage of working with NGOs is that conducting a series of studies in the same area (such as the series recently conducted in Kenya) enhances comparability by allowing researchers to compare the cost-effectiveness estimates of different interventions in the same setting.

External Validity

Without a theory to explain why a program has the effect it has, it may be unwarranted to generalize from one well-executed randomized evaluation. However, similar issues of generalizability arise no matter what evaluation technique is used. One way to determine whether a program's effects can be generalized is to encourage adapted replications of randomized evaluations in key domains of interest in several different settings. Although it will always be possible that a program unsuccessful in one context would have been successful in other adapted replications, replication of evaluations, if guided by a theory of why the program was effective, will go a long way toward alleviating concerns about generalizability.

The results of the first phase of a project often may be difficult to interpret because of circumstances that are unique to the first phase. If the project is unsuccessful, it may be because it faced implementation problems that could be avoided in later phases of the project; if the project is successful, it may be because more resources were allocated to it than would have been under a more realistic situation or in a less favorable context. Even if the choice of comparison and treatment groups ensures internal validity of estimates, the external validity of any method of evaluation may be problematic due to the specific circumstances of implementation—the results may not be able to be generalized to other contexts. Problems specific to randomized evaluations include the members of the treatment group changing their behavior (known as the Hawthorne effect) and members of the comparison group having their behavior affected (known as the John Henry effect) as a result of participation in the randomized evaluation.

Some of these concerns can be addressed by implementing adapted replications of successful (and potentially unsuccessful) programs in different contexts. Adapted replications present two advantages: first, in the process of "transplanting" a program, circumstances change, and robust programs will show their effectiveness by surviving these changes; second, obtaining several estimates in different contexts will provide some guidance about whether the impacts of the program are notably different in different groups. Replication of the initial phase of a study in a new context does not necessarily entail a delay in the full-scale implementation of a program if the latter is justified on the basis of existing knowledge. More often than not, the introduction of a program must proceed in stages, and the evaluation only requires that participants be moved into the program in random order. Even within a single study, it is possible to check whether program effects vary with covariates; for example, a program may have differential effects in small and large schools.

One example of adapted replication is the work in India of Bobonis, Miguel, and Sharma (forthcoming) who, as discussed previously, conducted an adapted replication of the de-worming study in Kenya. The baseline revealed that, although present, the levels of worm infection were substantially lower than in Kenya (in India, "only" 27 percent of children suffer from

some form of worm infection). However, 70 percent of children had moderate-to-severe anemia; thus, the program was modified to include iron supplementation. The program was administered through a network of preschools in urban India. After one year of treatment, the researchers found a nearly 50 percent reduction in moderate-to-severe anemia, large weight gains, and a 7 percent reduction in absenteeism among 4–6 year olds (but not for younger children). This supports the conclusion of the de-worming research in Kenya (Miguel and Kremer, 2004) that school health programs may be one of the most cost-effective ways of increasing school participation and, importantly, suggests that this conclusion may be relevant in low-income countries outside of Africa.

A different external validity issue is that randomized evaluation may be unable to accurately predict the cost of a program if it were implemented on a broader scale. For example, if a program initially implemented by an NGO were scaled up, the relative increase or decrease in costs might be unclear due to issues of corruption, overhead, or supervision.

Issues That Can Affect Both Randomized and Retrospective Evaluations

Sample-selection bias, attrition bias, subgroup variability, and spillover effects can affect both randomized and retrospective evaluations. In the author's opinion, it is often easier to correct for these limitations when conducting randomized evaluations than when conducting retrospective studies.

Sample-selection problems could arise if factors other than random assignment influence program allocation. Even if randomized methods have been employed and the intended allocation of the program was random, the actual allocation may not be. For example, parents may attempt to move their children from a class or school without the program to one with the program. Conversely, individuals allocated to a treatment group may not receive the treatment (for example, because they decide not to take up the program). This problem can be addressed through intention-to-treat (ITT) methods or by using random assignment as an instrument of variables for actual assignment. The problem is much harder to address in retrospective studies because it is often difficult to find factors that plausibly affect exposure to the program that would not affect educational outcomes through other channels.

A second issue affecting both randomized and retrospective evaluations is differential attrition in the treatment and the comparison groups, where participants in the program may be less likely to move or otherwise drop out of the sample than non-participants. At minimum, randomized evaluations can use statistical techniques to bound the potential bias and can attempt to track down individuals who drop out of the sample (e.g. administer tests to students who have dropped out of school), which is often not possible with retrospective evaluations.

A third issue is subgroup variability, the possibility that a program will affect some individuals more than others. The issue of subgroup variability is important, but plausible theoretical mechanisms for its presence often exist. For example, Glewwe et al. (2003) find no evidence that provision of the offi-

cial textbooks issued by the Kenyan government increased scores for the typical student. However, they do find evidence that textbooks led to higher test scores for the subset of students who scored well on a pretest. The authors note that English, the language both of instruction in Kenyan schools and of the textbooks, was the third language for most pupils. They cite evidence that many weaker pupils likely had difficulty reading the books.

Fourth, programs may create spillover effects on people who have not been treated. These spillovers may be physical, as found for the Kenyan de-worming program. De-worming interferes with disease transmission and thus makes children in treatment schools—and in schools near treatment schools—less likely to have worms, even if they were not themselves given the medicine. Spillovers may also operate through prices: Vermeersch and Kremer (2004) find that provision of meals in some schools led other schools to reduce school fees. Finally, there might also be learning and imitation effects (Duflo and Saez, 2003; Miguel and Kremer, 2003).

If spillovers are global (for example, due to changes in world prices), identification of total program impacts will be problematic with any methodology. However, if spillovers are local, randomization at the level of groups can allow estimation of the total program effect within groups and can generate sufficient variation in local treatment density to measure spillovers across groups. For example, the solution in the case of the de-worming study was to choose the school (rather than the pupils within a school) as the unit of randomization, and to look at the number of treatment and comparison schools within neighborhoods. Of course, this requires a larger sample size.

One limitation of randomized evaluations is that the evaluation itself may cause the Hawthorne effect or John Henry effect. Although these effects are specific to randomized evaluations, similar effects can occur in other settings. For example, the provision of inputs could temporarily increase morale among students and teachers, which could improve performance. Although this would create problems for randomized evaluations, it would also create problems for fixed-effect or difference-in-difference estimates.

A final issue is that the program may generate behavioral responses that would not occur if the program were generalized. For example, children may switch into a school that is provided additional inputs. This may affect the original pupils by increasing class size, if class size affects the outcome of interest. Nationwide adoption of the policy would not have this effect.

Although randomized evaluation is not a bulletproof strategy, potential biases are well known and can often be corrected. This stands in contrast to most other types of studies, where the bias due to non-random treatment assignments could be either positive or negative, and cannot be estimated.

CONCLUSIONS AND AVENUES FOR FURTHER WORK

As illustrated by the substantive examples discussed above, a number of educational interventions have been shown to expand school participation quite effectively. Randomized evaluations of school-based health programs and

remedial education programs suggest that these are extraordinarily cost-effective means of increasing the quantity of schooling attained in developing countries. Programs that reduce the cost of schooling or provide incentives for school attendance—whether implicitly, through school meals, or explicitly, through conditional grants—have been shown to have sizable impacts on school participation. Finally, school choice seems to have increased educational attainment in Colombia.

Randomized evaluations are needed on other means of increasing school participation rates, as there are a number of other promising avenues through which significant progress towards universal basic and secondary education can be made. For example, a recent study suggests that great potential likely exists on the margin of decreasing teacher absenteeism. A new representative survey of primary schools in India indicates that 25 percent of teachers in government primary schools are absent on a typical day. Two key interventions could take advantage of randomized evaluations: increasing community control in various ways (i.e., increasing the powers of parent-teacher associations) and increasing the frequency and quality of inspections, which preliminary evidence suggests can reduce teacher-absence rates.

References

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.

Angrist, Joshua, Eric Bettinger, and Michael Kremer. Forthcoming. "Long-term Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia." *American Economic Review*.

Angrist, Joshua, and Alan Krueger. 1999. "Empirical Strategies in Labor Economics." In *Handbook of Labor Economics*, Vol. 3A, eds. Orley Ashenfelter and David Card. Amsterdam: North Holland.

Angrist, Joshua, and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Children's Academic Achievement." *Quarterly Journal of Economics* 114 (2): 533–576.

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden. 2005. "Remedying Education: Evidence from Two Randomized Experiments in India." National Bureau of Economic Research Working Paper No. 11904. Cambridge, MA: NBER.

Bennett, Andrew, and Helen Guyatt. 2000. "Reducing Intestinal Nematode Infection: Efficacy of Albendazole and Mebendazole." *Parasitology Today* 16 (2): 71–75.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Difference in Difference Estimates?" *Quarterly Journal of Economics* 119 (1): 249–276.

Besley, Tim, and Anne Case. 2000 "Unnatural Experiments? Estimating the Incidence of Endogenous Policies." *Economic Journal* 110 (467): F672–F694.

Bettinger, Eric. 2001. "The Effect of Charter Schools on Charter Students and Public Schools." Mimeo. Case Western Reserve University.

Bobonis, Gustavo, Edward Miguel, and Charu Sharma. Forthcoming. "Iron Deficiency Anemia and School Participation." *Journal of Human Resources*.

Buddlemeyer, Hielke, and Emmanuel Skofias. 2003. "An Evaluation on the Performance of Regression Discontinuity Design on PROGRESA." Institute for Study of Labor, Discussion Paper No. 827.

Butterworth, A.E., et al. 1991. "Comparison of Different Chemotherapy Strategies against *Schistosoma mansoni* in Kachakos District, Kenya: Effects on Human Infection and Morbidity." *Parasitology* 103: 339–344.

Campbell, Donald. 1969. "Reforms as Experiments." *American Psychologist* 24: 407–429.

Card, David. 1999. "The Causal Effect of Education on Earnings." Pp. 1801–1863 in *Handbook of Labor Economics*, Vol. 3A, eds. Orley Ashenfelter and David Card. Amsterdam: North Holland.

Carceles, Gabriel, Birger Fredriksen, and Patrick Watt. 2001. "Can Sub-Saharan Africa Reach the International Targets for Human Development?" Africa Region Human Development Working Paper Series. Washington, DC: The World Bank.

Coalition for Health and Education Rights. 2002. "User Fees: The Right to Education and Health Denied." New York: CHER.

Diaz, Juan-Jose, and Sudhanshu Handa. 2003. "Estimating the Evaluation Bias of Matching Estimators Using Randomized-out Controls and Nonparticipants from PROGRESA." Mimeo. University of North Carolina at Chapel Hill.

Duflo, Esther. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review* 91 (4): 795–814.

Duflo, Esther, and Michael Kremer. Forthcoming. "Use of Randomization in the Evaluation of Development Effectiveness." Proceedings of the Conference on Evaluating Development Effectiveness, July 15–16, 2003. Washington, DC: World Bank Operations Evaluation Department (OED).

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

Gallego, Francisco. 2005. "Voucher-School Competition, Incentives, and Outcomes: Evidence from Chile." Mimeo. Massachusetts Institute of Technology.

Gertler, Paul, and Simone Boyce. 2003. "An Experiment in Incentive-based Welfare: The Impact of PROGRESA on Health in Mexico." Royal Economic Society Annual Conference 2003, no. 85. Royal Econometric Society.

Glazerman, Steven, Dan Levy, and David Meyers. 2002. "Nonexperimental Versus Experimental Estimates of Earnings Impacts." Mimeo. Mathematica Policy Research, Inc.

Glewwe, Paul, and Michael Kremer. Forthcoming. "Schools, Teachers, and Education Outcomes in Developing Countries." In *Handbook on the Economics of Education*, ed. E. Hanushek and F. Welch, forthcoming.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. 2003. "Textbooks and Test Scores: Evidence from a Randomized Evaluation in Kenya." Development Research Group, World Bank. Washington, DC: World Bank.

Glewwe, Paul, Michael Kremer, Sylvie Moulin, and Eric Zitzewitz. 2004. "Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya." *Journal of Development Economics* 74: 251–268.

Hall, Andrew, et al. 2001. "Anemia in Schoolchildren in Eight Countries in Africa and Asia." *Public Health Nutrition* 4 (3): 749–756.

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

Hoxby, Caroline. 2000. "Does Competition among Public Schools Benefit Students and Taxpayers?" *American Economic Review* 90 (5): 1209–1238.

Hseih, Chang-Tai, and Miguel Urquiola. Forthcoming. "The Effects of Generalized School Choice on Achievement and Stratification: Evidence from Chile's School Voucher Program." *Journal of Public Economics.*

Jacoby, Hanan. 2002. "Is There an Intrahousehold Flypaper Effect? Evidence from a School Feeding Program" *Economic Journal* 112 (476): 196–221.

Jimenez, Emmanuel, and Marianne Lockheed. 1995. "Public and Private Secondary Education in Developing Countries." World Bank Discussion Paper no. 309. Washington, DC: World Bank.

Kremer, Michael. 2003. "Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons." *American Economic Review Papers and Proceedings* 93 (2): 102–115.

Kremer, Michael, and Edward Miguel. 2003. "The Illusion of Sustainability." Mimeo. Harvard University.

Kremer, Michael, Edward Miguel, and Rebecca Thornton. 2004. "Incentives to Learn." Mimeo. University of California, Berkeley.

Kremer, Michael, Sylvie Moulin, and Robert Namunyu. 2002. "Decentralization: A Cautionary Tale." Mimeo. Harvard University.

Lacey, Marc. 2003. "Primary Schools in Kenya, Fees Abolished, Are Filled to Overflowing." *The New York Times*, January 7: A8.

LaLonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training with Experimental Data." *American Economic Review* 76 (4): 604–620.

Lokshin, Micahel, Elena Glinskaya, and Marito Garcia. 2000. "The Effect of Early Childhood Development Programs on Women's Labor Force Participation and Older Children's Schooling in Kenya." Policy Research Report on Gender and Development, Working Paper Series no. 15. Washington, DC: World Bank.

Long, Sharon K. 1991. "Do the School Nutrition Programs Supplement Household Food Expenditures?" *The Journal of Human Resources* 26: 654–678.

Meyer, Bruce D. 1995 "Natural and Quasi-experiments in Economics." *Journal of Business and Economic Statistics* 13 (2): 151–161.

Miguel, Edward, and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72 (1): 159–217.

——. 2003. "Networks, Social Learning, and Technology Adoption: The Case of Deworming Drugs in Kenya." Mimeo. Harvard University.

Morduch, Jonathan. 1998. "Does Microfinance Really Help the Poor? New Evidence from Flagship Programs in Bangladesh." Mimeo. Princeton University.

Morley, Samuel, and David Coady. 2003. *From Social Assistance to Social Development: Education Subsidies in Developing Countries*. Washington, DC: Institute for International Economics.

Nokes, C., S. M. Grantham-McGregor, A. W. Sawyer, E. S. Cooper, B. A. Robinson, and D. A. P. Bundy. 1992. "Moderate-to-heavy Infection of Trichuris richiura Affects Cognitive Function in Jamaican School Children." *Parasitology* 104: 539–547.

Pitt, Mark, and Shahidur Khandker. 1998. "The Impact of Group-based Credit Programs on Poor Households in Bangladesh: Does the Gender of Participants Matter?" *Journal of Political Economy* 106 (5): 958–996.

Powell, Christine, Sally Grantham-McGregor, and M. Elston. 1983. "An Evaluation of Giving the Jamaican Government School Meal to a Class of Children." *Human Nutrition: Clinical Nutrition* 37C: 381–388.

Schultz, T. Paul. 2004. "School Subsidies for the Poor: Evaluating the Mexican PROGRESA Poverty Program." *Journal of Development Economics* 74: 199–250.

Sen, Amartya. 2002. "The Pratichi Report." Pratichi India Trust.

United Nations Development Programme (UNDP). 2003. *Human Development Report*. New York: UNDP.

United Nations Educational, Scientific, and Cultural Oraganization (UNESCO). 2000. *Informe Final, Foro Mundial Sobre la Educación*, Dakar, Senegal. Paris: UNESCO Publishing.

———. 2002. *Education for All: Is the World On Track?* Paris: UNESCO Publishing.

Vermeersch, Christel, and Michael Kremer. 2004. "School Meals, Educational Achievement, and School Competition: Evidence from a Randomized Evaluation." World Bank Policy Research Working Paper No. 3523. Washington, DC: The World Bank.

World Food Programme. 2002. "Global School Feeding Report 2002." Rome: World Food Programme School Feeding Support Unit.

World Health Organization (WHO). 1992. *Model Describing Information: Drugs Used in Parasitic Diseases*. Geneva: WHO.

Contributors

Eric Bettinger is an assistant professor in the department of economics at Case Western Reserve University. He is also a faculty research fellow at the National Bureau of Economic Research. From 2002–2003, Bettinger was a Visiting Scholar at the American Academy of Arts and Sciences. His work focuses on determinants of student success in primary and secondary school. He has written several papers on the effects of educational vouchers on student outcomes in Colombia. He has also written on the academic and non-academic effects of educational vouchers in the United States. His most recent work focuses on the determinants of college dropouts and the effectiveness of remediation in reducing dropout behavior.

Henry Braun is a distinguished presidential appointee at the Educational Testing Service (ETS) and served as vice-president for research management at ETS from 1990–1999. He has published in the areas of mathematical statistics and stochastic modeling, the analysis of large-scale assessment data, test design, expert systems, and assessment technology. His current interests include the interplay of testing and education policy. He has investigated such issues as the structure of the Black-White achievement gap, the relationship between state education policies and state education outputs, and the effectiveness of charter schools. He is a co-winner of the Palmer O. Johnson award from the American Educational Research Association (1986), and a co-winner of the National Council for Measurement in Education award for Outstanding Technical Contributions to the Field of Educational Measurement (1999).

Anil Kanjee is an executive director at the Human Sciences Research Council (HSRC), South Africa. He is head of the HSRC Education Quality Improvement Initiative, which aims to support government and other key role-players in the implementation of evidence-based policies and practices to improve education quality. His research interests include education change and school reform in developing countries, the use of assessment to improve learning, the application of Item Response Theory for test development and score reporting, and the impact of globalization on knowledge creation and utilization. He also works on an initiative to establish and strengthen links among researchers in Africa and other developing nations for the purpose of sharing expertise and experience in improving education quality.

Michael Kremer is Gates Professor of Developing Societies at Harvard University, senior fellow at the Brookings Institution, and a non-resident fellow at the Center for Global Development. He founded and was the first executive director (1986–1989) of WorldTeach, a non-profit organization that places two hundred volunteer teachers annually in developing countries. He previously served as a teacher in Kenya. A Fellow of the American Academy of Arts and Sciences, Kremer received the MacArthur Fellowship in 1997. His research interests include AIDS and infectious diseases in developing countries, economics of developing countries, education and development, and mechanisms for encouraging research and development.

The Project on Universal Basic and Secondary Education

Directed by Joel E. Cohen (Rockefeller and Columbia Universities) and David E. Bloom (Harvard University), the Academy's project on Universal Basic and Secondary Education (UBASE) is sponsoring a series of multidisciplinary studies of the rationale, means, and consequences of providing an education of high quality to all children in the world. Working groups are investigating a number of topics including: basic facts and data on educational expansion, history of educational development, consequences of attaining universal education, means of educational expansion, goals and assessment of universal education, politics and obstacles to educational reform, costs of universal education, and health and education. The UBASE project is supported by grants from the William and Flora Hewlett Foundation, John Reed, the Golden Family Foundation, Paul Zuckerman, an anonymous donor, and the American Academy of Arts and Sciences.

Advisory Committee

Leslie Berlowitz
American Academy of Arts and Sciences

Nancy Birdsall
Center for Global Development

Joan Dassin
Ford Foundation

Howard Gardner
Harvard University

George Ingram
Academy for Educational Development

Kishore Mahbubani
National University of Singapore

Katherine Namuddu
Rockefeller Foundation

Kenneth Prewitt
Columbia University

John Reed
New York, NY

Jeffrey Sachs
Earth Institute, Columbia University

Gene Sperling
Council on Foreign Relations

Paul Zuckerman
Zuckerman & Associates, LLC

For more information, please contact:
Project on Universal Basic and Secondary Education
American Academy of Arts and Sciences
136 Irving Street, Cambridge, MA 02138
Phone: 617-576-5024
email: ubase@amacad.org
http://www.amacad.org/projects/ubase.aspx

The American Academy of Arts and Sciences

Founded in 1780, the American Academy of Arts and Sciences is an international learned society composed of the world's leading scientists, scholars, artists, business people, and public leaders. With a current membership of 4,000 American Fellows and 600 Foreign Honorary Members, the Academy has four major goals:

- Promoting service and study through analysis of critical social and intellectual issues and the development of practical policy alternatives;
- Fostering public engagement and the exchange of ideas with meetings, conferences, and symposia bringing diverse perspectives to the examination of issues of common concern;
- Mentoring a new generation of scholars and thinkers through the newly established Visiting Scholars Program;
- Honoring excellence by electing to membership men and women in a broad range of disciplines and professions.