2013

Improving Policies and Programs for Educational Quality: An Example from the Use of Learning Assessments

Daniel A. Wagner
University of Pennsylvania, wagner@literacy.upenn.edu

Recommended Citation (OVERRIDE)

Wagner, D.A. (2013). Improving Policies and Programs for Educational Quality: An Example from the Use of Learning Assessments. In Britto, P., Engle, P., & Super, C. (Eds.), *Handbook of Early Childhood Development and Its Impact on Global Policy*, 389-406. Oxford University Press. DOI:10.1093/acprof:oso/9780199922994.001.0001

Improving Policies and Programs for Educational Quality: An Example from the Use of Learning Assessments

Abstract

It is early morning in Kahalé village, about 45 kilometers from the capital city. It has been raining again, and the water has been flowing off the tin corrugated roof of the one-room schoolhouse at the center of the village. The rain makes it difficult for Monsieur Mamadou, a teacher, to get to his school on this Monday morning, as the rural taxi keeps getting stuck in the mud, forcing the six other passengers to help the driver get back on the road to the village. Once at school, Monsieur Mamadou waits for his school children to arrive. At 9 a.m., the room is only half-full, probably not a bad thing, as a full classroom would mean 65 children, and there are only benches enough to seat 50.

Disciplines

Curriculum and Instruction | Early Childhood Education | Education | Educational Assessment, Evaluation, and Research | Educational Methods | International and Comparative Education

{ 20 }

Improving Policies and Programs for Educational Quality

AN EXAMPLE FROM THE USE OF LEARNING ASSESSMENTS

Daniel A. Wagner

Introduction

It is early morning in Kahalé village, about 45 kilometers from the capital city. It has been raining again, and the water has been flowing off the tin corrugated roof of the one-room schoolhouse at the center of the village. The rain makes it difficult for Monsieur Mamadou, a teacher, to get to his school on this Monday morning, as the rural taxi keeps getting stuck in the mud, forcing the six other passengers to help the driver get back on the road to the village. Once at school, Monsieur Mamadou waits for his school children to arrive. At 9 a.m., the room is only half-full, probably not a bad thing, as a full classroom would mean 65 children, and there are only benches enough to seat 50.

Now about 35 students have arrived. Those with proper sandals and clean shirts that button are in the first row or two; those with no sandals and not-so-clean shirts sit further back. The children, all in first grade, range in age from 6 to 10 years. Monsieur Mamadou speaks first in Wolof, welcoming the children, telling them to quiet down and pay attention. He then begins to write a text on the blackboard in French, taking his time to get everything just so. The accuracy of the written text is important since only a few children (all in the front row) have school primers in front of them. Mamadou's writing takes about 15 minutes, during which time the children are chatting, looking out the window, or have their heads bent down with eyes closed on their desks. Some are already tired and hungry as they have had nothing but a glass of hot tea and stale bread or mash in the morning. When Monsieur Mamadou finishes his writing, he turns around to address the class in French: "You are now to copy this text into your *carnets* (notebooks)." The children

begin to work and Monsieur Mamadou steps outside to smoke a cigarette. It is April, and the rains have come, but he is tired—it has been a long year.

Aminata, 8-years-old, sits in row three. She has her pencil out, and begins to work in her *carnet*, carefully writing down each word written on the blackboard. She is thankful to make it to school that day, since her little baby sister was going to need Aminata to be a caretaker at home—except that her Auntie was visiting, so Aminata could go to school after all. Although going to school is better than staying home, Aminata has a sense that she is not making very good use of her time. She can copy the text, but doesn't understand what it says. Aminata can only read a few French words on the street signs and wall ads in her village. Thus, even as the only "school" child in her family, she is not much help to her mother, who wants to know what the writing on her prescription bottle of pills really says. Aminata feels bad about this, and wonders how it is that her classmates in the first row seem to already know some French. She also wonders why Monsieur Mamadou seems only to call on those pupils to come to the front of the class and work on the blackboard, and not her. She's heard that there is a school after primary school, but only the first-row kids seem to get to enroll there. What is the point of studying and staying in school, she wonders?

In the above story, there is nothing remarkable about Monsieur Mamadou or Aminata. The vignette tells an all too familiar tale that is repeated in countries around the world. Although dysfunctional classroom contexts exist in all nations, their consequences are exacerbated when resources for learning are so limited, as in the poorest countries in Africa. This vignette is about poverty, the cultural context of failing educational systems, and the communities that fail to notice what is wrong in their midst.

For many of us, it tells a story about non-learning, non-reading, and incipient school failure (for both children and the school). Most young children similar to Aminata will not be adequately assessed for learning before they drop out of school. Many children similar to Aminata do not even exist, when considered from a national statistical perspective. They will not make it to secondary school, will not go to university, and will not get a job in the global economy. This year or next will likely be Aminata's *last* in school. She will likely marry around puberty and begin a similar cycle of non-education for her own children. This is not true of all children, but it is true of most children in poor parts of poor countries. This sad and familiar story needs to be addressed and changed. Aminata's story is at the heart of the education problem in the low-income countries (LICs).

The above vignette is also a contextual "data point" (or really data *points*) that represent how to *collect* data, *interpret* the data, and *inform* policy making. It is also a story about how culture intersects with theories based on data collected in relatively affluent (Western) contexts quite different from those that may be the focus of improving the lives of children in poor contexts. The conceptual and practical responses to the above problems have been with us for many decades and

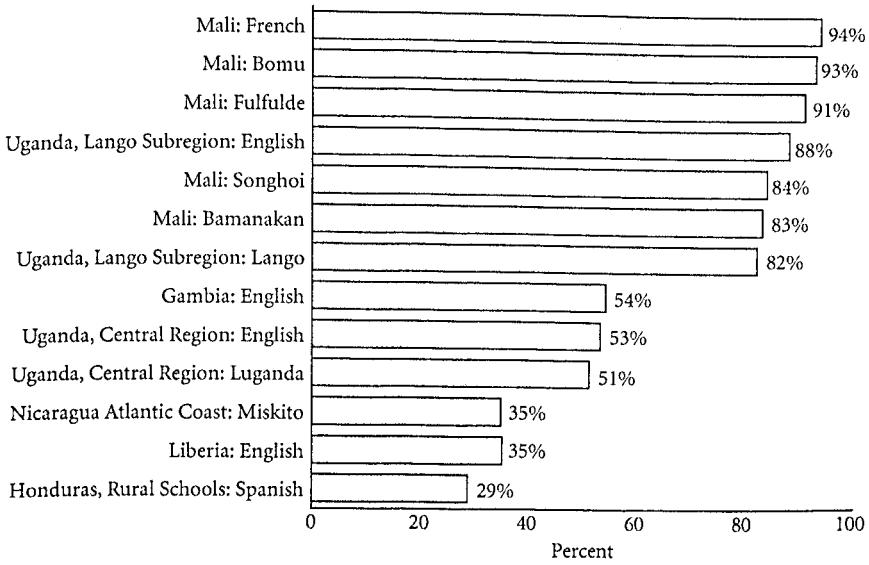

FIGURE 20.1 *Percentage of students who could not read a single word, 2008–2009.* Reprinted with permission from Gove & Cvelich, 2010, p. 10. Children were at end of grade 2 (about 7–9 years of age) from nationally representative samples; "language" indicates which language was used in the reading assessment.

are not easy to resolve. Irrespective of various scientific perspectives (e.g., anthropological, neuroscientific, maturational) around what to do from a perspective of improving child development, we do know this: lots of children are like Aminata, and cannot read a single word after a year or more in school (Figure 20.1). As a first step in thinking about Aminata's "problem" (that is, to Aminata herself, her family, her community, and national and international policymakers), we need first to consider what data count.

Measurement and Change in Two Developments

> We regard culture as a patterned configuration of routine, value-laden ways of doing things that make some sense as they occur together in the somewhat ordered flux of a community's ways of living. Cultural processes do not function in isolation or in mechanical interaction among independently definable entities. Research efforts that try to control for all but a few aspects of community functioning—to be able to separately examine the effects of stand-alone variables—overlook the meaning that is given to each aspect by their integration. (Rogoff & Angelillo, 2002, p. 216)

Generating knowledge about whether a program achieved its basic aims requires impact evaluation, which analyzes and documents the extent to which changes in the well-being of the target population can be attributed to

> a particular program or policy. Such evaluation tries to answer the question: "What difference did this program make?" Impact evaluation asks about the difference between what happened with the program and what would have happened without it... (Savedoff, Levine, & Birdsall, 2006, p. 12)

The two quotations above illustrate a dilemma, or rather two different ways of thinking about the meaning of data and how change can be measured. In the first, Rogoff and Angelillo (2002) assert that culture is not a single indicator, but rather is made up of a complex web of beliefs and behaviors that cannot be extracted and studied independently. By contrast, drawn from the field of economics, Savedoff and colleagues (2006) make the point that it is absolutely essential to determine which specific inputs (fiscal, cultural, policy, etc.) determine specific outcomes (i.e., impact evaluation) in trying to affect behavioral change. In the case of Aminata, the anthropologist would be focused on the various cultural, social, and cognitive (e.g., cannot read a single word) factors that might keep her in school now, or might influence her to leave school prematurely. An educational psychologist or economist might tinker, adjust, or change the type of textbook used in class or the incentives that might be put into play (e.g., conditional cash transfer; Behrman, Parker, & Todd, 2009) to persuade parents that schooling has value.

In the context of childhood development, the term "development" connotes what develops over *chronological time*—over the experiences that young people go through in their homes, schools, and societies (what is called *developmental science* in psychology). However, when employed by United Nations (UN) and donor agencies, the second meaning of "development" is generally thought of as representing international *economic* development. These two different connotations of development underscore another dilemma—namely, that change over individual chronological (lifespan) time, and that of societal change overlap—both take chronology seriously. But child development and international development have almost completely different conceptual and empirical bases. (For early explorations of the intersection of human development and international development, see Wagner [1983, 1986].)

One way to pursue this inquiry in the education field is to consider one of the core goals of all development agencies and ministries of education—namely, the production of children who can read.

Reading: A Globally Desirable Outcome and Developmental Challenge

The goal of reading—and a literate world—is at the top of UN Millennium Development Goals (MDGs) for education and economic development. Indeed, it is widely accepted that, despite its importance, literacy rates have not changed very much over several decades (Table 20.1), especially in LICs. If one were to engage in a substantive internet-based search of publications in the field of

TABLE 20.1 Estimates of adult illiterates and literacy rates (population aged 15+) by region, 1990 and 2000–2004

	Number of Illiterates (thousands)		Literacy Rates (%)		Change from 1990 to 2000–2004		
					Number of Illiterates		Literacy Rates
	1990	2000–2004	1990	2000–2004	thousands	(%)	(Percentage points)
World	871,750	771,129	75.4	81.9	−100,621	−12	6.4
Developing countries	855,127	759,199	67.0	76.4	−95,928	−11	9.4
Developed countries	14,864	10,498	98.0	98.7	−4,365	−29	0.7
Countries in transition	1,759	1,431	99.2	99.4	−328	−19	0.2
Sub-Saharan Africa	128,980	140,544	49.9	59.7	11,564	9	9.8
Arab States	63,023	65,128	50.0	62.7	2,105	3	12.6
Central Asia	572	404	98.7	99.2	−168	−29	0.5
East Asia and the Pacific	232,255	129,922	81.8	91.4	−102,333	−44	9.6
South and West Asia	382,353	381,116	47.5	58.6	−1,237	−0.3	11.2
Latin American and the Caribbean	41,742	37,901	85.0	89.7	−3,841	−9	4.7
Central and Eastern Europe	11,500	8,374	96.2	97.4	−3,126	27	1.2
North America and Eastern Europe	11,326	7,740	97.9	98.7	−3,585	−32	0.8

Note: Figures may not add to totals because of rounding.

Adapted from UNESCO, 2005, p. 63

Source: Statistical annex, Table 2A.

reading today, the outcome would surely show millions of articles, books, and chapters. Yet, the vast majority of these would be in only a handful of languages, largely contained within a dozen major languages of the world. This statistic would leave the remaining 2,000–3,000 languages most commonly used in the world with near-zero research as to how reading is acquired or utilized. As with other developmental phenomena, such as language, motor skills, and personality, one might (indeed should) ask the question of how much of a global sample of humanity is necessary before we can reach generalizable conclusions about a particular domain of behavior. Given a long-standing tendency in scientific psychology to look for universals, education and child development specialists have mainly not been very concerned (with some notable exceptions) about whether their conclusions might apply to peoples in far-away places, or even to ethnic groups much closer to home.

As one example of this universalistic trend, the study of reading acquisition remains heavily biased in favor of research undertaken in the industrialized world.[1] Further, much of this research is actually on the acquisition of cognitive skills, such as perception and memory, and reading subskills, such as decoding and comprehension (Kamil, Mosenthal, Pearson, & Barr, 2000). Most of this work has been carried out with school-aged children who are learning to read in English or in a handful of other languages, with relatively little research on reading acquisition undertaken in the large variety of the world's languages and scripts. The role of culture in theories of reading, although considered important, has often been marginalized in Western-dominated approaches to reading (however, see Street, 2001; Wagner, Venezky, & Street, 1999; Wagner, 2004).

In a more culture-sensitive model, a number of key learner characteristics need to be taken into account, most particularly what a child has learned at home before arriving at school and the cultural context of learning outside of the schooling process. The school provides, in addition, a set of inputs that includes time, teaching methods, teacher feedback, learning materials, and so forth. In the child, then, a set of outcomes would include cognitive skills learned (such as reading and writing), as well as social attitudes and values. This model points to the importance of measuring a variety of outcomes, but leaves out which outcomes depend on which intermediate contextual variables and how one might measure them. By understanding the factors that promote learning, a path toward improvement begins to come into focus. As one example, multiple studies have confirmed the role of a mother's education in the academic success of her children. Many claim that maternal education is one of the most powerful determinants of children's staying in school and learning achievement (Figure 20.2; UNESCO, 2005). Yet, how does a mother actually transmit skills, attitudes, and values to her children, especially if she herself is poorly educated? Recent evidence suggests that a key causal variable is how the mother communicates with her children (LeVine, LeVine, Schnell-Anzola, Rowe, & Dexter, 2012).

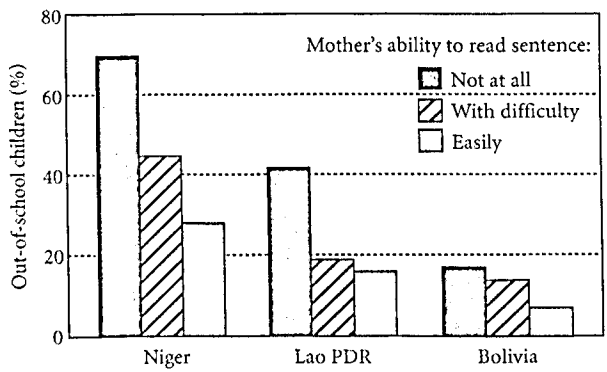

FIGURE 20.2 *Mother's literacy and schooling status in the Niger, the Lao PDR, and Bolivia, 2000.* Adapted from UNESCO, 2005, p. 130.

In addition, we know that the presence of qualified teachers, well-prepared curricula and textbooks, supportive parents, and engaged communities are all factors that can and do affect children's learning. What is less than clear is how to determine what inputs and outputs need to be studied in which cultural contexts, and then to decide what implementation steps are needed to reinforce and expand policies that support them. Improved measurement tools necessarily will play an important part in this process.

Measurement of Learning Outcomes

Educational measurement intersects with the world of population variation in ways that are predictable, but also can be difficult to address. This is not only a matter of international or cross-cultural comparability. Rather, variation in populations is endemic in each and every context where children are raised. Even variation in what households contain, such as the availability of books, has been found to be highly related to reading outcomes in Sub-Saharan Africa (see Figure 20.3). Each household may also contain significant variation in the learning environments of children.

Reading test scores serve as a proxy for general educational quality. Therefore, the use of such indicators can provide solid information on how well the content in a school curriculum is being understood as *learned* cognitive skills, a formative measure of teaching and learning policies, and a benchmark for how well children

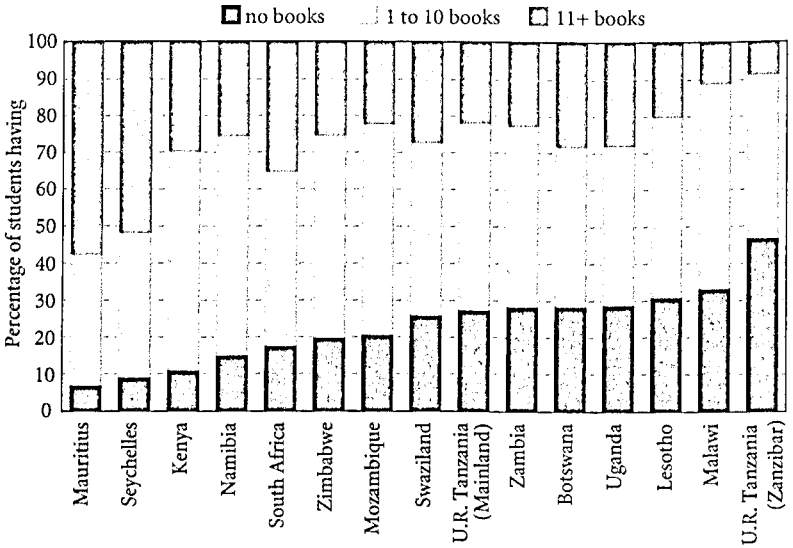

FIGURE 20.3 *Grade 6 student reports of quantity of books in their homes in 15 SACMEQ African education systems, 2000.*

Adapted from UNESCO, 2004. p. 208.

have done at the main exit points from the school system. This latter type of summative assessment may be used as a means of determining (and legitimizing) access to educational (and hence economic) advancement. The policy upside is that tests may help to ensure that the intended curriculum is taught and learned, whereas the downside is that they can provide ways for nonachieving students to be forced out the schooling system altogether if their learning is poor. Thus, one key goal of understanding how children learn to read is to determine better ways to intervene in the learning process and to remediate skills before dropout occurs.

Educational assessments come in a wide variety of styles, contents, and purposes. At the dawn of the 20th century, Alfred Binet (also known as one of the fathers of intelligence testing) was requested by the French government to develop an assessment instrument that could help predict which students would be most likely to succeed in public schools. This element of prediction—of success, or not, in schooling—was a watershed moment in the use of testing for policy making. Over the next century, educators and policymakers across the world have endeavored to make similar decisions based on examinations. As a consequence, even countries with relatively low incomes and poorly financed educational systems have begun to actively participate in such learning assessments. For the present purposes, with a focus on reading in LICs, we will consider only three principal types of assessments: international, regional, and hybrid.

International Assessments

International assessments (sometimes termed large-scale educational assessments or LSEAs) are designed to measure learning in multiple countries. Their aims include (a) cross-national comparisons that target a variety of educational policy issues, (b) provision of "league tables" that rank-order achievement scores by nation or region or other variables, and (c) within-country analyses that are then compared to how other countries operate at a subnational level. Such assessments gather data principally from learners, teachers, and educational systems—parameters that help to provide better ways of interpreting test results. These studies, many of which include reading tests, are planned and implemented by various international organizations and agencies, including the International Association for the Evaluation of Educational Achievement (IEA) that conducts the Progress in International Reading Literacy Study (PIRLS), and the Organization for Economic Cooperation and Development (OECD), which is responsible for the Program for International Student Achievement (PISA) studies. These assessments may also be characterized by their attention to high-quality instruments, rigorous fieldwork methodology, and sophisticated analyses of results. Each of these international reading assessments is now in use in dozens of countries and is expanding to LICs, well beyond the OECD country user base that formed the early core group of participation. International assessments often attract media attention and thus provide an opportunity for greater focus and debate on the education sector and national outcomes relative to other countries.

Regional Assessments

As part of an effort to extend the use of LSEAs into LICs, regional and international organizations have collaborated to create three major regional assessments: the Latin American Laboratory for Assessment of Quality in Education (LLECE), the Southern and Eastern African Consortium for the Monitoring of Education Quality (SACMEQ), and Program for the Analysis of Educational Systems of the CONFEMEN (Francophone Africa) countries (PASEC). These regional assessments have much in common with the international assessments, but there are several important differences, including the relatively greater proximity in content between test and curriculum, normative scales that may or may not be tied to local (normed) skill levels, and attention to local policy concerns (such as the role of the French language in PASEC countries). The overlap in expertise between the specialists working on the international and regional levels has generally meant that these regional tests are given substantial credibility, and they are largely used by specialists within national ministries of education.

Hybrid Assessments

In recent years, a new approach to assessment has sought to focus more directly on the needs of poor LIC assessment contexts. Initially, this approach was conceptualized under the acronym *smaller, quicker, cheaper* (SQC) methods of literacy assessment (Wagner, 2003; subsequently in Wagner, 2011). The idea was to see whether LSEA methodologies could be reshaped into hybrid methods that are just big enough, faster at capturing and analyzing data, and cheaper in terms of time and effort.[2] The resulting methodology would be flexible enough to be adaptable to local LIC contexts, and, in particular, be able to deal with key problems such as ethnolinguistic diversity in many of the world's poor countries. The Early Grade Reading Assessment (EGRA; Research Triangle Institute, 2009) contains a number of the above features and is probably the best-known current example of a hybrid assessment in reading acquisition. The EGRA was initially designed with three main assessment goals: early reading (grades 1–3), local context focus (rather than comparability across contexts), and local linguistic and orthographic variation. Hybrid assessments (such as those used to produce the data in Figure 20.1) can provide relatively simple outcome indicators that may be conducted on specific population samples.

What Is Compared in Assessments?

Comparability is at the heart of all assessment instruments. It is also a core function for ensuring early child development and to determining which children need more attention in order to achieve a benchmark of sufficient progress. It is possible to identify four key areas that allow assessments themselves to be compared with one another: credibility (in terms of validity and reliability), sampling, scaling, and

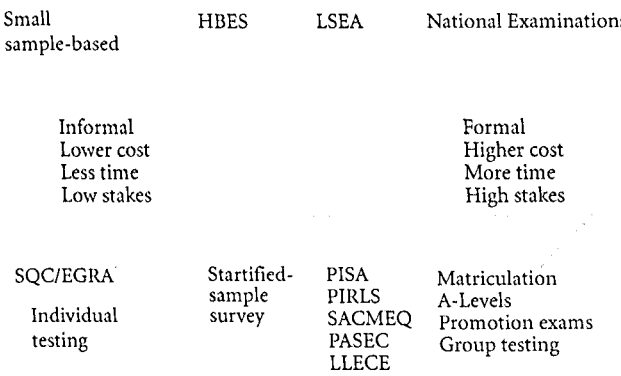

FIGURE 20.4 *Assessment continuum. Ranging from SQC hybrid assessments to LSEA and National Examinations.* Adapted from Kanjee, A. (2009, October). *Assessment overview.* Presentation at the First READ Global Conference, Moscow.
(*Note:* HBES refers to household-based educational surveys; see text for other acronyms.)

implementation. Each will be considered in turn; a schematic diagram of assessment types is shown in Figure 20.4.

Credibility

All assessments depend on the credibility through which well-trained scientists and experts can achieve consensus on the merits of a particular set of findings, even if they might disagree with the interpretation of such findings. The two most oft-cited components of assessment science are validity and reliability. The *validity* of an assessment instrument is the degree to which items on a test can be credibly linked to the conceptual rationale for the testing instrument. Thus, having read a paragraph in an assessment, does the child's answers to questions (on, say, a multiple-choice test) really relate to a child's ability to read, or to the ability to remember what he or she has read earlier? Validity can vary significantly by setting and by population, since a test that might be valid in London may have little validity in Lahore. A reading test used effectively for one language group of mother-tongue speakers may be quite inappropriate for children who are second-language speakers of the same language.

Reliability is typically measured in two ways. Generically, *reliability* refers to the degree to which an individual's score on a test is consistently related to additional times that the individual takes the same (or equivalent) test. High reliability usually means that the rank ordering of individuals taking a given test would, on a second occasion, produce a very similar rank ordering. In the psychometrics of assessment, it is not unusual to obtain relatively high test–retest reliability on LSEAs. This result stems in large part from the fact that assessments of human cognitive function (of many kinds) tend to be highly stable. A second way to measure reliability is in terms of the internal function of the test items: Do the items in each part of an assessment have a strong association with one another? This is

inter-item reliability (measured by Cronbach's alpha statistic). And, if the test is administered by two different individuals, then the reliability of the instrument can also be judged by interrater reliability.

Overall, there are numerous ways of thinking about the credibility of any assessment. Within the measurement community, credibility is typically thought of as a combination of validity and reliability. Yet, in the non-statistical sense, credibility implies more than the particular statistical tools available to test designers. This is so largely due to the fact that many of the difficult decisions about credibility are made *before* statistical tests are employed. For example, is an assessment credible if many of the poorest children are excluded from participation? Is an assessment credible if the enumerator does not speak the child's language? Is an assessment credible if some children have taken many such tests before, while for others this is the first time? These are not merely choices that are internal to the test, but rather are related to the context in which the assessment is deployed, and who is the user of the assessment.[3]

Sampling of Skills and Populations

The majority of LSEAs tend to utilize standardized tests in a particular domain, such as reading, math, or science. The approach relative to a domain can vary widely across tests, even if the same domain is tested in multiple different assessments. The assessments mentioned earlier—PIRLS, PISA, LLECE, SACMEQ, and PASEC—are essentially based on the school programs of the countries concerned. These assessments generally try to evaluate the match between what should have been taught (and learned), and what the student has actually learned (as demonstrated by the assessment). All are administered in writing as group-administered tests in school settings, with a main focus on reading comprehension. By contrast, the EGRA contains a set of measures that are individually administered and are primarily based on a number of reading fluency skills developed originally for diagnostic purposes in beginning reading (see Wagner [2011] for a more in-depth analysis of these measures).

The representativeness of the sample population is a fundamental part of all assessments, and all the assessments mentioned above take seriously this aspect of measurement design.[4] Nonetheless, it is often the case that many of the populations of children most in need are systematically excluded from measurement in LSEAs. This seems to be both the result of, and indeed a cause of, exclusion from LSEAs of vulnerable and marginalized populations. The rationales vary from assessment to assessment, and from one national policy to another, and yet the result is the same—those least likely to succeed on tests, and those who are most disadvantaged, represent the groups most often excluded from the sample population for assessment. In particular, it is not unusual for children who speak "minority" languages to be excluded from assessments. This may be particularly accentuated in areas where civil conflict or economic distress leads to substantial cross-border migration, where immigrant groups (and their children) are treated as "transients,"

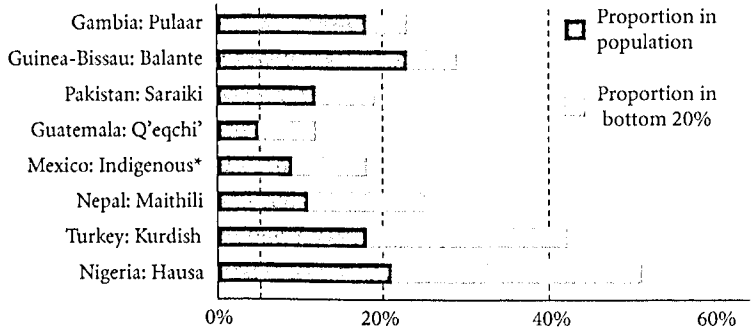

FIGURE 20.5 *Percent of selected language groups in the bottom 20% of the education distribution, selected countries.*

Note: *The indigenous language category in Mexico consists of those who speak indigenous languages only and do not speak Spanish

Adapted from UNESCO, 2010, p. 152.

and where groups may be provided with little or no schooling. This is particularly unfortunate since current evidence suggests that such ethnolinguistic minorities make up a disproportionate percentage of illiterates in LICs (Figure 20.5).

Further, each of the LSEAs described above selects children from those already enrolled in school, thus excluding out-of-school children, the group probably most in need of assistance. In addition, international and regional LSEAs contribute in other ways to exclusion, in singling out children already determined to be dyslexic or with mental or physical handicap (PISA), those who are enrolled in "small schools" (SACMEQ), those who have not sufficiently mastered the language of the assessment, and those too young to be tested in group format. The EGRA, with its focus and testing in local languages, individualized testing, and the propensity to sample among the most disadvantaged young children, has the least statistical need to make population exclusions.

Comparability Between and Within Countries

International statistical reports on education typically base their datasets on national reports, where data may have many different ways of being collected. In contrast (and as one of the attractions of LSEAs) is that nations may be rank-ordered in *league tables* (as in PISA and PIRLS). Naturally, there can be problems in applying a common skill sampling scale across widely differing populations. In the 2006 PIRLS study of reading achievement, the median score of South African grade 4 students was below the "0" percentile of the high-income OECD nations (Crouch, 2009). Also, EGRA scores used in English in rural Kenya are far lower than for same-age (or grade) English-speaking students in suburban Washington, D.C. (Research Triangle Institute, 2008). Such dramatic disparities raise considerable concern about the gap that will need to be closed for low-income countries to catch up to high-income countries.[5]

Can both comparability and context sensitivity be appropriately balanced in assessments? Should countries with low average scores be tested on the same

scales with countries that have much higher average scores? If there are countries (or groups of students) at the "floor" of a scale, some would say that the solution is to lower the scale of difficulty. Others might say that the scale itself is flawed, and that there are different types of skills that could be better assessed, especially if the variation is strongly influenced by race, ethnicity, and language. Having different scales for different groups (or nations) seems to some specialists to be an unacceptable compromise of overall standards or international benchmarks.

To the extent that comparability can be achieved (and no assessment claims perfect comparability), the results allow policymakers to consider their own national (or regional) situation relative to others. This seems to have most merit when there are proximal (as opposed to distal) choices to make. For example, if a neighboring country in Africa has adopted a particular bilingual education program that appears to work better in primary school, and if the African minister believes that the case is similar enough to his or her own national situation, then comparing the results of, say, primary school reading outcomes makes good sense. A more distal comparison might be to observe that a certain kind of bilingual education program in Canada seems to be effective, but there may be more doubt about its application in a quite different context in Africa. But, proximity is not always the most pertinent feature: There are many cases (the United States and Japan, for example) where rivalries between educational outcomes and economic systems have been a matter of serious discussion and useful debate over the years.

The key issue here is the degree to which it is necessary to have full comparability, with all individuals and all groups on the same measurement scale. Or, if a choice is made to not "force" the compromises needed for a single unified scale, what are the gains and losses in terms of comparability? Alternatively, one might ask whether the assessments need to measure the same attributes: For example, the EGRA focuses mainly on cognitive prereading skills (such as phonemic awareness), whereas international LSEAs focus mainly on reading comprehension. Can international statistics be maintained as stable and reliable if localized approaches are chosen over international comparability? This question has led to situations in which some LICs, although tempted to participate in international assessments, nevertheless hesitate due to the possible appearance of very low results or the feeling that the expense of participation is not worth the value added to decision making at the national level. Others may participate because they do not want to be viewed as having "inferior" benchmarks to those used in OECD countries (Greaney & Kelleghan, 1996).

Implementation

School-based assessments are typically implemented with two key parameters in mind. First, there are "break points" at which a student will leave one level of education for another more advanced stage. Thus, there exist in many countries national examinations held at the end of primary, lower secondary, and upper secondary school, to determine who will be allowed into the next stage of the

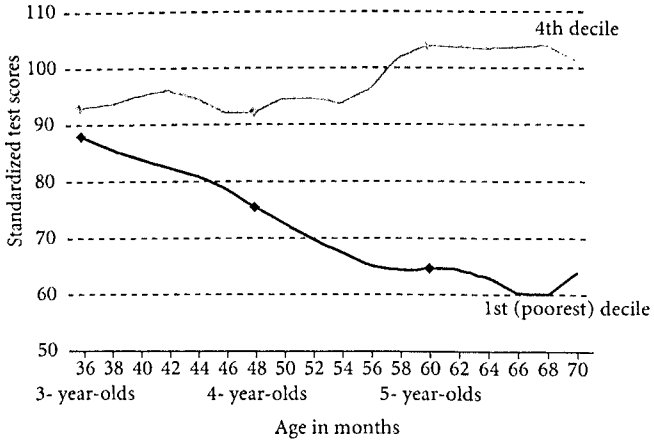

FIGURE 20.6 *Wealth-based gaps: Test scores across ages for the poorest and the fourth deciles in Ecuador, 2003–2004.*

Adapted from UNESCO, 2010, p. 50.

schooling system. Second, there are exams that view the level of competency as a more appropriate cognitive point in which students should be tested. As noted earlier, PIRLS tests children at the end of grade 4, which is the point at which (in OECD countries) it was determined that most children should have learned the basics of reading, writing, and math. Hybrid assessments like the EGRA focus mainly on the period from grades 1 to 3, which permits one to ascertain serious reading problems much earlier than do the other LSEAs. This aspect of early detection is made possible in part due to the one-on-one and largely oral assessments given to children. There is a very important policy rationale as well. In the field of early childhood education, there is growing consensus on the positive impact of early intervention (Heckman, 2006; see also Behrman & Urzúa, 2013, Chapter 6, this volume). Further, research in LIC contexts shows that wealth-based gaps in children's cognitive development grow over time (see Figure 20.6). Taken as a whole, it is widely accepted that the earlier one can detect and remedy educational problems, the more effective the intervention.

International and regional assessments are typically carried out on a 3-, 5-, or even 10-year cycle. If the goal is for a tighter relationship between findings and policies that can be implemented during the annual school cycle, or within the mandate of a typical minister of education, then greater frequency of assessment is required. Achieving this latter aim will likely necessitate instruments such as hybrid instruments whose turnaround time is usually less than 1 year and whose smaller sample size (and lower cost) will allow greater frequency of repetition (Wagner et al., 2011).

One of the most difficult implementation questions concerning LSEAs is how much data and of which kind to collect. The idea that one collects "just enough" data is easier said than done. What some term "right-sizing" data collection has

been more recently called "evidence-centered design" (Braun & Kanjee, 2006). Each of the international and regional assessments utilizes a survey that is undertaken at the school level, with techniques that allow the use of extended passages, like a newspaper article, in the assessment of reading comprehension. But newspaper content can vary by language, topic, intent, and more. As has been pointed out by McCall (2009), comparing implementations of very similar interventions (such as assessments) can be remarkably problematic due to the varied contexts in which such activities are actually carried out in real-life settings.

Each of the three types of assessments reviewed above varies by sampling, scaling, and implementation parameters—with an overall impact on assessment credibility. Further, each assessment approach provides for a degree of comparison within and between population groups (or nations). The ultimate value of a given assessment will depend on the policy purpose to which it is put, such as international comparison or local validity. Finally, there is the key issue of who the end-users of such learning assessments are and how the information gathered impacts policy planning. This brings us to the matter of stakeholders, those persons who have an interest in the data collected.

Stakeholders, Learning Assessments, and Policy Outcomes as Public Goods

In most countries (and perhaps especially so in LICs), educational specialists and statisticians (composing a rather narrow group of individuals) are the primary guardians of learning assessment results. This restricted access to knowledge about learning achievement is likely due, at least in part, to the complexities of carrying out large-scale assessments and the difficulty of interpreting complex datasets. In addition, there may be reticence among policymakers who might worry about publicized assessment differences between the *internal* societal backgrounds of children (such as by ethnolinguistic groups, private and public schools, etc.) in national policy debates on education. Even *external* national comparisons (e.g., when Singapore or Finland clearly outdistances U.S. schoolchildren on various science and math tests) can be used to promote national policies by demonstrating that more investments are needed to "catch up" with top global educational winners. As has been studied in OECD countries, when parents and community groups become aware of the poor scores in their schools relative to others (perhaps through newspaper accounts), they, too, can become consumers of results and actors in social and political change.

In other words, the benefits of greater transparency in learning assessment outcomes are becoming more widely understood and sought after across the globe—and by a wider array of stakeholders. Whether due to improved accountability by governments, influences of international agencies, efforts of Non-Governmental Organizations (NGOs), or greater community activism, there is little doubt that

interest in children's learning and educational success has become increasingly important. Parents the world over—rich and poor, in OECD or LIC contexts—increasingly recognize the importance of learning and learning assessment results in determining the future of their children in personal, social, and economic development. Collectively, parent and community involvement leads to empowerment, as outcomes have an impact on how policymakers collect data on learning and manage the results that come from assessments. In other words, educational outcomes, and the policies that derive from them, are rapidly becoming *public goods* that can empower community action in LIC contexts where centralized educational systems have held sway for decades.

Learning assessments, as part of educational impact measures, are playing an increasingly important role in the growth of policy transparency across the world. Although relatively recent in LICs, the utility of learning assessments will depend on the value of the information collected to specific groups of stakeholders—so that change is possible, negotiable, and expected. In this way, learning assessments will break new ground in educational, social, and (therefore) economic accountability. In order to achieve the educational priorities of the UN MDGs, it is critical to sustain a significant policy and assessment focus on poor and marginalized populations, and to enable parents and communities to be able to interpret the findings of this work. In other words, learning assessments can and should provide timely and understandable feedback to those who care about young girls like Aminata—and thereby enhance educational quality.

Acknowledgments

Parts of this paper are drawn from Wagner (2010, 2011). Thanks to Patrice Engle for her very helpful editing comments throughout. All remaining views are those of the author.

Notes

1. There is some ambiguity on whether it is universalistic scientific approaches that constrain inquiry to normative samples in industrialized countries or whether the limited research in other countries is simply due to historical limitations on human and fiscal support for such research. My own view is that both are likely to be responsible for our limited ability to bring adequate research to bear on educational problems, such as improving reading. This is, as noted below, beginning to change.
2. Another distinction is that hybrid measures, such as the EGRA, tend to be individually administered (made necessary in large part due to the younger age of children typically assessed—aged 6–8 usually), whereas the LSEAs tend to be group-administered tests made possible by the older age of the children involved.

3. There are many possible users (or stakeholders) in such assessments, from the parent to the teacher to the headmasters up to the minister of education. Each might define credibility in a somewhat different way.
4. The PIRLS uses a sample of at least 150 schools with students in grade 4, but the sample may be heterogeneous by age, especially in LICs, where late school enrollment and/or grade repetition is frequent. The LLECE takes into account stratification criteria including type of geographical area and type of school (public or private); about 4,000 students are chosen, with half between the two grades tested (grade 3 and grade 4). The LLECE evaluates students in two adjacent grades (grade 3 and grade 4) as part of data collection. The SACMEQ evaluates students reading in grade 6, with a sampling technique similar to that of the PIRLS. The PASEC focuses on children enrolled in the grades 2 and 5. In the PISA, the main criterion for choosing students is their age (15 years). The EGRA assessments are done during grades 1, 2, and 3, with sample sizes typically ranging between 800 and 6,000 children.
5. In addition, floor and ceiling effects are much more likely when skill results vary significantly across population sampling, thus invalidating statistical comparisons.

References

Behrman, J., Parker, S., & Todd, P. (2009). Schooling impacts of conditional cash transfers on young children: Evidence from Mexico. *Economic Development and Cultural Change, 57*, 439–477.

Behrman, J. R., & Urzúa, S. S. (2013). Economic perspectives on some important dimensions of early childhood development in developing countries. In P. R. Britto, P. L. Engle, & C. M. Super (Eds.), *Handbook of early childhood development research and its impact on global policy* (Chapter 6). New York: Oxford University Press.

Braun, H., & Kanjee, A. (2006). Using assessment to improve education in developing nations. In J. E. Cohen, D. E. Bloom, & M. Malin. (Eds.), *Improving education through assessment, innovation, and evaluation* (pp. 1–46). Cambridge, MA: American Academy of Arts and Sciences.

Crouch, L. (2009). *Literacy, quality education, and socioeconomic development* (Powerpoint presentation). Washington, DC: USAID.

Gove, A., & Cvelich, P. (2010). *Early reading: Igniting Education for All.* A report by the Early Grade Learning Community of Practice. Washington, DC: RTI.

Greaney, V., & Kellaghan, T. (1996). *Monitoring the learning outcomes of education systems.* Washington, DC: World Bank.

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science, 312*(5782), 1900–1902.

Kamil, M. L., Mosenthal, P. B., Pearson, P. D., & Barr, R. (Eds.). (2000). *Handbook of reading research: Volume III.* Mahwah, NJ: L. Erlbaum.

Kanjee, A. (2009, October). *Assessment overview.* Presentation at the First READ Global Conference, Moscow.

LeVine, R. A., LeVine, S. E., Schnell-Anzola, B., Rowe, M. L., & Dexter, E. (2012). *Literacy and mothering: How women's schooling changes the lives of the world's children.* New York: Oxford University Press.

McCall, R. B. (2009). Evidence-based programming in the context of practice and policy. *SRCD Social Policy Report, 23*(3), 3–20.

Research Triangle Institute (RTI). (2008). *Early grade reading Kenya baseline assessment analyses and implications for teaching interventions design* (Final Report). Washington, DC: Author.

Research Triangle Institute (RTI). (2009). *Early grade reading assessment toolkit.* Washington, DC: Author.

Rogoff, B., & Angelillo, C. (2002). Investigating the coordinated functioning of multifaceted cultural practices in human development. *Human Development, 45*, 211–225.

Savedoff, W. D., Levine, R., & Birdsall, N. (2006). *When will we ever learn? Improving lives through impact evaluation.* Washington, DC: Center for Global Development.

Street, B. V. (2001). *Literacy and development: Ethnographic perspectives.* London: Routledge.

UNESCO. (2004). *EFA global monitoring report 2005. The quality imperative.* Paris: Author.

UNESCO. (2005). *EFA global monitoring report 2006. Literacy for life.* Paris: Author.

UNESCO. (2010). *EFA global monitoring report 2010. Reaching the marginalized.* Paris: Author.

Wagner, D. A. (1983). (Ed.). *Child development and international development: Research-policy interfaces.* San Francisco: Jossey-Bass.

Wagner, D. A. (1986). Child development research and the Third World: A future of mutual interest? *American Psychologist, 41*, 298–301.

Wagner, D. A. (2003). Smaller, quicker, cheaper: Alternative strategies for literacy assessment in the UN Literacy Decade. *International Journal of Educational Research, 39*(3), 293–309.

Wagner, D. A. (2004). Literacy(ies), culture(s) and development(s): The ethnographic challenge. *Reading Research Quarterly, 39*(2), 234–241.

Wagner, D. A. (2010). Quality of education, comparability, and assessment choice in developing countries. *COMPARE: A Journal of Comparative and International Education, 40*(6), 741–760.

Wagner, D. A. (2011). *Smaller, quicker, cheaper: Improving learning assessments in developing countries.* Paris/Washington: UNESCO-IIEP and EFA-FTI.

Wagner, D. A., Babson, A., & Katie M. Murphy. (2011). How much is learning measurement worth? Assessment costs in low-income countries. *Current Issues in Comparative Education, 14*, 3–23.

Wagner, D. A., Venezky, R. L., & Street, B. V. (Eds.). 1999. *Literacy: An international handbook.* Boulder, CO: Westview Press.