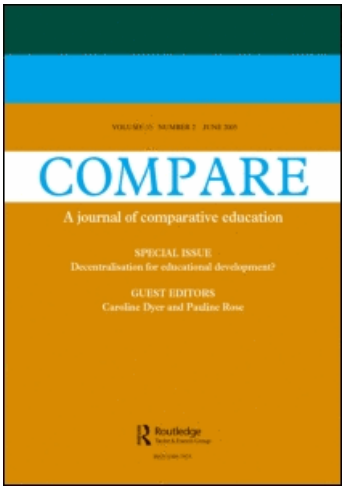

Compare: A Journal of Comparative and International Education
Publication details, including instructions for authors and subscription information: http://www-intra.informaworld.com/smpp/title~content=t713410984

Quality of education, comparability, and assessment choice in developing countries

Daniel A. Wagner[a]
[a] Graduate School of Education, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Online publication date: 19 November 2010

To cite this Article Wagner, Daniel A.(2010) 'Quality of education, comparability, and assessment choice in developing countries', Compare: A Journal of Comparative and International Education, 40: 6, 741 — 760
To link to this Article: DOI: 10.1080/03057925.2010.523231
URL: http://dx.doi.org/10.1080/03057925.2010.523231

Compare
Vol. 40, No. 6, December 2010, 741–760

Routledge
Taylor & Francis Group

Quality of education, comparability, and assessment choice in developing countries[1]

Daniel A. Wagner*

Graduate School of Education, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Over the past decade, international development agencies have begun to emphasize the improvement of the quality (rather than simply quantity) of education in developing countries. This new focus has been paralleled by a significant increase in the use of educational assessments as a way to measure gains and losses in quality. As this interest in assessment has grown, low-income countries have begun to adopt and adapt international and other assessments for a variety of uses, including the comparability of national quality with other countries, improved ways of measuring reading achievement and further attempts to reach marginalized populations within a country. The present paper reviews a number of international, national and 'hybrid' assessments, and considers their merits in terms of how learning is measured, as well as their credibility, sampling and scaling methodologies. The new hybrid assessments, for example, provide innovative opportunities for early intervention for children in their local languages. They also put a premium on local validity over international comparability. The review concludes that there is no single assessment with a dominant scientific superiority, nor is strict comparability across populations or nations a requirement. Rather, different assessments have different policy and practical purposes, and can be used in important and differing ways to improve educational quality. Educational decision makers working in developing countries have important assessment needs and priorities, and will have to choose carefully in order to address them.

Keywords: assessment; quality of education; comparability; choice

The World Conference on Education for All (EFA) in Jomtien (Thailand) in 1990 was considered to be a watershed moment in international education and development. It is important to signal two key themes of this event: first, across several educational goals, there would be a focus on the education of children in poor countries; and second, there would be a cross-cutting effort to promote the *quality of learning* in education, not just counting who was or was not in school.[2] In 2000, at a further EFA conference in Dakar, these same two themes were reinforced in a more detailed list of six education targets.[3] They were reinforced again in the UN Millennium Development Goals (MDGs) for 2015.[4]

With these goals and themes in place, it was realized that improved ways of measuring learning outcomes were going to be required, especially in the poorest developing country contexts. It was thought that with improved assessment methodologies and greater capacity for data collection and analysis, it would be possible to address the increased need for credible data on learning achievement in a truly global

*Email: wagner@literacy.upenn.edu

ISSN 0305-7925 print/ISSN 1469-3623 online

DOI: 10.1080/03057925.2010.523231
http://www.informaworld.com

perspective. Indeed, in the years following Jomtien and Dakar, various initiatives began that would devote substantial new resources to learning achievement and its measurement.

Educational quality is not, however, only a matter of international political commitment, sufficient funding, technical expertise and human resources. Rather, there are important choices to be made about which information (that is, data) will be sought and listened to, and for which stakeholders. One may consider, for example, the following types of stakeholder questions:

- At the international level. A donor agency might ask: How can we (the international/donor community) better judge the current status of learning across countries? Further, which countries should be compared? Or, what kind of learning is common enough across countries that would allow 'fair' comparison?
- At the national (country) level. A Minister of Education might ask: How can we improve the flow of talent through the pyramid of education, ensuring that all pupils at least attain some threshold amount of learning, while assuring that those with most talent rise as high as possible in the education system? How can we help our system do better?

Such questions will vary not only by type of stakeholder, but also by country, gender, ethnic and linguistic group, as well as by region within and across countries. This variation begins to point toward the inequalities that exist (and, importantly, are *perceived* by stakeholders to exist) across and between various group memberships. In other words, the assessment of learning necessarily begins to play a substantial role in helping to shape policies that can drive educational quality and educational change.

The goal of improved quality of education

Educational *quality*, the subject of the 2005 EFA *Global Monitoring Report* (UNESCO, 2004), has several core components, including:

- *what* learners should know – the goals of any 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 (for example, 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, and so on);
- *how* learning takes place – the characteristics of learner-teacher interactions (for example, the roles learners play in their learning, teacher and learner attitudes towards learning, other teacher practices, and so forth); and
- *what* is actually learned – the outcomes of education (for example, the knowledge, skills, competencies, attitudes and values that learners acquire).[5]

A second way that educational quality may be considered is through the use of input-output models – where a number of key learner characteristics are taken into account, most particularly what a child has learned at home before arriving at school. The school provides a set of inputs that includes time, teaching methods, teacher feedback, learning materials and so forth. The outcomes of this process, in the learner, may

be a set of cognitive skills learned (such as reading and writing), social attitudes and values, and more. This model points to the importance of the measurement of a variety of outcomes, but leaves out which outcomes depend on which intermediate contextual variables and how we might measure them.

A third way to consider the promise of improved quality of education is to consider how learning achievement has been linked to economic development. For example, numerous studies have demonstrated how the measure the 'return on investment' (ROI) of investments in schooling (measured by basic skills learning) can be applied in developing countries.[6] This is yet another way that international and national government agencies rationalize increases in the quantity and quality of education.

The present analysis is about assessments that compare, as well as a comparison of such assessments. The view of this paper is that there is no single agreed upon type of learning assessment, even though there are some general scientific principles to which most adhere. There is much to consider about assessment choice, but first there must be some agreement on what is it that needs to be measured.

Measuring learning

Countries across the world comprise a multiplicity of populations that vary along ethnic, linguistic, social class, economic and other dimensions. Each country has its own unique history of socio-political development, and its own experiences with formal schooling and broader educational development. The international policy community has its interests as well, mostly in trying to guide and support national decision making, especially in less developed countries (LDCs), to reach EFA and MDG targets. The world of educational measurement intersects with a world of population variation in ways that are often predictable, but also difficult to address. This is not only a matter of international comparability. Rather, variation in populations is endemic in each and every context where children are raised. Each household itself may also contain significant variation, especially if one considers how differently boys and girls may be treated in many societies.

The measurement of learning in education has never been uncontroversial, and it remains so today. Whenever an educational assessment is reported in the media, it is not surprising to hear from critics who challenge the results by claiming a contradictory piece of evidence, or that the assessment itself was flawed for a variety of technical reasons. Thus, when it was learned that French adults scored more poorly than adults in other European countries that participated in the International Adult Literacy Survey (IALS), French officials withdrew from the study, claiming technical flaws in the study itself.[7] Similar stories can be told in nearly every country when educational news is 'negative'. Of course, what might be called 'political defensiveness' is the other side of 'policy sensitivity', and shows, among other things, that measurement can be an important source of debate and social change.

In discussions of learning, test scores themselves often serve as indicators of overall educational quality. Indeed, such indicators can provide solid information on such issues as: how well content in the school curriculum is being learned and understood, a 'formative' measure on teaching and learning policies, and how well learners have done at the main exit points from the school system. This latter type of 'summative' assessment may be criterion- or norm-referenced, and may be used as a means of facilitating (and legitimizing) access to social and economic hierarchies. Since literacy is

a core feature in both the EFA and MDG basic education goals, reading is the indicator that will receive the most attention in the present discussion.[8]

Research has demonstrated that much (or even most) of the statistical variance associated with school success or failure results from inputs that are outside of the school walls, even far outside.[9] Naturally, there are a whole host of experiences that a child brings to school – experiences that involve not only learned facts about his/her life and community, but also attitudes and values, support structures that implicate language, cultural processes, and much more. These inputs are sometimes, for some children, acknowledged when they finally arrive at the primary school door (such as language of instruction, if it matches what is spoken in the home), and sometimes not. In fact, as more is learned about children's lives at home, more is understood about a multitude of types of inputs, as well as mismatches between children and schools. In today's world, where the MDGs try to guarantee universal basic education, it is no longer possible to ignore context – the personal, social, and ethno-linguistic characteristics that children bring to the classroom. Further, there is a growing recognition that reaching the most difficult to reach, or 'marginalized', populations will likely require special attention and increased funding.[10]

In terms of context, research has shown that actual instructional hours in school are often far less than those intended by the educational system. In one recent study, it was found that there were huge losses in quality instructional time for children in a rural village setting, not just from loss of schooling hours (government schools were non-operational for about 25% of the days of the school year), but also due to teachers being off-task (that is, not directly working with the pupils) more than half the time.[11] As a consequence, it is not surprising that this study found that more than one-third of pupils in grade 3 could not read a single word. Similarly, in the area of language exposure, it has been found that, despite national policies, there is great variability in teachers' actual use of the language of instruction in classrooms, resulting in highly significant differences in children's language mastery by region and by instructor.[12] These are precisely the types of dramatic results that have inspired an increased focus on the importance of early learning in poor countries.

Types of assessments

Educational assessments come in a wide variety of styles, contents, and purposes – and they have been around at least since the beginning of national systems of public education that began in France in the nineteenth century.[13] 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.[14] Over the next century, educators and policy makers across the world have endeavoured to make similar decisions based on examinations – hence the growth in the use of assessment instruments in educational planning (see Figure 1). As a consequence, even countries with relatively low incomes and poorly financed educational systems have begun to participate actively in such assessments. Indeed, as shown in Table 1, a substantial number of EFA-FTI[15] countries, among the poorest in the world, have begun during the past decade to invest in a range of assessments.

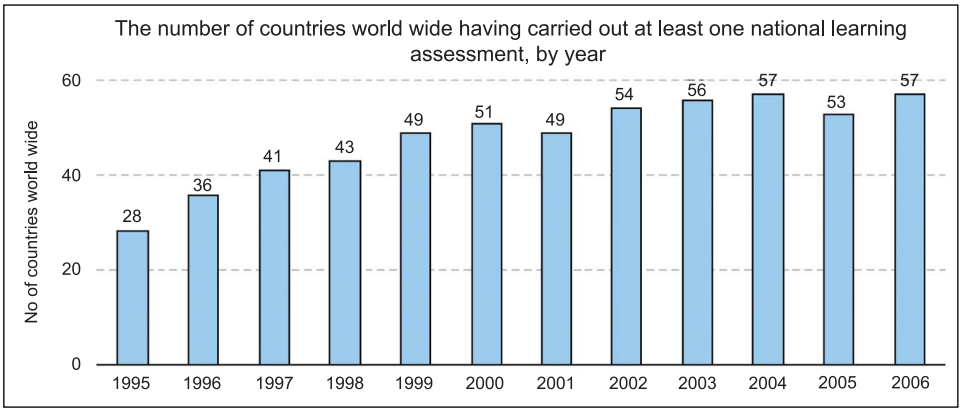

Figure 1. Growth in use of national assessments of learning (1995–2006). (Adapted from Benavot and Tanner 2007, 6).

Large-scale educational assessments

Large-scale educational assessments (LSEAs) have been increasingly used by national and international agencies beginning in the 1980s. Previously, only a small number of cross-national large-scale assessments had been conducted, mostly by the IEA (International Association for the Evaluation of Educational Achievement).[16] Along with political pressure to improve educational systems generally, technological and methodological advances in assessment have especially spurred this trend in LDCs.[17] The 1990 Jomtien Conference demanded more accountability and systemic evaluation in LDCs, and LSEAs became increasingly a key tool for meeting this demand.[18] Further, in 2000, the UNESCO Dakar Framework for Action called for the achievement of 'measurable' learning outcomes, and that such progress should be 'monitored systematically'.[19]

Despite this momentum, the increasing complexity and expense of LSEAs have led some to question the utility of conducting LSEAs in low-income countries.[20] Although a number of agencies have carried out LSEAs in the OECD countries, it was not until the 1990s that the capacity to participate in LSEAs (international and regional) became more available to LDCs.[21] The diversity of stakeholder interests, in addition to resource constraints, has limited growth of LSEAs in LDCs. However, various agencies, such as the World Bank, have become increasingly important funders of LSEAs, making it more affordable and more likely for such assessments to be utilized even when national budgets are very constrained.[22]

For the present purposes, with a focus on reading in low-income countries, the present discussion centres on three main types of LSEAs: international, regional and hybrid. Each of these is described below.

International assessments

International assessments 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 order by rank achievement scores

Table 1. EFA-FTI countries' participation in international, regional and hybrid assessment studies, during the past decade (Adapted from Encinas-Martin 2008, 30–1; and from RTI 2009).

COUNTRY	INTERNATIONAL	REGIONAL	HYBRID
AFRICA			
Benin			
Burkina Faso		PASEC	
Cameroon		PASEC	
Central African Rep.		PASEC	
Ethiopia			
Gambia			EGRA
Ghana	TIMSS 2003, SISS		EGRA
Guinea		PASEC	
Kenya		SACMEQI & II	EGRA
Lesotho		SACMEQII	
Liberia			EGRA
Madagascar		PASEC	
Mali			EGRA
Mozambique		SACMEQII	
Niger		PASEC	EGRA
Rwanda			EGRA
Sao Tome & Principe			
Senegal		PASEC	EGRA
Sierra Leone			
ARAB STATES			
Djibouti	TIMSS 2003, 2007	PASEC	
Mauritania			
Yemen	TIMSS 2003, 2007		
ASIA & PACIFIC			
Cambodia			EGRA
Mongolia	TIMSS 2007		
Tajikistan			
Timor-Leste			EGRA
VietNam			EGRA
LATIN AMERICA & CARRIB.			
Guyana			EGRA
Haiti		LLECE, SERCE	EGRA
Honduras	TIMSS 2007		EGRA
Nicaragua		LLECE	EGRA

by nation or region or other variables; and (c) within-country analyses that are then compared to how other countries operate at a sub-national 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 IEA that conducts the *Progress in International Reading Literacy Study*[23] (PIRLS), and the OECD (Organization for Economic Cooperation and Development) that 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 assessments is now in use in dozens of countries, and is expanding well beyond the OECD country user base that formed the early core group of participation.[24] 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 developing countries, 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.

Hybrid assessments

In recent years, a new approach to assessment has sought to focus more directly on the needs of poor LDC assessment contexts. Initially, this approach was conceptualized under the acronym *smaller, quicker, cheaper* (SQC) methods of literacy assessment.[25] The idea was to see whether LSEA methodologies could be reshaped into *hybrid*[26] methods that are: just big enough, faster at capturing and analysing data, and cheaper in terms of time and effort. The resulting methodology would be flexible enough to be adaptable to local contexts, and in particular be able to deal with key problems such as ethno-linguistic diversity in many of the world's poor countries. The *Early Grade Reading Assessment* (EGRA)[27] contains a number of the above features, and is probably the best-known current example of a hybrid assessment in reading. 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. EGRA, as a hybrid assessment, has different goals than those generally put forward by LSEAs. Hybrid methods do not necessarily make the assessment task simpler or easier, but they do put the emphasis in different places.

What is compared in assessments?

Comparability is at the heart of assessment. From early work on intelligence testing to the current debates about children and school achievement worldwide, the role of

comparison and test 'fairness' has never ceased to be challenged. Similarly, in LSEAs used cross-nationally, there are legitimate concerns as to what constitutes an appropriate science of comparison. While an in-depth discussion of this topic is beyond the space constraints of this article, 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 implementation. Each will be considered in turn.

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, do questions on 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. With respect to international LSEAs, there have been a number of critiques of content validity, around the choice and appropriateness of items, given their application to local cultures and school systems.[28] It seems that regional tests do somewhat better on this aspect of validity, as they have tended to use material from the stated national curricula as items in the test itself.[29] Translation of international LSEAs remains a problem, as it often uncertain whether an equivalent translated item will have the same statistical properties as an indigenous word chosen independently.[30]

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 human cognitive function is highly stable. A second, and easier, 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). Of course, reliability implies little about the validity of the instrument, wherein agreement must be reached concerning the relevance of the instrument for educational outcomes. Nonetheless, reliability is crucial to achieve in any LSEA, and failure to achieve a relative high level may indicate serious ceiling or floor effects.

Credible comparability is central to global education data collection, such as the data collection carried out by the UNESCO Institute for Statistics (UIS). Nonetheless, if comparability becomes the primary goal, while less attention is paid to the (local and cultural) validity of the definitions and classifications of learning, then the data may become less meaningful and potentially less applicable at the ground level. This is a natural and essential tension between 'emic' (within-culture) and 'etic'

(cross-culture) approaches to measurement, and is particularly relevant to marginalized populations.[31]

Overall, there are various ways of thinking about the credibility of any assessment. Within the measurement community, credibility is defined 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.

Sampling of skills and populations

The majority of LSEAs tend to utilize standardized tests in a particular domain, such as reading, mathematics or science. The approach relative to a domain can vary widely across tests, even if the same domain is tested in multiple different assessments. Assessments such as PIRLS, LLECE, SACMEQ, and PASEC are essentially based on the school programmes of the countries concerned. The 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). For example, PIRLS assesses achievement in reading comprehension. Reading comprehension processes include the following areas: locating and explaining particular items of information; drawing inferences from logical or chronological sequences and interrelated events; interpreting and integrating ideas and information; and, examining and evaluating content, language and textual elements. In LLECE, tests include both multiple choice and open-ended items; language components include reading comprehension; metalinguistic skill; and production of written text.[32] SACMEQ adopted the definition of reading literacy used in PIRLS.[33] PASEC tests were constructed in French on the basis of elements that are common to curricula in francophone countries in Africa.[34] In PISA, skills tested include: knowledge and skills applied in personal, public, occupational and educational settings; content or structure of texts (continuous, or in tables, charts or forms); and processes that need to be performed, such as retrieval, reflection, evaluation and interpretation of written text. All of the above assessments were administered in writing as group-administered tests in school settings. By contrast, 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.[35]

The representativeness of the sample population is a fundamental part of all LSEAs. PIRLS uses a sample of at least 150 schools with students in fourth grade. The sample may be heterogeneous by age in some of the countries, and in particular in developing countries where late school enrolment and/or grade repetition is frequent.[36] LLECE takes into account stratification criteria including: type of geographical area (metropolitan, urban area, rural area) and type of school (public or private). About 4,000 students are chosen (40 per school), with half between the two grades tested (grade 3 and grade 4). LLECE evaluates students in two adjacent grades (grade 3 and grade 4) as part of data collection. Depending on the particular country, students were

either 8 or 9 years old.[37] SACMEQ evaluates students' reading in grade 6, with a sampling technique similar to that of PIRLS.[38] PASEC focuses on children enrolled in grades 2 and 5. The sampling was carried out at two levels: first, a sample of schools is selected that is proportional to their weight in the number of students in each of the two grades; second, schools are chosen by stratification, in such a way as to be representative of the national education system as a whole.[39] PASEC evaluates grades 2 and 4; in addition, the students are tested at the beginning and the end of the school year for each of the two grades. In PISA, the main criterion for choosing students is their age (15 years), independent of their schooling level and type of institution. This can result in substantially different combinations of learning experiences between countries.[40] EGRA assessments are typically done orally, and during grades 1, 2 and 3. EGRA tends to have smaller sample sizes on average than the other LSEAs, but has a fairly wide range: from 800 children (in Kenya) to up to about 6,000 (in Nicaragua).

It is a persistent irony that many children most in need of better education are systematically excluded from measurement in LSEAs. As is sometimes said among assessment specialists: 'if you are not measured, you do not exist'. This seems to be both the result of, and indeed a cause of, exclusion from LSEAs of vulnerable and marginal 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. To understand why this is so, it is useful to disaggregate what is meant by the term exclusion.

Gender, for example, has been a leading factor in school non-participation in LDCs, though significant progress has been made over recent decades. Nonetheless, it is clear that in the poorest countries, girls continue to be less enrolled in school than boys, both at the point of primary school entry and by about grade 5. Systematic exclusion of girls in poor LDCs, as well as discrimination, usually results in lower participation in schooling among adolescent girls. Similar trends show important differences in assessments when comparing rural and urban areas in LDCs.[41] Further, language variation across ethnic groups exists in nearly all countries, as a result of historical trends and more recent migrations. Many of these groups – termed *ethnolinguistic* minorities – are well integrated into a national mix (for example, in Switzerland), but at other times this may result in civil strife (for example, in Rwanda). Often, there exist social and political forces that try to help resolve differences, usually including policy decisions that result in a hierarchy of 'acceptable' languages to be used in schools and in governance structures. In such situations, whether in OECD countries or LDCs, it is not unusual for children who speak 'minority' languages to be excluded from assessments.[42] This may be particularly accentuated in areas where civil conflict or economic distress lead to substantial cross-border migration, where immigrant groups (and their children) are treated as 'transients' and where groups may be provided with little or no schooling.

Each of the LSEAs described above selects children from those already enrolled in school, thus excluding out-of-school children, the group most in need. In addition, the international and regional LSEAs have further instances of exclusion, such as: children already determined to be dyslexic or with mental or physical handicap (PISA); those who are in 'small schools' (SACMEQ);[43] and, as noted earlier, those who have not mastered sufficiently the language of the assessment. EGRA, with its focus and testing on local languages, and the propensity to sample amongst the most disadvantaged young children, seems to have the least exclusions.

Scaling

International statistical reports on education (such as those produced by UIS, Unicef or the World Bank) typically base their datasets on national reports, where there may have been many different ways of collecting data. In contrast, and one of the attractions of LSEAs is that nations may be ordered by rank in league tables (as in PISA and PIRLS). Yet, as noted above, there may 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.[44] Such dramatic disparities raise considerable concern about the gap that will need to be closed for LDCs to catch up to high-income countries. Naturally, floor and ceiling effects are possible any time when skill results vary significantly across population sampling.[45] For example, EGRA scores used in English in rural Kenya are far lower than for same-age (or grade) English-speaking students in suburban Washington, DC.[46]

As noted earlier, the international and regional LSEAs typically involve *group* based testing in schools, requiring students to be skilled enough to complete a written examination independently. In poor LDCs, especially in the early grades, this approach is nearly impossible, even if one simplifies the content. If the purpose is to assess children at the level of *beginning* reading (which is where many learners in poor countries remain even after two or more years at school), the EGRA methodology makes most sense.

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 drop the scale to a lower level 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 variables are evidently caused by race, ethnicity, language, and related variables that lead one to question the test as much as the group that is tested. Having different scales for different groups (or nations) seems to some to be an unacceptable compromise of overall standards.

To the extent that comparability can be achieved (and no assessment claims perfect comparability), the results allow policy makers 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 neighbouring country in Africa has adopted a particular bilingual education programme that appears to work better in primary school, and if the African minister believes that the case is similar enough to his/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 programme 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 USA and Japan, for example) where rivalries between educational outcomes and economic systems have been a matter of serious discussion and debate over the years.[47] In another example, closer to present purposes, senior officials in Botswana were interested in knowing how Singapore came to be first in mathematics.[48]

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 scales need to measure the same skills: for example, EGRA focuses on cognitive 'pre-reading' skills (such as phonemic awareness), while international LSEAs focus 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 where some LDCs, while tempted to participate in international assessments, nevertheless hesitate due to the 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.[49] Others may participate because they do not want to be viewed as having 'inferior' benchmarks to those used in OECD countries.[50]

Implementation

School-based assessments are typically implemented with two key parameters in mind. First, there are 'break points' when 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 – to determine who will be allowed in the next stage of the school system. Second, there are examinationss 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 (about age 9 in OECD countries), which is the point at which it was determined that most children should have learned the basics of reading, writing and mathematics; PASEC, LLECE and SACMEQ are similarly clustered around the mid-end of primary school. On the other hand, PISA assesses at age 15 in order to capture higher levels of attainment.

Hybrid assessments like EGRA[51] focus mainly on the period from grades 1 to 3, which allows it the ability to ascertain serious reading problems much earlier than 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 evidence of the impact of early interventions, such as those indicating that a dollar spent in the early years will pay off many times over in later life.[52] Further, it is clear from additional studies that the wealth-based gaps in children's cognitive development grow over time (see Figure 2). Taken as a whole, it is widely accepted that the earlier one can detect and remedy educational problems, the more effective can be the intervention.

International and regional assessments are typically carried out on a 3- or 5- or even 10-year cycle for repetition. 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 is likely to necessitate instruments such as hybrid instruments whose turnaround time is usually less than one year, and whose smaller sample size (and lower cost)[53] will allow greater frequency of repetition.

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'.[54] Each of the international and regional

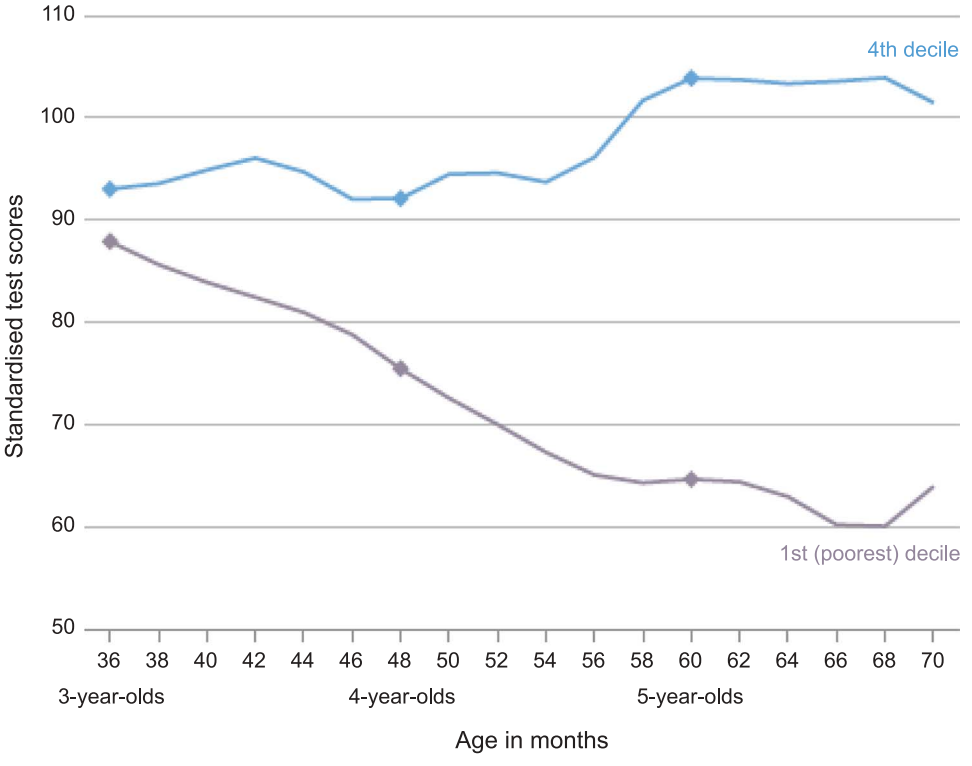

Figure 2. Wealth-based gaps: Test scores (Peabody Picture Vocabulary Test in Spanish) across ages for the poorest and the fourth deciles in Ecuador, 2003–2004. (Adapted from UNESCO 2010, 50).

assessments described earlier is a survey that is undertaken at the school level. These LSEAs utilize a common statistical technique called item response theory (IRT), an approach that increases skill test coverage by allowing more total items in the assessment, but fewer for each individual student to take.[55] In this way, it also allows the use of extended passages, like a newspaper article, in the assessment of reading comprehension. In assessments without IRT (such as PASEC and EGRA), all students respond to a full set of items, providing a transparent comparison across identical sets of items, but also restricting the breadth and depth of what is assessed.[56] As with all such statistical techniques, the IRT, as employed in international assessments, is not without its critics.[57] The EGRA, and other hybrids, by contrast, try to adhere to the SQC model of right-sizing data, by opting for considerably less data collection than that of the other international and regional tests. This approach has much in common with what is called 'short-form' test development wherein longer tests are reduced to smaller ones, with various statistical risks to both validity and reliability.[58]

Each of the 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 comparsion within and between population groups (or nations), but the attractiveness of a given assessment will depend on the policy purpose to which it is put.

Conclusions: a matter of choice

Given the increasing attention now being paid to the quality of education in developing countries, it is not surprising that there is a concomitant growth in interest concerning assessments, and how to choose amongst them. What are the appropriate criteria for such a choice?

Some responses have been addressed above. All ministries of education and those in the broader educational community will insist on the credibility, proper sampling and scaling, and effective implementation of the assessment used. As we have seen, assessments may be compared in differing ways, but no single assessment can be said to be best, since each assessment is designed along a set of compromises to achieve a specific set of policy goals. PIRLS and PISA (and their regional compatriots) have strong empirical designs, and achieve scientifically credible approaches within the domains they assess. But even these well-known tests have made compromises in terms of narrow age and grade focus, population exclusions, languages utilized, training of enumerators, and so forth. For example, league tables, while of value to some nations, may be seen as less useful for LDCs that have scores so close to the floor that comparison with OECD countries is of limited policy value. In other words, international comparability, in terms of 'horse race' comparisons, may be of minimal value to low-income countries. International LSEAs are also too complex to be undertaken on a frequent (say annual or even biennial basis), rendering them of very limited near-term policy or educational change utility.

Hybrid assessments, by contrast, are able to assess children in a one-on-one oral and highly valid way, but they do not try (as yet) to achieve strong cross-national comparability.[59] By focusing on classroom and context level assessments, hybrids can provide a more nuanced understanding of individual and classroom level variables. These relationships can then be compared (or contrasted) with other similar or different contexts.[60] Various reliable indicators (with high face and consequential validity) may be included in, or derived from, hybrid assessments, and these may avoid some of the difficulties of cross-national comparability in LSEAs. Even so, various kinds of comparison need to be a part of any good hybrid assessment, such as comparability across students in a defined sample, within a specified linguistic context, and over time (that is, in longitudinal studies).

In the end, all assessments seek comparability, but in different ways. International and regional LSEAs are aimed at cross-national comparability, while hybrid assessments are more focused on local contexts and increased validity. Hybrids offer some kinds of comparability that LSEAs do not, such as among marginalized populations or younger children. Which types of comparability are most important depends on the policy goals desired, as well as timing and cost considerations. As in comparative education more generally, cultual context will determine whether and when empirical interpretations are deemed credible.[61] Overall, there is little doubt that hybrid assessments put a premium on local validity over international comparability.

In most countries (and especially in LDCs), educational specialists and statisticians are the primary guardians of learning assessment results. This restricted access to knowledge about learning achievement is due, at least in part, to the complexities of carrying out large-scale assessments, but also perhaps to a reticence among policy makers who may worry about publicized assessment differences between groups of children (such as between ethno-linguistic groups, private and public schools, and so forth). The importance of involving multiple stakeholders in education decision

making is today more widely recognized. Whether due to improved transparency by governments, influences of international agencies, efforts of non-governmental organisations (NGOs), or greater community activism, there is little doubt that interest in children's learning has become increasingly important. The response to this growing interest will require both more focused and 'real time' data – results that can be implemented in the near term. With multiple stakeholders there will be greater awareness of both the benefits of, and deficiencies in, schooling. SQC or hybrid assessments have the potential of breaking new ground in accountability and local ownership, largely by having as a clear policy goal the provision of information that matters to specific groups in a timely manner, such that change is possible, negotiable and expected.

Assessments are here to stay, and increasingly will be used globally and locally for a variety of policy and practical purposes. The present discussion has tried to lay out some of the pros and cons of different approaches and choices in learning assessment, with a particular focus on poor and developing countries. There is not a single way to do an assessment, and countries may have very different purposes for their assessments. There is no ideal assessment – rather, there are a variety of scientific approaches that can and will provide solid and credible avenues towards improving the quality of education. One size does not fit all.

Notes

1. Parts of this paper are adapted from a report initially prepared for the International Institute of Educational Planning (IIEP)-United Nations Educational, Scientific and Cultural Organization (UNESCO) and the Fast Track Initiative (FTI); see Wagner (2010).
2. In the present discussion, it should be understood the education focus is almost entirely on schooling, even though education, under EFA, covers a broader range of educational efforts.
3. The six goals of Dakar EFA Framework for Action were: early childhood care; compulsory primary school; ensuring learning needs for all; adult literacy; gender disparities; quality of measurement of learning outcomes. See UNESCO, 2004, 28.
4. United Nations (2000).
5. Adapted from Braun and Kanjee (2006, 5)
6. For example, UNESCO (2004, 42).
7. Blum, Goldstein, and Guérin-Pace (2001). France participated in the 1995 and 1998 IALS. Apparently, there were also differences between the Swiss and French francophone translations.
8. It must be admitted here that while reading is often seen as the most essential of school-based cognitive learning (as evidenced by its inclusion in both EFA and MDG goals), it should not be taken as the only type of learning of relevance in schools. As noted, there is a wide variety of skills, attitudes and values that are 'transferred'in the schooling process. It is clear that reading is important; it is also important that reading is a skill that may be more easily measured than the 'softer' metrics of, say, attitudes and values.
9. Of course, there are many who have looked at the role of socio-economic status (SES) and in-school factors (such as textbooks, teacher training, management and use of resources, and so forth) for explanations of educational outcomes. See, for example, Heyneman & Loxley (1983) a recent more review by Gamaron & Long (2006).
10. See the 2010 GMR report entitled *Reaching the marginalized*, (UNESCO 2010).
11. DeStefano and Elaheebocus (2009, 13), also report that 'students who reported having missed school the previous week had reading fluency rates half those of the students who said they had not missed school. …By itself, student self-reported attendance explains 35% of the variation in a schools average reading fluency.'
12. See, among others, Muthwii (2004), in Kenya and Uganda; also Commeyras & Inyega (2007).
13. One reviewer of this paper correctly noted that curriculum-derived tests orginated from Imperial China. Yet the Chinese examinations were not focused on universal public education (as was the case in post-revolutionary France), but rather on a version of meritocratic selection for public administration.

14. Others, with a less technocratic and more political perspective, would say that the main purpose of such testing is the legitimation of the distribution of the scarce public good of education. Thanks to one of the reviewers for pointing this out.
15. FTI is the Fast Track Initiative. See http://www.educationfasttrack.org/.
16. See Chromy (2002, 84) for a listing of the major studies; also Lockheed (2008, 6).
17. Chromy (2002); Kelleghan and Greaney, (2001, 32).
18. Lockheed and Verspoor (1991).
19. UNESCO (2000, 21).
20. Braun and Kanjee (2006, 8).
21. Greaney and Kellaghan (2008, 8–9).
22. According to a survey of national policy makers Gilmore (2005, 45), World Bank funding has been a key determinant of decision making in LSEA adoption for low- and middle-income countries.
23. While the emphasis is on PIRLS reading studies, some reference is also made to the Third International Math and Science Study (TIMSS) and Second International Science Study (SISS) math achievement studies, also undertaken by the IEA.
24. Kamens and McNeely (2010) point out that increased globalization is one reason for the dramatic increase in the number of countries now participating in international testing. They further claim that globalization has fostered a 'world educational ideology' as well as a 'hegemony of science' – both of which have led to an acceptance of educational testing that is much greater than heretofore seen.
25. ILI/UNESCO (1998). (Wagner 2003, 2010).
26. In this instance, hybrid refers to drawing together some of the elements of LSEAs, national curricular assessments and tests that were initially designed as cognitive assessments of reading and other basic skills.
27. For more on EGRA, see RTI (2009).
28. Sjoberg (2007) claimed that some test items deviated substantially from the stated PISA goal of evaluating competencies for the workforce. Howie and Hugues (2000) found that the TIMSS covered only a very small fraction (18%) of the curriculum of science in grade 7 in South Africa, while as much as 50% in grade 8.
29. See Ross and Genevois (2006) on SACMEQ.
30. See Hambleton and Kanjee (1995) for a discussion on translation issues in international assessments.
31. 'Emic' approaches are those that are conciously focused on local cultural relevance, such as local words or descriptors for an 'intelligent' person. 'Etic' approaches are those that define 'intelligence' as a universal concept, and try to measure individuals across cultures on that single concept or definition. Some also see this as one way to think of the boundary between the disciplines of anthropology (emic) versus psychology (etic). See Harris (1976).
32. UNESCO-LLECE (2008).
33. Elley (1992).
34. CONFEMEN (2008).
35. Most EGRA sub-tests require students to read aloud and therefore require the intervention of an enumerator. The reading aloud tasks involve fluency (that is, accuracy and speed) measured by the mean of correct items processed in one minute. The various subtasks typically include: (a) Engagement and relationship to print. Indicate where to begin reading and the direction of reading within a line and a page; (b) Letter name knowledge (1 minute test) – provide the name (and sometimes the sound) of upper- and lower-case letters distributed in random order; (c) Phonemic awareness – segment words into phonemes (pronunciation of the different phonemes of a word containing from 2 to 5 phonemes), by identifying the initial sounds in different words; (d) Familiar word reading (1 minute test); – read simple and common one-and two-syllable words; (e) Unfamiliar non-word (or pseudo-word) reading (1 minute test) – use of grapheme-phoneme correspondences to read simple nonsense words; (f) Oral reading fluency (ORF) in text reading (1 minute test) – read a short text with accuracy; (g) Reading comprehension – respond correctly to different type of questions (literal, and inferential) about the text they have read (above); (h) Listening comprehension – respond to different type of questions (similar to those used to assess reading comprehension) about a story told by an adult enumerator; and (i) Dictation - write, spell, and use grammar properly through a dictation exercise. For more detail, see the RTI Toolkit (RTI 2009). Not all EGRA studies have used each of these subtests, and changes in subtests are under development.

36. Two additional criteria are important: the geographical location where the school is situated and the status of the school (public school, private school, religious). In some countries, these status criteria are not clear, leading to various problems in comparison.
37. The second LLECE assessment is known as SERCE, and evaluated in grades 3 and 6. See UNESCO-LLECE (2008).
38. In SACMEQ countries, students at lower grades transition between the usage of local and national languages in classrooms in primary school. This language transition occurs generally around grade 3 (or grade 4), with the assumption that the national language has been learned sufficiently for most or all students by grade 6. See Ross and Genevois (2006, 39–41). Of course, this assumption is quite variable from one location to another, and is one of the principal reasons why EGRA assessments in local languages have proven attractive.
39. Stratification is implemented by the type of school or the type of geographical area (rural, urban), but without differentiating the geographical area. When a school is chosen, PASEC proceeds by pooling a fixed number of student groups by each level tested. In all, a minimum of 150 schools is required.
40. For example in France certain 15-year-old students are at the upper secondary level while others are at the lower secondary level. In many countries (especially in LDCs), this results in students being chosen from more than one grade in school; for example, in West Africa, it is not unusual to have 15-year-olds in the lower grades of primary school. It should further be noted that comparisons of LSEAs over time (that is, across years) can be problematic as well, since some countries participate only on an irregular basis.
41. For example, according to Greaney and Kellaghan (2008, 71), various sampling problems for the TIMSS appeared in the Republic of Yemen, where a number of schools did not have grade 4 classes and where nomadic children could not be located.
42. In the United States, for example, in the 2003 National Assessment of Adult Literacy, only English and Spanish literacy were assessed, even though dozens of other languages are used by adult learners in American adult education classes. US Department of Education, NCES, 2009.
43. In Lesotho, if a school had less than ten students in grade 6, it was excluded from the population sample. In Seychelles, Botswana and Tanzania, schools with fewer than 20 students were excluded. In Uganda, students were excluded if they were in zones where a civil conflict was in process. See Ross et al. (2005). See also the SACMEQ II report on Kenya Onsumu, Nzomo, and Obiero (2005).
44. Crouch (2009).
45. In the upcoming 2011 Pre-PIRLS study, lower benchmarks (easier vocabulary, shorter passages, and so on) will be utilized so that more explanatory (statistical power) will be available at the bottom end of the scale. According to Mullis et al. (2009), Pre-PIRLS will also gather more background information on home, schools and classrooms, as well as opportunity to learn.
46. RTI (2008).
47. Stevenson and Stigler (1982).
48. Gilmore (2005, 26).
49. See Greaney & Kellaghan (1996) for a useful overview on this issue.
50. It should be noted that donor agencies often play a role in this decision making by supporting certain assessments as part of a 'package' of support for evaluation capacity building.
51. Another well-known hybrid assessment is that of ASER (2009).
52. Heckman (2006).
53. For an in-depth and comparative assessment of costs, see Wagner (2010). While the cost per learner is currently not very different between international and hybrid assessments, the costs of the latter are dropping as research costs are being reduced. Further, the total cost of carrying out international and regional assessments is much higher typically than in hybrid (more focused and local) assessments.
54. 'The basic idea of evidence-centered design 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'. (Braun and Kanjee 2006, 13).

55. See Hambleton, Swaminathan, and Rogers (1991).
56. There are also some disadvantages with IRT, especially for LDCs beginning an assessment programme. Administration (for example, printing and distribution) is more complex, as is scoring and scaling of scores, while analyses involving individual students or school data can be problematic and require more sophisticated personnel. See Greaney and Kellaghan (2008, 42).
57. See, for example, Goldstein (2004); Goldstein, Bonnet, and Rocher (2007); and Mislevy and Verhelst (1990).
58. Smith, McCarthy, and Anderson (2000). This review describes how various well-known tests have been manipulated into shorter forms, and provides methodological suggestions on how to improve the 'short form' versions.
59. There is some effort, at present, to create a benchmark using one of EGRA's sub-tests, on oral reading fluency (ORF), where tentative norms of 40 or 60 correct words per minute may become a cross-national (and comparative) indicator. At the time of writing, there is considerable debate on this matter.
60. It is also possible to focus on generic benchmarks rather than summative total scores on an international test. For example, the indicators recently advocated by the FTI suggest a school-based benchmark as the 'proportion of students who, after two years of schooling, demonstrate sufficient reading fluency and comprehension to "read to learn"'. One could also use 'read a short text in your first language' as a international benchmark. See http://www.educationfasttrack.org/themes/learning-outcomes/. These indicators are also tied to the use of ORF as a possible benchmark (see previous footnote).
61. See Steiner-Khamsi (2010) for a recent discussion on comparability in comparative education.

References

ASER. 2009. Evaluating the reliability and validity of the ASER testing tools. Unpublished draft. New Delhi. http://www.asercentre.org. (accessed 31 October 2009).

Benavot, A., and E. Tanner. 2007. *The growth of national learning assessments in the world, 1995–2006.* Background paper prepared for the EFA Global Monitoring Report 2008. Paris: UNESCO.

Blum, A., H. Goldstein, and F. Guérin-Pace. 2001. International Adult Literacy Survey (IALS): An analysis of adult literacy. *Assessment in Education* 8, no. 2: 225–46.

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

Chromy, J.R. 2002. Sampling issues in design, conduct, and interpretation of international comparative studies of school achievement. In *Methodological advances in cross-national surveys of educational achievement,* ed. A.C. Porter, and A. Gamoran, 80–116. Washington, DC: The National Academies Press.

Commeyras, M., and H.N. Inyega. 2007. An integrative review of teaching reading in Kenyan primary schools. *Reading Research Quarterly* 42, no. 2: 258–81.

CONFEMEN. 2008. *Vers la scolarisation universelle de qualité pour 2015: Evaluation diagnostique: GABON.* Programme d'analyse des systèmes éducatifs de la CONFEMEN (PASEC). Dakar: CONFEMEN.

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

DeStefano, J., and N. Elaheebocus. 2009. School effectiveness in Woliso, Ethiopia: Measuring opportunity to learn and early grade reading fluency. Unpublished draft, Save The Children.

Elley, W. 1992. *How in the world do students read?* The International Association for the Evaluation of Educational Achievement. The Hague: IEA.

Encinas-Martin, M. 2008. *Overview of approaches to understanding, assessing and improving the quality of learning for all.* Paris: UNESCO.

Gameron, A., and D.A. Long. 2006. *Equality of educational opportunity: A 40-year retrospective.* WCER Working Paper No. 2006-9. Madison, WI: WCER

Gilmore, A. 2005. *The impact of PIRLS (2001) and TIMSS (2003) in low- and middle-income countries: An evaluation of the value of World Bank support for international surveys of reading literacy (PIRLS) and mathematics and science (TIMSS).* New Zealand: IEA.
Greaney, V., and T. Kellaghan. 1996. *Monitoring the learning outcomes of education systems.* Washington, DC. World Bank.
Goldstein, H. 2004. International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education* 11, no. 3: 319–30.
Goldstein, H., G. Bonnet, and T. Rocher. 2007. Multilevel stuctural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioral Statistics* 32, no. 3: 252–86.
Greaney, V., and T. Kellaghan. 2008. *Assessing national achievement levels in education: National assessment of educational achievement.* Vol. 1. Washington, DC: World Bank.
Hambleton, R.K., and A. Kanjee. 1995. Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptation. *European Journal of Psychological Assessment* 11, no. 3: 147–57.
Hambleton, R.K., R. Swaminathan, and H.J. Rogers. 1991. *Fundamentals of item response theory.* Newbury Park, CA: Sage.
Harris, M. 1976. History and significance of the emic/etic distinction. *Annual Review of Anthropology* 5: 329–50.
Heckman, J.J. 2006. Skill formation and the economics of investing in disadvantaged children. *Science* 312, no. 5782: 1900–2.
Heyneman, S.P., and W.A. Loxley. 1983. The effect of primary-school quality on academic achievement across twenty-nine high- and low-income countries. *American Journal of Sociology* 88, no. 6: 1162–94.
Howie, S., and C. Hughes. 2000. South Africa. *The impact of TIMSS on the teaching and learning of mathematics and science,* ed. D. Robitaille, A. Beaton, and T. Plomb, 139–45. Vancouver, BC: Pacific Educational Press.
ILI/UNESCO. 1998. *Literacy assessment for out-of-school youth and adults.* ILI/UNESCO Technical Report from Expert Seminar, Paris, June 1998. Philadelphia, PA: International Literacy Institute, University of Pennsylvania.
Kamens, D.H., and C.L. McNeely. 2010. Globalization and the growth of international educational testing and national assessment. *Comparative Education Review* 54, no. 1: 5–25.
Kellaghan, T., and V. Greaney. 2001. *Using assessment to improve the quality of education.* Paris: International Institute for Educational Planning.
Lockheed, M. 2008. *Measuring progress with tests of learning: Pros and cons for "cash on delivery aid" in Education.* Working Paper Number 147. Washington, DC: Center for Global Development.
Lockheed, M., and A. Verspoor. 1991. *Improving primary education in developing countries.* Oxford: Oxford University Press.
Mislevy, R.J., and N. Verhelst. 1990. Modeling item responses when different subjects employ different solution strategies. *Psychometrika* 55, no. 2: 195–215.
Mullis, I.V.S., M.O. Martin, A.M. Kennedy, K.L. Trong, and M. Sains. 2009. *PIRLS 2011 Assessment Framework.* Boston, MA: Boston College, TIMSS & PIRLS International Study Center.
Muthwii, M. 2004. Language of instruction: A qualitative analysis of the perception of parents, pupils, and teachers among the Kalenjin in Kenya. *Language, Culture, and Curriculum* 17: 15–32.
Onsumu, E., J. Nzomo, and C. Obiero. 2005. *The SACMEQ II Project in Kenya: A study of the conditions of schooling and the quality of education.* Harare: SACMEQ.
RTI (Research Triangle Institute). 2008. *Early grade reading Kenya baseline assessment: analyses and implications for teaching interventions design.* Final Report. Washington, DC: RTI International.
RTI (Research Triangle Institute). 2009. *Early grade reading assessment toolkit.* Washington, DC: RTI International.
Ross, K.N., and I.J. Genevois. 2006. Cross-national studies of the quality of education: Planning their design and managing their impact. Paris: IIEP-UNESCO.
Ross, K.R.., M. Saito, S. Dolata, M. Ikeda, L. Zuze, S. Murimba, T.N. Postlethwaite, and P. Griffin. 2005. The conduct of the SACMEQ II project, ed. E. Onsomu, J. Nzomo, and C.

Obiero. *The SACMEQ II project in Kenya: A study of the conditions of schooling and the quality of education.* Paris: SACMEQ/IIEP.

Sjoberg, S. 2007. PISA and 'real life challenges': Mission impossible? In *PISA according to PISA: Does PISA keep what it promises?,* ed. S.T. Hopmann, G. Brinek, and M. Retzl. Vienna: LIT Verlag. http://folk.uio.no/sveinsj/Sjoberg-PISA-book-2007.pdf. (accessed 23 October 2010).

Smith, G.T., D. M. McCarthy, and K.G. Anderson. 2000. On the sins of short-form development. *Psychological Assessment* 12, no. 1: 102–111.

Steiner-Khamsi, G. 2010. The politics and economics of comparison. *Comparative Education Review* 54, no. 3: 323–42.

Stevenson, H.W., and J.W. Stigler. 1982. *The learning gap: Why our schools are failing and what we can learn from Japanese and Chinese education*. New York: Summit.

UNESCO. 2000. *Dakar framework for action. Education for all: Meeting our collective commitments.* Dakar/Paris: UNESCO.

UNESCO. 2004. *EFA Global Monitoring Report 2005: The quality imperative.* Paris: UNESCO.

UNESCO. 2010. *EFA Global Monitoring Report 2010: Reaching the marginalized.* Paris: UNESCO.

UNESCO-LLECE. 2008. Student Achievement in Latin America and the Caribbean. Results of the Second Regional Comparative and Explanatory Study (SERCE). Santiago, Chile, Unesco, LLECE, Regional Bureau of Education in Latin America and the Caribbean.

United Nations. 2000. United Nations Millennium Declaration. Resolution adopted by the General Assembly. (United Nations A/RES/55/2). http://(www.un.org/millennium/declaration/ares552e.htm) (accessed 23 October 2010).

U.S. Department of Education, NCES. 2009. *Basic reading skills and the literacy of America's least literate adults: Results from the 2003 National Assessment of Adult Literacy (NAAL) Supplemental Studies.* Report NCES 2009–481. Washington, DC: US Department of Education.

Wagner, D.A. 2003. Smaller, quicker, cheaper: Alternative strategies for literacy assessment in the UN Literacy Decade. *International Journal of Educational Research* 39, no. 3: 293–309.

Wagner, D.A. 2010. *Smaller, quicker, cheaper: Improving learning assessments for developing countries.* Paper commissioned by the IIEP-UNESCO and the FTI. Philadelphia, PA: International Literacy Institute.