

Measuring progress with tests of learning: Pros and Cons for “Cash on Delivery Aid” in Education¹

Marlaine E. Lockheed
Center for Global Development

This paper reviews, in non-technical terms, the case for and against using tests of learning for measuring annual educational progress within programs of “Cash on Delivery aid.” It examines the evidence supporting the first of two main assumptions behind Cash on Delivery aid in education -- that progress in learning can be measured validly and reliably --by examining three questions: whether valid and reliable measures of student learning are currently available in developing countries, whether existing tests are capable of registering the changes in educational results called for under “Cash on Delivery aid,” and whether developing countries have the administrative capacity to undertake annual assessments of learning.

The paper does not address a second key assumption, that developing countries have the technical capacity to effect improvement in their education systems, if progress incentives are in place. The evidence regarding the impact of progress incentives, such as are found in the No Child Left Behind (NCLB) Act of 2002, is mixed. Earlier studies tended to find no impact of performance incentives on increased student learning (e.g. Amrein and Berliner 2002) while more recent analyses find a positive effect (Braun 2004, Carnoy and Loeb 2003, Hanushek and Raymond 2005, Rosenshine 2003). But several analysts have questioned whether student learning outcomes can improve within a reform framework that lacks the material and pedagogical support required by the new curriculum (Goertz and Duffey 2003; Kelley, Odden, Milanowski and Heneman 2000). And one study found accountability less cost-effective than other approaches for boosting achievement (Yeh 2007).

The paper is organized as follows: the next section provides a brief description of existing testing activities in developing and transition countries. Section 2 outlines the essential requirements for a measure of learning “progress” at the national level and compares these requirements with what is known about the measures typically provided by

¹ Thanks to Carol Owen (Educational Testing Service) and Abigail Harris (Fordham University) for their constructive comments on an early version of this paper. Support for this paper was provided by the Center for Global Development. Any errors of fact or interpretation are solely those of the author. Comments are welcome.

national, regional and international learning assessments, as carried out in developing and transition countries. Section 3 discusses some technical topics in greater detail. The final section discusses testing costs and options for the use of learning assessments in the context of “Cash on Delivery aid.”

1. Building capacity for measuring learning outcomes

Nearly two decades ago, the first recommendation of the World Bank’s 1990 *Primary Education Policy Paper* called for education systems to:

“Emphasize learning. Developing countries need to increase the number of children who acquire the skills specified in their nation’s curriculum and who successfully complete the primary cycle. To this end, countries must emphasize students’ learning as the key policy objective.” (World Bank 1990: 54).

Regrettably, back in 1990, few developing countries had the capacity to measure student learning, so the donor community undertook to build that capacity through loans and grants. At the World Bank, for example, the number of education projects providing support for national learning assessments increased from no projects funded before 1988 to 27 percent of projects funded in 1991, 70 percent funded during 1990-1994 and around 60 percent of projects funded since 1995 (Larach and Lockheed 1992; Nielsen 2006). Other multilateral donors, including UNESCO and IDB, and bilateral donors, including USAID and CONFEMEN, have supported capacity building for student assessments. These efforts have met with some success, and the context for a heightened interest in “results based aid” where learning is an explicit indicator of results has greatly changed.

This change is clearest at the country level. Whereas in 1990, only a handful of developing countries regularly carried out national learning assessments at the primary level², none had participated in regional learning assessments at the primary level, and fewer than a half dozen had participated in any international learning assessment, in 2008 a variety of assessment systems flourish in developing countries. Increasing shares of developing and transition countries³ are implementing national learning assessments. The percentages of developing and transition countries carrying out at least one national learning assessment have risen dramatically: from 28 percent of developing and 0 percent of transition countries in 1995-1999, to 51 percent of developing and 17 percent of transition countries in 2000-2006 (Benevot and Tanner 2007).

² In 1988, Chile established SIMCE, which assessed math, Spanish and writing at Grade 4; in 1990, Colombia established a sample-based assessment for Grades 3 and 5 in math and Spanish; and in 1985, Thailand established a Grade 6 assessment in math, Thai language and science (Murphy and others 1996). Many countries, often former British colonies, had well-established systems of selection and certification examinations, but these were “high stakes” tests designed as gatekeepers for further education, rather assessments of learning. For example, in 1950, Jamaica instituted an 11+ (Common Entrance Examination) for selection into lower secondary schools.

³ Developing and transition countries include those countries designated as low and middle-income countries by the World Bank.

In addition, many developing and transition countries are participating in international learning assessments; 17 developing countries participated in the 2006 Progress in Reading Literacy Study (PIRLS 2006), 26 in the Programme for International Student Assessment (PISA 2006) and 37 in the 2007 Trends in Mathematics and Science Study (TIMSS 2007). Furthermore, regional learning assessments are ongoing in several regions: 15 countries in Southern Africa participate in the Southern African Consortium for Monitoring Education Quality (SACMEQ III), 22 countries in Francophone Africa participate in the Programme d'Analyse des Systemes Educatifs des Pays de la CONFEMEN (PASEC) and 16 countries in Latin America participate in the Laboratoria Latinoamericano de Evaluacion de la Calida de la Education (LLECE). Can such national, regional or international learning assessments provide suitable indicators of progress, for a program of "Cash on Delivery aid" at the primary level? Or, as some argue, should the international community prepare measurement instruments for this purpose (Annual Status of Education Report 2006, RTI International 2007). This paper will address this question.

2. Measuring learning progress

"If you want to measure change, don't change the measure". This observation, attributed to Otis Dudley Duncan in 1969 and applied to national assessments of learning by Albert Beaton in 1988, is central to any discussion of measuring progress. Measuring learning progress (that is, positive change) requires measurement instruments that are stable over time, in at least six ways:

- Testing the same cohorts (e.g. age cohort) for T_1 and T_2
- Measuring the same academic content or competencies at T_1 and T_2
- If sampling is used, using the same sampling procedures for T_1 and T_2
- Using measurement instruments having the same levels of difficulty at T_1 and T_2
- Using measurement instruments having the same reliability at T_1 and T_2
- Confirming the equivalence of the tests through empirical equating (Linn 2005)

In addition, to measure progress annually, a country must have the capacity to construct psychometrically valid and empirically equated tests, and to administer, score and report test results on an annual basis.

Achieving stability in learning assessments is generally carried out through the process of standardization. As Braun and Kanjii (2006: 317) note: "Standardization is a prerequisite for fairness when scores must be comparable. It demands, at a minimum, that the tests be administered under uniform conditions and graded according to a fixed set of rules or rubrics." Few developing countries can ensure that this is the case, and lack of experience leads to underestimation of the complexity of the processes for designing, developing, administering, scoring, analyzing and reporting the results of standardized tests (Braun and Kanjii 2006).

Much has been written about the difficulty of measuring change. Ragosa (1995) identified nine widely-held myths about measuring change, and concluded that such

measurement is possible, if individual growth curves are taken as the starting point. However, Braun and Bridgeman are skeptical, observing that models of individual growth

“assume that the psychometric meaning of test scores stays stable over time, which is a dubious proposition at best. Physical growth may be easily measured because exactly the same ruler can be used to measure the height of a person at 2 years old and then at 25 years old, but the same test could not be used to measure the reasoning or reading skills of the child and the adult.” (Braun and Bridgeman 2005: 4).

Rock, in discussing “the uncritical use of gain scores and their interpretation” and arguing for the use of adaptive testing, observes that:

“Gain scores that **are not** based on an adaptive testing approach are likely to give erroneous results because of floor and ceiling effects. Tests that are too hard (floor effects) or too easy (ceiling effects) will be unreliable in both tails of the score distribution and are likely to underestimate the amount of gains for those children in both tails” (Rock 2006:1).

In the United States, state-based accountability systems, in particular, have been questioned for their capacity to evaluate annual growth in student performance (Martineau 2006, Way 2006). One paper notes that volatility of test score measures “can wreak havoc in school accountability systems.” (Kane and Staiger 2002: 236 as quoted in Way 2006), although others debate this conclusion (Ragosa 2002).

Alternatives to individual growth monitoring include:

- status models (that describe the status of a school or educational system and report average scores or the share of students that have reached a particular performance standard rather than reporting change),
- successive groups models (that compare the performance of the same grade of students across consecutive years, e.g. grade 4 students in 2007 with the performance of grade 4 students in 2006; differences in cohort composition need to be taken into account), and
- longitudinal models (that follow a particular cohort as they age or progress through the grades).

Longitudinal models place a huge burden on education systems, which are plagued by three main technical limitations (Way 2006): (a) vertical scaling of tests⁴, (b) that all children be followed, and (c) an infrastructure and data capable of tracking students longitudinally. Student dropout and mobility and the lack of a suitable infrastructure lead to serious missing data problems. Because of difficulties in implementing longitudinal assessments, in the U.S., implementation of state-based accountability systems has

⁴ Vertical scaling simply means that a common scale exists against which student performance at increasingly older ages or higher grades can be measured.

generally relied on the “successive groups” approach for measuring year to year change, despite the technical limitations associated with measuring different cohorts of students (Marion et al 2002 as cited in Way 2006). Successive group models are also used for most national, regional and international learning assessments, and can provide information regarding differential learning outcomes among subgroups of students (such as gender, ethnicity, socio-economic status or geographical location). Successive group models require, at a minimum, that the tests be horizontally equated from one year to the next.⁵

What do we know about the tests used in learning assessments in developing countries? Are they aligned with respect to the six dimensions of stability noted above? How do the tests used in national, regional and international assessments differ? Regrettably, little public information is available about the technical characteristics of **national learning assessments**. Some information is provided in the Statistical Annex of the *2008 UNESCO Global Monitoring Report*, the single most comprehensive listing of countries with national learning assessments (UNESCO 2007). It presents information regarding 117 countries’ national learning assessments, including the target populations, the academic content/ competencies assessed and the regularity of the assessments. Detailed information about the tests used and the degree to which they are equated from administration to administration, however, is not provided in this report. Another recent report on testing in Latin America, specifically, is also relatively silent about the technical dimensions of the tests and notes that “Technical validation of test items or questions is a critical element in developing assessment instruments; unfortunately, national technical reporting on this subject is not particularly detailed” (Ferrer 2006: 28).

Table 1. Major regional and international learning assessments

Assessment (Sponsor)	Countries (most recent)	Target population	Content tested	Frequency	Years implemented
TIMSS (IEA)	37	Grade 4, 8	Math, Science	4 year cycle	1995, 1999, 2003, 2007
PIRLS (IEA)	17	Grade 4	Reading	5 year cycle	2001, 2006
PISA (OECD)	26	15-year olds	Math, Science, Reading	3 year cycle	2000, 2003, 2006
SACMEQ (IIEP and African ministers of education)	15	Grade 6	Math, Language	variable	1995-97, 2000-2002, 2007
LLECE (OREALC/UNESCO)	16	Grade 3,6	Math, Reading, Writing, Science*	10 year cycle	1997, 2007
PASEC (CONFEMEN)	22**	Grades 2,5	Math, French, National Language	variable	1993-95, 1997-2001, 2003-06

⁵ That is, although the test takers from one year to the next are the same age or grade, they are actually different individuals who take different versions of the test; these versions need to be empirically equated for the scores to have the same meaning from one year to the next.

*optional, **member countries; assessments in 1-4 countries per year; 9 countries have participated in one assessment, 6 countries have participated in 2 assessments and 2 countries have participated in 3 assessments; in all 17 of the 22 member countries have participated in an assessment.

The major **regional assessments** are Laboratoria Latinoamericano de Evaluacion de la Calidad de la Education (LLECE) in Latin America, the Southern African Consortium for the Measurement of Education Quality (SACMEQ) in Southern Africa, and Programme d'Analyse des Systemes Educatifs des Pays de la CONFEMEN (PASEC) in Francophone Africa. The major **international assessments** are the Trends in International Mathematics and Science Study (TIMSS), Progress in Reading Literacy Study (PIRLS) and Programme for International Student Assessment (PISA). Websites of regional and international learning assessments provide the most up-to-date information about these exercises (table 1). Such regional and international assessments utilize standardized tests with high degrees of reliability achieved through modern psychometric methods.

Criteria that National Learning Assessments often meet

Three of the criteria mentioned above appear to be met by national learning assessments: stability of target populations, stability of content, and administrative capacity. All are met by regional and international learning assessments. One other criterion, stability in sampling methods, is rarely applicable to national learning assessments, as most test entire grade cohorts.

Stability of target populations. Most countries target the same age or grade for testing, from one year to the next, within a narrow range of primary grades.⁶ All countries in Sub-Saharan Africa, East Asia and the Pacific, Latin America and the Caribbean and Central Asia that carried out any national learning assessment, 2000-2006, tested children in enrolled in at least one of grades 4-6 and many countries also tested children in grades 8 or 9 (Benevot 2007). Many countries also conduct annual assessments, but alternate among grades to be tested. For example, Chile tested 4th grade students in 2002, 2nd grade students in 2003 and 8th grade students in 2004 (Ferrer 2006). By comparison, regional and international assessments typically test students in the same grades or age cohorts consistently from one assessment to the next.

Content stability. Most countries with national learning assessments report stability in terms of test content. This means that the tests cover the same general curricular content areas, typically literacy and numeracy, from one assessment to the next. Over 90 percent of countries with national learning assessments test mathematics, reading and writing, and about 50 percent of countries test science, with some regional variations (Benavot and Tanner 2007). However, there is no guarantee that the tests cover exactly the same content from one year to the next, that test questions covering the same content are comparable with respect to difficulty from one year to the next, or that the curriculum includes “clear (or even operational) definitions of what students are expected to be able

⁶ Age cohorts are more comparable over time than are grade cohorts, since variations in grade cohorts can result from demographic changes in grade enrollments, and differences in repetition and dropout rates from one year to the next.

to do with the conceptual knowledge contained in the curricula” (Ferrer 2006:20). Thus, although the content may appear to be stable, it may also change dramatically. Change can come from the process of annual test development, the absence of curricula content and performance standards to guide test development, and curriculum reforms. Ferrer (2006) notes that Colombia, Ecuador and Uruguay have made efforts to specify curricular standards and link them with assessments, but these are exceptions in Latin America.⁷

Regional and international assessments also measure student learning in reading, writing mathematics and science, but the tests are typically constructed (and empirically equated) so as to be stable over time. TIMSS and PIRLS also are constructed to reflect the academic curricula of the participating countries. International reports from TIMSS, PIRLS and PISA already include sections that document change in achievement over time for participating countries.

Administrative capacity. Few countries have the administrative capacity to undertake annual assessments of students at the primary level. UNESCO’s 2008 EFA Global Monitoring Report identifies 20 developing or transition countries that report carrying out annual national learning assessments (table 2), and 10 others that report having implemented national learning assessments for two consecutive years over the period 2005-2007.⁸ These numbers may be overstated, however. In reporting out the countries in Western Europe and North America that have annual national learning assessments, UNESCO classifies United States’ National Assessment of Education Progress (NAEP) as an annual assessment, although it actually takes place biannually (NAEP website). Regional and international learning assessments typically take place at intervals of 3 or more years, reflecting the greater technical demands of such studies.

Table 2. Low and middle-income countries reporting yearly national learning assessments for a minimum of three consecutive years since 2000

Region	Name of Countries
Sub-Saharan Africa	Gambia, Seychelles
Arab states	Jordan
East Asia and the Pacific	Indonesia, Malaysia, Thailand
Latin America and the Caribbean	Anguilla, Bahamas, Belize, Brazil, Chile, Guyana, Jamaica, Mexico
Central and Eastern Europe and Central Asia	Estonia, Hungary, Mongolia, Poland, Turkey

Source: UNESCO 2007

Sampling. Most national learning assessments do not utilize scientific sampling methods for selecting schools or children for assessment. Rather, entire cohorts or populations of students are assessed, a practice that reduces technical complexity regarding sampling and weighting of results while increasing administrative complexity and cost. Among countries in Latin America, those that utilized sample-based assessments in the 1990s have generally shifted to the use of national censuses in the 2000s. Regional and

⁷ Jamaica, not reviewed by Ferrer, also has a long-standing effort in this regard.

⁸ Malawi, Madagascar, Uganda, Egypt, Mauritania, Myanmar, Pakistan, Philippines, El Salvador, Albania

international assessments, by comparison, utilize scientific samples, although the actual sampling is done by international experts rather than by experts within the country.

Criteria that National Learning Assessments rarely meet

Criteria that ensure stability of measurement instruments over time are rarely met in national learning assessments. Constructing instruments capable of measuring change over time is technically complex⁹. National learning assessments in developing or transition countries rarely employ such complex measurement instruments because such countries rarely have the requisite domestic capacity or can afford to purchase expertise from abroad.¹⁰ Even in countries where the capacity exists, the results of the assessment may not be able to be used effectively. For example, with respect to the National System of Basic Education Evaluation (SAEB) in Brazil, Wilcox and Ryder (2002:217, citing Crespo and others 2000) observe that “the SAEB, though a world-class performance measurement, has failed to fulfill its potential,” largely due to the absence of technically qualified personnel to interpret results, and little capacity to communicate the results to policy-makers and the general public. By comparison, recent regional and international learning assessments, which typically employ technical expertise from OECD countries, use measurement instruments that often are explicitly designed to measure change and provide interpretive materials aimed at the policy makers.

Three areas where national learning assessments are likely to fall short are in terms test stability in difficulty, reliability and comparability over time.

Stability of difficulty. Available studies of national learning assessments do not discuss technical issues of test construction, including how levels of difficulty are held constant year to year. However, some evidence from developing countries suggests that even relatively sophisticated national assessment units construct tests with “volatile” levels of difficulty. For example, average mathematics and language scores of grade 6 students in Jamaica showed remarkable variation, 1999-2004.¹¹ The main explanation for this variation was that the test was not stable with respect to difficulty, and the test developers had not designed the test in such a way that differences in difficulty could be empirically adjusted *post hoc* (Lockheed and others 2005). By comparison, all three major international assessments develop tests that are designed to use modern test theory and psychometric methods to ensure stability with respect to difficulty.

Reliability. National learning assessments rarely report information about the reliability of the tests, whereas regional and international assessments often include such information in technical manuals accompanying the assessment. Classical methods of test construction typically report internal consistency reliability (Cronbach’s alpha) or test-

⁹ *Educational Measurement, 4th Edition*, (Brennan 2006) discusses classical and modern techniques for establishing test stability over time. Topics include validity, reliability, item response theory (IRT), test bias, scaling, norming and equating.

¹⁰ There are exceptions; Qatar, for example, employs Educational Testing Service to conduct its national assessments of learning

¹¹ For a test of written communication graded on a scale of 1-6, average scores were significantly different across the four years: 5.0 in 2001, 3.7 in 2002, 4.5 in 2003 and 3.0 in 2004.

retest reliability, while modern approaches emphasize Item Response Theory (IRT) for establishing tests that measure the same constructs reliably. For example, PIRLS 2001 reports the classical Cronbach's alpha reliability scores for each country in the assessment, with a high median overall reliability of .88 (Mullis, Martin, Gonzalez and Kennedy 2003). Reliability is not reported for tests equated through IRT, which provides other, more sophisticated indicators of reliability.

Empirical equating. In order for tests to have the same meaning from one administration to the next, they must be equated (see Braun and Holland 1982 for a discussion of test equating) and tests must be designed in advance for this purpose, using calibrated items. Applying the same exact test from one time to the next does not guarantee their equivalence, as individual questions may change in their degree of difficulty from one year to the next.¹² Psychometric professional consensus is that equating can occur only when tests measure the same constructs and have the same reliability and when the equating process is symmetrical, equitable and population invariant (Linn 2005). Because equating requires the application of complex psychometric and statistical techniques, this is the area in which most national learning assessments show greatest weaknesses, and where IRT is generally applied in regional and international assessments, such as SACMEQ, PIRLS, PISA and TIMSS. Regional and international learning assessments expose participants to many of the technical issues for ensuring the stability of measurement instruments over time. In addition, specific programs for building national assessment capacity have been established by donors, such as the World Bank, and international testing bodies, such as Educational Testing Service and the International Association for the Evaluation of Educational Achievement (IEA).

But building the capacity to develop valid and reliable measures of student learning achievement faces significant obstacles in developing countries. Among these are: (a) the technology and psychometrics of test development are evolving continuously, requiring continuous professional development for test developers, (b) test development software, including item banking software, is not supported after newer versions are available, (c) the lack of ongoing doctoral programs in psychometrics in developing countries means that specialists are often sent overseas for training, but receive little support when they return home, and (d) the recurrent costs for training and upgrading test development staff are often unsustainable. Moreover, as the complexity of testing and assessment has grown – with greater numbers of domains, higher levels of performance, and more variety in performance measures -- all the processes required to construct, score and equate tests become more difficult and more expensive, and less within the capacity of developing countries to achieve.

To summarize, the minimum requirements for monitoring change over time are rarely satisfied by existing national learning assessments in developing countries. A few middle-income countries have the technical and administrative capacity to measure learning progress on an annual basis. Specifically, 11 countries, indicated in bold in Table 3, have been exposed to the technical skills needed to build stable tests, through

¹² For example, a question regarding the distance between the earth and the moon was difficult before any astronauts had been to the moon, and easy immediately after the first moon landing.

participation in a recent regional or international assessment study, and also have demonstrated their administrative capacity to administer, score and report the results of tests on an annual basis, through having done so for a minimum of three consecutive years. These countries are: Brazil, Chile, Estonia, Hungary, Indonesia, Jordan, Malaysia, Mexico, Poland, Thailand and Turkey. Egypt and El Salvador could also be included, as they have administered national learning assessments for two consecutive years during 2005-2007 and have participated in a recent international assessment. Two countries, Brazil and Chile, are also recognized for the high quality of their national learning assessment.

Table 3. Low and middle-income countries participating in recent international learning assessments

Region	PIRLS 2006	PISA 2006	TIMSS 2007
Sub-Saharan Africa	South Africa		Botswana, Ghana, South Africa
Arab states	Morocco	Jordan , Tunisia	Algeria, Egypt, Jordan , Morocco, Oman, Syria, Tunisia, Yemen
East Asia and the Pacific	Indonesia	Indonesia, Thailand	Indonesia, Malaysia, Thailand
Latin America and the Caribbean	Trinidad and Tobago	Argentina, Brazil, Chile , Colombia, Mexico , Uruguay	Colombia, El Salvador, Trinidad and Tobago
Central and Eastern Europe and Central Asia	Bulgaria, Georgia, Hungary , Iran, Latvia, Lithuania, R. of Macedonia, Moldova, Poland, Romania, Russian F. Slovak R. Slovenia	Azerbaijan, Bulgaria, Croatia, Czech R., Estonia, Hungary , Kyrgyz R., Latvia, Lithuania, Montenegro, Poland, Serbia, Slovak R., Romania, Russia F. Slovenia, Turkey	Armenia, Bosnia, Bulgaria, Czech Rep, Georgia, Hungary , Iran, Kazakhstan, Latvia, Lithuania, R. of Macedonia, Moldova, Poland , Romania, Russian F., Serbia, Slovak R. Slovenia, Turkey , Ukraine

Source: IEA 2007, OECD 2007

3. A discussion of some terminology

National learning assessments typically utilize tests that purport to measure what has been learned. I now digress with a brief discussion of terminology related to national (or regional or international) learning assessment, beginning with the term “**test.**” Any discussion of “tests” must take into account that the term covers a wide range of assessment instruments, used for a variety of purposes. Webster’s defines a test as “any series of questions or exercises or other means of measuring the skill, knowledge, intelligence, capacities, or aptitudes of an individual or group.” Anderson and others (1986: 425) observe that “common elements seem to be (a) an experience that is reproducible across two or more people or groups and (b) some means of characterizing individuals or groups in comparable terms on the basis of that experience.”

These simple definitions hide the complexity of tests, which can be described in a great variety of ways, among them the 16 dimensions outlined in table 4. Other analysts

provide additional dimensions (see *Educational Measurement*, 4th Edition, Brennan 2006). National, regional or international assessments typically use a common constellation of test characteristics (table 4, column 3).

Table 4. Dimensions of tests as applied to national, regional or international assessments

Dimension	Example	Application to assessments
Proposed use	Individual diagnosis, selection, certification, program evaluation	Evaluation
About whom test-based decisions are to be made	Individuals or groups	None (group implications)
What construct is to be measured	Personality, aptitude, mental abilities, interests, skills	Mental abilities and skills
What subject matter or content is to be measured	Mathematics, reading, art	Mathematics, reading
Whether the focus is on maximal or typical performance		Maximal
How heterogeneous are the test questions or exercises in terms of constructs or subjects	Batteries of tests, single subject tests	Batteries of tests
How the score or performance is to be interpreted	Subjective standards, criterion referenced, norms	Criterion referenced, norm referenced
Type of response the student is to provide	Performance (an essay, a drawing, a recital), recognition (multiple choice, true-false, item matching)	Performance and recognition
How the student's response is scored	Objective vs. subjective, "machine" vs. hand, quantitative vs. qualitative	Objective and subjective
Whether there are standards for the acceptability of the response	Correct or "best" answers, scales or agreement	Correct, best answers
Whether the student's and the tester's perceptions of the tests are congruent	Usually for clinical purposes only	Congruent
When the test is administered	Annually, periodically, before or after an instructional program	Annually or periodically
Emphasis on the speed of response	Speeded with shorter time limits vs. power with longer time limits	Power
To whom the test is administered	Individuals or groups	Groups
Who constructs the test	Teacher made vs. professional	Professional

Source: adapted from Anderson and others 1986

Standardized tests are tests that are "administered under uniform conditions and graded according to a fixed set of rules or rubrics"; they are not simply tests that use multiple-choice test items (or questions). They are tests whose results are not contingent on the time or location of the test or on the scoring of the results. Standardized tests are often administered with multiple choice (recognition) formats, to ensure consistent scoring as well as lower costs. Detailed scoring guides, or rubrics, combined with improvements in scanning technology has meant that consistent scoring can be achieved for constructed response items (i.e. open-ended questions), allowing standardized tests to include a higher share of such items, although the cost of professional scoring of constructed response items is many times greater than that for scoring multiple choice items. Major

international assessments use both types of items in their standardized tests. For example, about one-third of TIMSS 2003 items and two-thirds of PISA 2003 items were constructed-response items (Hutchison and Schagen 2007).

Standardized tests can be referenced to norms, to criteria or to both. **Norm referencing**, in its original sense, refers to a process whereby a random sample of individuals is drawn from the reference population and these individuals are tested. If the test is well constructed, scores will be distributed according to a normal curve. Results from subsequent administrations of the test to different individuals or groups will be compared with the “norm reference group.” For example, infant growth curves are norm-referenced with respect to the height and weight progress of a norm-reference sample of infants. More recently, with the advent of large-scale testing, the term norm-referencing is used when an individual score is compared with the distribution of all scores (such scores are often referred to as percentile scores). Norm-referenced tests are constructed to spread questions across a broad range of difficulty. **Criterion referencing** is completely different, as it refers to categories of performance that are ordered, such as hurdles of different heights; individuals either meet or do not meet the criterion at different levels. Much debate can surround decisions regarding the cut-off points for various criteria, and criteria-referenced tests are constructed to have greater discrimination around cut-off points. A single test can provide both “norm referenced” and “criterion-referenced” information, although there is debate about this practice.¹³

Standards based assessment is related to criterion referencing, and pertains to a test that seeks to assess performance relative to a set of pre-established performance standards or criteria (Tognolini and Stanley 2007). Performance standards are intrinsic to criterion referenced tests (Berk 1986, Hambleton and Plake 1995). Standards based assessments (or tests) comprise items that sample from a broad range of performance, but even meaningful standards-based assessments “cannot represent the depth and breadth of skills reflected in the standards documents for any one domain or grade level” (Rupp and Lesaux 2006: 316-7). Establishing standards can require decades of deliberation, and building standards-based assessments requires considerable technical skill on the part of test developers:

“The challenge for test developers is that standards-based assessment must be broad enough to address in sufficient detail the complex aspects of proficiency that are laid out in the standards – those that require complex reasoning and problem-solving skills— while still addressing basic knowledge and skills” (Rupp and Lesaux 2006:317).

The US experience with standards-based assessment shows that different jurisdictions can have very different definitions of standards, with “high standards” in one jurisdiction barely meeting “average standards” in another (Fuller and Wright 2007). A New York Times editorial observes that “many states have gamed the system – and misled voters –

¹³ Consider household income. A household can be described as being in the top quintile of the distribution (norm referencing) or as below the poverty line (criterion referencing).

devising weak tests, setting low passing scores or changing tests from year to year to prevent accurate comparisons over time” (New York Times 2007).

Some discussions of assessment distinguish between **“competency” assessments** and **“curriculum” assessments**. These distinctions, however, suggest greater differences than have been empirically verified. A well-designed school learning program will address both curriculum content and cognitive demands, and good tests will include questions that assess both.

For example, consider the content of two international mathematics tests (in TIMSS 2003 and PISA 2003) that are widely believed to be very different with respect to what they measure: TIMSS2003 measures “curriculum” and PISA 2003 measures “competency.” Comparing the actual content of the test questions on the two tests, however, Hutchinson and Schagen (2007) find that both tests include questions that measure competencies in mathematics (using concepts, reasoning) and both tests include questions that cover the curriculum (knowing facts and solving routine problems). About a quarter of questions on both tests measure reasoning competency, while TIMSS 2003 has a higher share of questions that measure knowing facts, and PISA 2003 has a higher share of questions that measure using concepts. Moreover, scores on TIMSS 2003 can be summarized into levels of student performance ranging from advanced (students can organize information, make generalizations, solve non-routine problems, and draw and justify conclusions from data) to low (students have some basic mathematical knowledge), with such benchmarks indicating levels of competencies (Mullis and Martin 2007).

Another term is **“high stakes”** testing. An important feature of sample-based learning assessments is that they are not used to make decisions about either individuals or groups (although they do provide information that can inform decisions about groups). Because students (and teachers and schools) have no direct incentives to perform well on these tests, test performance is not influenced by potentially distorting incentives. Sample-based learning assessments are typically considered “low stakes” tests, in sharp contrast with “high stakes” tests that are used for selection or certification purposes, and – in the United States--for accountability under the No Child Left Behind (NCLB) legislation. In a seminal paper, “Will national tests improve student learning?” first presented at the 1991 Annual Meeting of the American Educational Research Association and later published, Laurie Sheppard delineated six reasons why “high stakes” tests may fail to reform education. She argued that high stakes tests: (a) become inflated without actual improvement in learning, (b) narrow the curriculum, (c) misdirect instruction even for basic skills, (d) deny students opportunities to develop thinking and problem-solving skills, (e) result in hard-to-teach children being excluded from the system, and (f) reduce professional knowledge and the status of teachers. In addition, high stakes tests are subject to distortions due to cheating by students, teachers and schools. Testing agencies in both developed and developing countries have gone to great lengths to prevent such from occurring and to identify situations in which they have occurred.

Many developing countries have well-established institutions for constructing, administering, scoring and reporting the results from “high stakes” selection and

certification tests (often referred to as “**examinations**”), and many of the national learning assessments reported by UNESCO are, in fact, examinations. Such tests are not used for purposes of educational accountability, and are “high stakes” only for the students taking the tests. Regional bodies, such as the Caribbean Examinations Council and the West African Examinations Council, conduct or provide support for national examinations. Selection examinations were once commonly administered at the end of primary education (Primary School Leaving Examinations) to regulate the flow of students into junior secondary education. Examinations at this level, however, have typically disappeared since junior secondary education has been incorporated into EFA.

Selection examinations are widely administered at the end of secondary education to regulate the flow into tertiary (higher) education, and some countries still have examinations at the end of junior secondary education, which are used for certifying completion of schooling at this level and for purposes of “guidance” – that is, determining which course of study the student should pursue at the secondary level. Selection examinations often serve dual purposes, both for screening and for certifying successful completion of the level of schooling. In general, when two different tests are used, selection examinations contain a higher share of difficult test questions than do certification examinations, since the purpose of the selection examination is to limit access.

4. Using testing for Cash on Delivery aid

The above discussion is intended to draw attention to the administrative and technical complexities surrounding the application of tests intended to measure progress in learning. We have not yet discussed the costs associated with testing, and will do so in this section. What are some of the implications of this discussion for using test results for Cash on Delivery aid, and what might some alternatives entail?

Costs of testing

Testing comprises a relatively small share of total education expenditures. For example, Ilon and Harris (1992) used the “ingredients” method (Levin and McEwan 2004) to estimate costs associated with test development, registration, production, administration, scoring and reporting for a sample of 20,000 test takers in Jamaica in 1992. The total cost amounted to US\$196,250 (\$9.80 per test taker) which was less than 7 percent of Jamaica’s average annual expenditure on education, 1993-96.¹⁴ More recent data on costs suggest that participation in international or regional assessments, such as TIMSS or PIRLS, is also relatively inexpensive. The annual fee for countries participating in the international activities that support capacity development in assessment was \$30,000 per grade assessed, or \$120,000 for the four-year testing cycle, about US\$20-25 per test

¹⁴ Education expenditure as a share of GNP, 1993-1996 = 7.5%; GNP per capita in 1995 = \$1,510; population in 1995 = 2.5 million.

taker.¹⁵ At the same time, there is ample evidence that testing units within governments are only weakly funded and that only strong political incentives – such as, for example, the incentive that EU or OECD membership provides governments for participating in PISA – shake loose adequate funds. Underfunding is particularly widespread in low-income countries, including those with long histories of participation in regional examinations councils, such as the West African Examinations Council or the Caribbean Examinations Council.

Implications

The major implications of the above are that (a) cost is not a determining impediment for using testing in Cash on Delivery aid, but that (b) the technical and administrative requirements for using a valid and reliable measure of learning progress pose significant impediments for testing in developing countries. Political will is also a constraint, but even with strong political will, capacity issues remain. While the capacity for carrying out national learning assessments has improved over the past two decades in developing countries, it remains fragile in most low-income countries. Even in the US, where the NCLB has created a strong incentive for regular learning assessment, the assessment capacity for meeting NCLB requirements has been severely stretched. In addition, three other factors argue against measuring progress directly with tests: the volatility of change scores, the risk of non-participation and “test pollution.”

Volatility in scores leads to incorrect rewards. National learning assessments utilize, for the most part, measurement instruments that are poorly suited for registering change, and are consequently volatile over time. Such volatility means that, on an annual basis, countries (or schools) could be rewarded or penalized erroneously. While a few, middle income, countries may have the technical capacity for developing, administering, scoring and reporting the results from valid measures of student learning on an annual basis, these countries are not the main focus of the donor community’s concern with respect to education. Countries for whom Cash on Delivery aid is most salient lack the requisite technical and scores will not reflect actual progress.

A risk of public embarrassment when scores are released can threaten participation. Countries have been embarrassed by their performance on international learning assessments, and this embarrassment has reduced participation and led to restrictions in publication and dissemination of results (for example, Mexico declined to release its scores on TIMSS 1999). The publicity that could surround “payments for progress” could also be embarrassing for some education authorities. The challenge would be to publicly disclose results without specifically identifying or penalizing schools, which may not be realistic. Although schools often report the results of “high stakes” tests – e.g. how many of their students passed an examination or qualified for a merit scholarship – the “high stakes” are for the students rather than the schools. While a school might perceive strong incentives in having many students perform well, resources flowing to the school generally are not contingent on the students’ performance.

¹⁵ Based on World Bank Development Grant Facility support to the IEA for 17 developing countries participating in TIMSS2007 and an average sample size of 5,000 students per country.

Public disclosure of test scores does not necessitate embarrassment, however, and can help target resources where they can be effective. In Chile, for example, a program provided the lowest performing 10 percent of schools, as identified from test scores, with a package of school inputs, including textbooks, in-service teacher training and tutoring for low-achieving children; between 1997 and 2000 the test score gap between indigenous and non-indigenous students had dropped nearly 30 percent due to these school reforms (McEwan 2006).

“Test pollution” when low stakes tests become high stakes tests. Linking test performance to budgetary support to education would, inevitably, make any test that is used to monitor change into a “high stakes” test, at least for school administrators. Regional and international learning assessments are, by their very nature, low stakes tests. But high stakes tests are subject to numerous distortions, which Pearson and others (2001) refer to as “test score pollution,” beginning with those associated with cheating,

Countries with weak testing capacity are simply not able to guarantee that the results of high stakes tests will not be distorted through cheating. Cheating occurs at all levels, and parents, students, teachers and schools all have been found to cheat on high stakes tests. Some countries apply heroic measures to avoid cheating, with astronomical costs. For example, to reduce cheating on a college admissions test, the National Assessment and Evaluation Center in Georgia established secure testing facilities and purchased video-cameras and monitors to observe both test takers and the environment surrounding the test facility (Maia Miminoshvili personal communication May 18, 2005). In other countries, cheating is routinely detected through statistical methods, and the cost is also substantial.

Cheating can occur at all stages of testing. Tests or test questions can be stolen and released to prospective test takers. Test takers can hire others to take their place for testing or use cell phones to communicate information about the test questions. Teachers can coach students or correct their answers after the test administration. Administrators can invite “low performing” students to skip school on the testing day. The list is endless, and cheating is caused by having so much ride on the results. As Pearson and others (2001:177) remark: “Given what we know about test score pollution, we are forced to believe that nothing is fair or objective and to trust no one.”

Alternatives to using national learning assessments in Cash on Delivery aid

Given the technical complexities involved in measuring change in learning, the lack of the necessary measurement and implementation capacity in most developing countries, and the inevitable “test pollution” arising from high-stakes testing, alternatives to using the results of national learning assessments for monitoring progress could be considered. The most promising approach would be for multiple indicators of progress to be adopted into a “country report card” in which test results would be only one of several indicators of education quality. Some alternatives are discussed in the following paragraphs.

Use results from assessments implemented by testing experts. One alternative would be for countries participating in Cash on Delivery aid programs to contract international measurement bodies to carry out their annual learning assessments, which could be based on nationally developed standards. Two advantages of this option are that technical capacity in measurement is no longer an issue and that the independence of the assessment could be ensured. Three disadvantages are that this option might generate little ownership of the assessment by the countries, would not necessarily help national policy makers understand the test results, and would entail significant costs. But donors could finance the costs, and there is precedent. The World Bank's Development Grant Facility and UNESCO have provided support to countries participating in IEA's international learning assessments (TIMSS and PIRLS), and USAID has supported countries participating in RTI's Early Grade Reading Assessment studies.

Use donor-developed tests of competency in early grades. A related alternative would be for international donors to develop tests of reading and math competencies in the early grades, which tests could be used to measure the results of Cash on Delivery aid. Any such tests, of course, would need to meet the key criteria listed above. Such tests have been developed, with reasonable evidence of internal consistency reliability. The earliest examples are tests of English and mathematics developed by Educational Testing Service and used in a study of educational achievement in Kenya and Tanzania in 1980 (Knight and Sabot 1990). These tests were subsequently adapted and used in Ghana, along with a test of local language developed by the University of Ghana in association with the University College of Education at Winneba, to evaluate World Bank lending for education in Ghana (Operations Evaluation Department 2004).¹⁶ During the 1990s, curriculum-based, tests were developed for use in a longitudinal evaluation of USAID supported projects in Malawi and Ghana; these tests were roughly equated, and data on the development of the tests are available (Dowd and Harris 2008, Calahan and Harris 2008). In India, an NGO called Pratham developed a reading test to assess literacy, but their reports provide little information about the quality of the test, including such basics as measures of dispersion or reliability (Annual Status of Education Report 2006).

Currently, a beginning reading test (Early Grade Reading Assessment, or EGRA) is under development by RTI International, with funding from USAID, the World Bank and various NGOs. EGRA has been applied in pilot versions for various languages in Afghanistan, Bangladesh, the Gambia, Haiti, Kenya, Mali, Nicaragua, Niger, and Senegal, with results reported for a few countries; estimates of internal consistency reliability are available for Kenya¹⁷ and Senegal (RTI International 2007, Springer-Charolles 2007). The developers of the EGRA, however, are quite explicit that their test is not suitable for accountability purposes, but is intended only for diagnostic use by and for teachers in early grades (Grove 2007).

¹⁶ All tests had reasonable internal consistency reliabilities as measured by Cronbach's alpha, although the reliabilities of tests developed by ETS and adapted for use in Ghana (.72 - .79 for English and .75 - .82 for mathematics) were slightly higher than those developed for local languages in Ghana (.64 and .70).

¹⁷ Cronbach's alpha of .72 - .89 for English and .88 - .92 for English and Kiswahili combined.

The advantages and disadvantages of such tests are similar to those associated with using the results of test implemented by international measurement bodies.

Use participation in regional or international assessments. A third alternative would be for these countries' participation in ongoing regional or international assessments to count as "progress." Collaborating with international measurement experts would enable "testing units" in developing countries to strengthen their capacity by forming ongoing relationships with professional colleagues while providing "low stakes" measures capable of tracking change over longer periods of time (3-5 years, for example). Performance would be measured by progress in implementing the technical steps associated with the regional or international assessment, supported by external technical assistance, rather than by annual testing of students. This approach has been successful in bringing more developing countries into such regional and international assessments, and could be expanded to low-income countries, in particular.

Publish national assessment data disaggregated by key indicators. A fourth alternative would focus on the gaps in performance among various population subgroups in the country, rather than on the actual level of performance. While this approach also involves technical complexities (related to weighting of results, for example) and tests would still need to meet the technical criteria listed in section 2, it could generate national discussion about the quality of education for "disadvantaged" groups. Care would need to be taken, however, to avoid reinforcing gender, ethnic, socio-economic or regional stereotypes regarding student abilities. Some countries might resist publication of disaggregated student test scores, from such a concern. For example, Malaysia has not published the results of national or international assessments, disaggregated by ethnicity (Chinese, Malay, Tamil) for this reason, and few countries publish results disaggregated by and indicator of "language spoken at home."

Measure progress in terms of the development of concrete curriculum standards. Most developing countries lack concrete curriculum standards, and many teachers in developing countries do not have a clear vision of what performance is to be expected from their students (Ferrer 2006). An important first step for improving the quality of education in developing countries could be initiating a process for developing and putting into place curriculum standards and their associated measures of performance. Progress could be indicated by reaching certain specific objectives, such as: (a) defining specific language and mathematics skills, (b) presenting them to the educational community and general public for discussion, as Colombia did in 2003 (Ferrer 2006), (c) establishing learning progress maps for students, as in Chile, (d) providing data on test specifications related to the skill objectives for review by technical experts, and (e) carrying out an assessment based on such learning standards.

Adopt non-testing alternative indicators of progress. Another alternative would be for Cash on Delivery aid to select an indicator of progress that does not involve the technical and administrative considerations surrounding testing. Such indicators include: (a) community monitoring of teacher attendance, (b) random checks on average daily student attendance, and (c) student placements at next higher level of education. As these

measures would also become “high stakes” care would need to be taken in collecting and reporting the relevant data.

The advantages of community monitoring of teacher attendance are two-fold: it involves the community in the school and is relatively simple to administer. This is also the case for random checks on average daily student attendance. Monitoring student progress in terms of placement at the next higher level of education has significant disadvantages insofar as access to the next higher level of education is frequently determined by supply factors that are not quickly amenable to change; in particular, since upper secondary education typically requires teachers having higher levels of skills and qualifications, it is difficult to expand secondary education without prior investment in tertiary education and teacher training, with considerable lag times before the results of these investments are visible.

Use multiple indicators of progress. Perhaps the most effective alternative would be a mix of indicators that combine testing and assessment with other, observable, indicators that are uncomplicated by the technical dimensions of testing. Tests are an important part of this package of indicators, since they can help focus attention on what learning objectives are important and what signals student accomplishment in attaining these objectives. But relying on tests alone places too much importance on an indicator that is easily “corruptible.”

In Short

Existing national learning assessments are poorly suited for measuring annual educational progress in developing countries, for both technical and administrative reasons. These tests could, however, be used in combination with other indicators of education quality to measure educational status and to identify within-country variations in student learning. “Test pollution” is an issue that would need to be addressed, as performance incentives can change low stakes tests into high stakes tests. Currently available technical evidence regarding tests of early reading is insufficient to judge whether such tests offer a high quality alternative to national assessments. Regional and international assessments offer high quality alternatives, but are poorly suited for monitoring annual improvements. Countries could purchase technical expertise, but the costs may be too high to consider.

Alternative indicators of progress suffer from fewer technical difficulties, but are also possibly subject to distortions caused by becoming “high stakes” indicators. Progress in implementing regional or international assessments, including independent sampling, data collection and analysis, could be an essential element of performance-based aid for education. Progress in establishing clear and measurable performance standards could be a measure of progress toward quality. Country “report cards” that include multiple indicators of progress – including test scores – may be more appropriate than a single indicator.

References

Amrein, A. L. and D. C. Berliner. 2002. "High-Stakes Testing, Uncertainty, and Student Learning." *Education Policy Analysis Archives* 10 (18). Retrieved December 2, 2007 from <http://epaa.asu.edu/epaa/v10no18>.

Annual Status of Education Report. 2006. New Delhi. www.Indiatogether.org.

Berk, R. 1986. "A consumer's guide to setting performance standards on criterion referenced tests." *Review of Educational Research*, 56, 137-172.

Braun, H. 2004. "Reconsidering the impact of high-stakes testing." *Educational Policy Archives*, 12 (1). Retrieved December 2, 2007 from <http://epaa.asu.edu/epaa/v12n1/>.

Braun, H. and Bridgeman, B. 2005. An introduction of the measurement of change problem. RM-05-01. Princeton, NJ. Educational Testing Service

Braun, H. and A. Kanjii. 2006. "Using assessment to improve education in developing nations." In J. Cohen, D. Bloom and M. Malin (Eds.) *Educating All Children: A Global Agenda*. Cambridge, MA: American Academy of Arts and Sciences.

Brennan, R. (Ed.) 2006. *Educational Measurement (4th Edition)*. Westport, CT: Praeger.

Cahalan, C. and A. Harris. 2008. "Passage difficulty and reading performance in Ghana." Paper to be presented at the Annual Meeting of the Comparative and International Education Society, New York, March 17-21.

Carnoy, M. and S. Loeb. 2002. "Does external accountability affect student outcomes? A cross-state analysis." *Educational Evaluation and Policy Analysis* 24 (4), 305-331

Dowd, A. and A. Harris. 2008. "Measuring reading proficiency in Malawi, 1999-2006" Paper to be presented at the Annual Meeting of the Comparative and International Education Society, New York, March 17-21.

Ferrer, G. 2006. *Educational Assessment Systems in Latin America: Current Practice and Future Challenges*. Washington, DC: PREAL

Fuller, B. and J. Wright. 2007. "Diminishing returns? Gauging the achievement effects of centralized school accountability." Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April 11.

Goertz, M. and M. Duffy. 2003. "Mapping the landscape of high-stakes testing and accountability programs." *Theory into Practice* 42 (1), 4-11.

Greene, J.P, M. A. Winters, and G. Forster. 2003. "Testing high stakes tests: Can we believe the results of accountability tests?" *Civic Report 33* (February). Retrieved December 2, 2007 from http://www.manhattan-institute.org/html/cr_33.htm.

Grove, A. 2007. RTI website article on Early Grade Reading Assessment.

Hambleton, R. and Plake, B. 1995. "Using an extended Angoff procedure to set standards on complex performance assessments." *Applied Measurement in Education*, 8 (1), 41-55

Hutchison, D. and I. Schagen. 2007. "Comparisons between PISA and TIMSS—Are We the Man with Two Watches?" In T. Loveless (Ed.) *Lessons Learned: What International Assessments Tell Us about Math Achievement*. Washington, DC: The Brookings Institution.

Ilon, L. and A. Harris. 1992. "Economic analysis of testing system for Jamaica." Paper prepared for World Bank.

Ilon, L. 1992. "A framework for costing tests in Third world settings." Population and Human Resources Department Discussion Paper PHREE-92-65. Washington, DC: The World Bank.

Kane, T. J. and D. O. Staiger. 2002. "Volatility in school test scores: Implications for test-based accountability systems." In D. Ravich (Ed.) *Brookings papers on education policy*, Washington, DC: Brookings Institution. (pp 235-260).

Koski, W. S. and H. A. Weis. 2004. "What educational resources do students need to meet California's educational content standards? A textual analysis of California's educational content standards and their implications for basic educational conditions and resources." *Teachers College Record 106* (10), 1907-1935

Lockheed, M., A. Harris, P. Gammill, K. Barrow, T. Jayasundera. 2006. "New Horizons for Primary Schools in Jamaica: Inputs, Outcomes and Impact" Washington, DC: Academy for Educational Development

Lockheed, M., A. Harris, P. Gammill, and K. Barrow. 2006. "Impact of New Horizons for Primary Schools on Literacy and Numeracy in Jamaica: Inputs." Washington, DC: Academy for Educational Development

Martineau, J. A. 2006. "Distorting value-added: The use of longitudinal vertically scaled student achievement data for growth-based, value-added accountability." *Journal of Educational and Behavioral Statistics*. 31 (1), 35-62

McEwan, P. 2006. "The fortuitous decline of ethnic inequality in Chilean schools." Wellesley College, Wellesley, MA.

Mullis, I. and M. Martin. 2007. "TIMSS in perspective: Lessons learned from IEA's four decades of international mathematics assessments." In. T. Loveless (Ed.) *Lessons Learned: What International Assessments Tell Us about Math Achievement*. Washington, DC: The Brookings Institution.

Murphy, P., V. Greaney, M. Lockheed and C. Rojas (Eds.). 1996. *National Assessments: Testing the System*. Washington, DC: The World Bank Economic Development Institute.

New York Times. 2007 (November 26). "Test and Switch". Editorial.

Operations Evaluation Department. 2004. *Books, buildings and learning outcomes: An impact evaluation of World Bank support to basic education in Ghana*. Washington, DC: The World Bank.

Pearson, P. D., S. Vyas, L. Sensale and Y. Kim. 2001. "Making our way through the assessment and accountability maze: Where do we go now?" *The Clearing House* 74 (4), 175-182

Porter, A. C., M. Chester, and M. Schlesinger. 2004. "Framework for an effective assessment and accountability program: The Philadelphia example." *Teachers College Record* 106 (6), 1358-1400

Ragosa, D. 1995. Myths and methods: "The myths about longitudinal research" plus supplemental questions. In J. M. Gotman (Ed.) *The analysis of change*. Mahwah, NJ: Erlbaum.

Ragosa, D. 2002, October. Irrelevance of reliability coefficients to accountability systems: Statistical disconnect in Kane-Staiger "Volatility in School test scores."

Resnick, L.B. 2006. "Making accountability really count." *Educational Measurement: Issues and Practice*. Spring, 33-37

Riffert, F. 2005. "The use and misuse of standardized testing: A Whiteheadian point of view." *Interchange* 36 (1-2), 231-252

Rock, D. 2006 "Some thoughts on gain scores and their interpretation in developmental models designed to measure change in the early school years." Princeton, NJ: Educational Testing Service Center for Global Assessment. Draft August 14, 2007.

RTI International. 2007. "Early Grade Reading Kenya Baseline Assessment: analyses and Implications for Teaching Interventions Design." Draft for discussion, August 20, 2007.

Sprenger-Charolles, L. 2007. "Senegalese and Gambian Early Grade Reading Assessments (EGRA): Analyses of the results and suggestions for future EGRA applications." Paris: CNRS and Descartes University.

Tognolini, J. and G. Stanley. 2007. "Standards-based assessment: a tool and means to the development of human capital and capacity building in education." *Australian Journal of Education*, 51 (2), 129-145.

Way, W. 2006. Precision and volatility in school accountability systems. RR-06-26. Princeton, NJ: Educational Testing Service.

Wilcox, K. C. and R. Ryder. 2002. "Standardized testing and improving educational opportunity in Brazil." *The Educational Forum* 66 (spring), 214-219

Yeh, S. 2007. "The cost-effectiveness of five policies for improving student achievement" *American Journal of Evaluation* 28 (4), 416-436.