Debating International Learner Assessments as a Proxy Measure of Quality of Education in the Context of EFA-A Review Essay

Godfrey Mulongo[1,*]

[1]Regional Monitoring and Evaluation Specialist at The International Potato Center (CIP) and Ph.D candidate, Witwatersrand University, South Africa

*Correspondence: Box 34424, Manara Road Plot 10, Dar-es-salaam, Tanzania. E-mail: mulongoe@gmail.com

Received: September 28, 2013 Accepted: December 12, 2013 Online Published: February 11, 2014

doi:10.5430/wje.v4n1p35 URL: http://dx.doi.org/10.5430/wje.v4n1p35

Abstract

This review essay looks at three publications that discuss the contentious issue of evaluating education quality (Note 1) by learner outcomes as a proxy indicator (Note 2). The essay explores the debates, gaps and proposes recommendations in the context of Education For All (EFA) (Note 3). The three articles reviewed are Harvey Goldstein's (2004) *"Education For All: the globalization of learning targets"*, Angeline Barret's (2009), *"The education Millennium Development Goal beyond 2015: Prospects for quality and learners"* and Daniel Wagner's. et.al. (2012) article on *"the debate on learning assessments in developing countries"*. Goldstein and Barret's articles argue against the adherence to numerical learner achievement targets and explore possible consequences of doing so, while Wagner et.al. (2012) articles argue in support of same.

Keywords: *assessment, learner outcomes, quality, indicator*

1. Background

The joint UNESCO/UNICEF project on Monitoring Education-For-All Goals with focus on Learning Achievement (Note 4) begun in September 1992. This was an immediate outcome of The World Declaration on Education-For-All adopted at Jomtien in March 1990 which pointed to the need *"to define acceptable levels of learning acquisition for educational programmed and to improve and apply systems of assessing learning achievement"*. The understanding was that merely improving the supply of education -- quantity -- was not enough, and an improvement in quality was considered vital, so was the means to assess progress made on this front (UNESCO/UNICEF, 1994). However, international learner assessments did not start with the UNESCO/UNICEF project, courtesy of EFA (Note 5). According to Larry Suter (Note 6), education researchers and policymakers from twelve countries first established a plan for making large-scale cross-national comparisons between countries on student performance way back in 1958 at the UNESCO Institute for Education in Hamburg, Germany. This led to the first successful large-scale quantitative international study in mathematics conducted in 1965 by the International Association for the Evaluation of Educational Achievement (IEA) and included Australia, Belgium, England, Finland, France, Germany, Israel, Japan, Netherlands, Scotland, Sweden, and the United States.

Between 1965 and 2001 the IEA sponsored studies of mathematics in 1965, 1982, 1995, and 1999; science in 1970, 1986, 1995, and 1999; reading in 1970, 1991, and 2001; civics in 1970 and 1998; and technology in 1990 and 1999. The Educational Testing Service conducted an International Assessment for Education Progress in science and mathematics in 1990. The first international learner assessment under the EFA framework begun in 1992. Today, the main international assessment frameworks include the International Association for the Evaluation of Education Achievement, or IEA (e.g. TIMSS and PIRLS), the OECD (e.g. PISA), the Laboratoria Latinoamericano de Evaluacion de la Calidad de la Education or LLECE (e.g. TERCE), and the Southern and Eastern Africa Consortium for Monitoring Education Quality (SACMEQ) among others.

It may therefore seem late in the day, especially now that the EFA vision of 2015 is drawing nigh, to re-open the contentious debate on the utilisation of the international numerical learner assessments as a proxy indicator of quality of education. However, it is my contention that this debate was not exhaustively concluded. The paper will analyse the arguments in the three papers and suggest some recommendations from the perspective of a Monitoring and

Evaluation (M&E) practitioner.

2. Harvey Goldstein's (2004) Article

Goldstein's article is a critique of the EFA literacy target; his major bone of contention is the inherent potential of the distorting effects that 'high stakes' target setting can lead to. While citing evidence from England and USA where national and state-wide testing were introduced in the 1990s, Goldstein adopts a theoretical descriptive and normative approache to argue that in England, whilst test score levels in those aspects of the curriculum that are tested in public examination results have risen, there has been a backwash effect on learners and teachers that is detrimental to other aspects of quality. This includes a tendency to de-motivate pupils, increased test anxiety especially amongst low achievers and teachers.

In Texas, Goldstein posits that although the high stake testing rewarded schools or teachers on the basis of pupils' test scores with subsequent large gains in student test scores, the gain in the scores over time for the students on the national test was much less than that implied by the Texas test scores. Technically, the writer contends that by adopting these quantitative targets, individuals are encouraged to adapt their behaviour in order to maximize perceived rewards ("even where this is dysfunctional in educational terms"). He also points out the near impossible challenges of creating achievement tests that are culturally or educationally specific and thus suitable for specific socio-economic contexts. Goldstein observes that "if a measuring instrument is restricted only to those items for which we might assume there are no locally specific differences, there is then a real question about whether such an instrument is measuring anything useful" (Goldstein, 2004:9). Also, under the section of *measuring targets*, Goldstein identifies the problematic issues of designing a monitoring framework for any given project. He observes that learning outcomes under the EFA framework do not have a clearly set-out description of what form the relevant assessments might take.

In essence, the writer is raising very pertinent issues related to technical questions surrounding the design of international achievement tests. Finally, and perhaps the most forthright section in the article is the one that explores some of the possible consequences of UNESCO's continued adherence to such targets: The writer strongly argues that the imposition of targets for institutions or school authorities can be viewed as an effective means of centralized control, an increasing control of individual systems by institutions such as the World Bank, which may not only lead to demoralization of poor performing countries but may also allow the imposition from outside of systemic reforms under the heading of 'remedies'. He strongly recommends that each educational system should consider "developing different criteria for assessing quality, enrolment, etc. instead of monitoring progress towards an essentially artificial set of targets.

3. Angeline Barret's (2009) Article

This article is an unequivocal supporter of Goldstein's (2004) paper which argues against quantitative measurement of learner outcomes as an indicator for quality of education. Specifically, the key issues on which these two writers are in agreement include; the difficulty of contextualizing the assessments/tests to suit the different international contexts; potential unintended outcomes of the assessments on learners and teachers and the inadequacy of the test instruments to comprehensively "measure learning". On the latter, Barret argues that using quantitative indicators to measure learning outcomes is tantamount to technically neglecting complex, often culture specific and political nature of education as a social practice. The writer, while arguing against the adoption of Filmer et. al's (2006) proposal for replacement of Millenium Development Goal (MDG) with a Millennium Learning Goal (MLG) in place of universal completion of primary education, asserts that when learning outcomes are used as an indicator of quality, there is a tendency to privilege cognitive learning outcomes that are amenable to measurement by standardized testing. The writer is not a lone voice as far as this issue is concerned (Note 7). Education experts agree that assessment needs to be 'fit-for-purpose'; that is, it should enable evaluation of the extent to which learners have learned and the extent to which they can demonstrate that learning (Brown & Smith, 1997, in Brown, 2004). Other writers further propose that educationists need to consider not just *what* they are assessing and *how* they are doing it (particularly which methods and approaches), but also *why* — the rationale for assessing and *who* (who to participate in the assessment) (Brown, 2004; Wagner, 2012). Finally, like Goldstein (2004), Barret concludes that international quantitative assessments can have unintended effect of impoverishing curricula and educational processes as teachers and learners come under pressure to maximize scores in pen and paper tests, irrespective of whether this enhances useful learning outcomes or not. In essence, the writer finds the quantitative measurement of learner outcomes an

inadequate indicator of quality of education.

4. Daniel Wagner's et. al (2012) Articles

This is a group of papers that provides multiple perspectives in support of learner achievement assessments. The writers adopt various approaches to present their points. For instance, Daniel Wagner's article utilizes a hypothetical theoretical approach to argue in favour of quantitative international assessments. However, he is quick to add that the current learning assessments are a work in progress, and that no one has yet the perfect tool for the many goals and contexts that need to be addressed. Meanwhile, Marlaine Lockheed in a paper titled *Policies, performance and panaceas: what international large-scale assessments in developing countries*, cites the conclusions of four evaluations of The IEA assessments as well as three evaluations of SAQMEC to theorize in support of the large-scale international assessments. Lockheed postulates that these assessments motivate regulatory and behavioural policy reforms around the content of teaching and learning, that they create a learning environment in which assessment specialists can improve their technical skills and related performance and that they increase transparency regarding education system outcomes and human capital development in a cross national context and support analytic work enabled by data sets. In essence, giving reasons, the writer clearly discusses the significance of the achievements assessments at the national level and is unequivocal in support of the same. However, Wagner's and Lockheed's papers are the most relevant in this group of papers (for this review essay). Other papers include Ina Mullis and MichaelMartin's who by citing prePIRLSas an empirical example, explore practical means of contextualizing international test instruments for local educational demands. Amber Gove's article, *Think global, act local: how early reading assessments can improve learning for all* analyses ongoing efforts to establish global-level learning indicators that would require countries to measure the percentage of children meeting locally set targets. The writer does this by drawing comparative examples from the work of Research Triangle Institute's (RTI)[1] Early Grade Reading Assessment (EGRA) conducted in various countries. Finally, Amy Jo Dowd's paper, *An NGO perspective on assessment choice: from practice to research to practice* relies on anecdotal evidence to discuss how assessment evidence changed 'business as usual' for Save the Children basic education programmes. The writer therefore uses Save the Children as a case study to demonstrate how assessments can instigate shifts in approach to technical guidance, national implementation, advocacy and equitable impact for this discussion; this essay will therefore give a lot of attention on these two (Wagner and Lockheed's).

5. Points of Convergence

The writers seem to have divergent views on most points. However, the points on which they hold similar views include the notion that "quality" of education is a multifaceted concept that must be understood holistically. The consensus amongst the writers is that quality education is one that improves learning outcomes, meet the social and affective as well as cognitive needs of learners and create the conditions in the classroom, school and systemic levels that are conducive to learning. The three articles also underscore the need for "quality" education as opposed to mere preoccupation with "quantity". The writers also agree on the general definition of "learning outcomes" to broadly imply literacy, numeracy and life skills, creative and emotional skills, values and the social benefits of education. It is for this reason that the writers postulate that the measurement of the quality of education can never be a linear process and that quantitative learning outcomes can only be a partial proxy measure. Finally, the writers concur that any attempt, especially by international institutions such as The World Bank to change these international assessments from low-stakes to high-stakes enterprises while dangling the "carrot and the stick" on the basis of student test performance has the potential to introduce distortions in the assessments and therefore compromise the ability of the assessments to provide both valid and reliable cross-national measures of human capital investments or valuable data for decision making.

6. Points of Divergence

As already mentioned, Goldstein (2004) and Barret (2009) are "fraternal" voices that strongly oppose reliance on numerical learner achievement targets as a proxy indicator in monitoring the quality of education especially in the context of EFA. In contrast, Wagner et.al (2009) argues in favour of the same. In their arguments, they posit the following reasons:

- Goldstein (2004) and Barret (2009) advance the argument that the eventual outcome of pursuing EFA targets may well be an increasing control of individual national systems by institutions such as the World

Bank or aid agencies, supported by global testing corporations. On the other hand, Wagner, gives this argument "a contempt card" by insisting that learning assessments have grown increasingly important as policy-makers and other educational consumers (agencies, schools, communities, parents, individuals, etc.) seek to understand what is (and isn't) learned as a function of information inputs and that these quantitative measures are important for transparency of education system outcomes that may be compared across national contexts, support analytic work based on solid data sets, support teacher professional development, improve instructional design and reduce learning inequities..

- Where learning outcomes are used as an indicator for quality, there is a tendency to privilege cognitive learning outcomes that are amenable to measurement by standardised testing. This "testing can have the unintended effect of impoverishing curricula and educational processes as teachers and learners come under pressure to maximize scores in pen and paper tests, irrespective of whether this enhances useful learning outcomes" (Barret, 2009:3)**.** In contrast**,** Wagner (2012) does not find anything wrong with competitive testing because in any case, "learning assessments have been around as long as parents have been trying to teach their children, and institutions have been trying to determine who is intellectually fit for a particular job" (ibid pg. 510). Wagner also argues that international large-scale assessments are neither targeted for improving the individual performance of students nor the individual effectiveness of teachers or schools because they are typically sample based – for content domains as well as by students, teachers and schools. Therefore, pressure on individual students and teachers should not arise since these tests "rarely provide information about all students, teachers or schools in a country, they are not good for holding schools or teachers accountable or for creating incentives to reward school or teacher performance" (Wagner et. al , 2012:515).
- Goldstein and Barret further theorize that these assessments may lead to demoralization of poor performing countries and also allow the imposition from outside of systemic reforms under the heading of 'remedies' to put those countries 'on track'. In contrast**,** Lockheed in Wagner et.al (2012) contends that international assessments can document the poor performance of a country relative to other countries at similar levels of economic development, which in turn could motivate a country to alter its investments in human capital development through education. According to him, a detailed analysis of the strengths and weaknesses of a nation's curriculum as compared to that of other countries could motivate regulatory reform regarding the content and methods of instruction and results related to differences in student achievement associated with cross-country differences in teaching strategies which in-turn could invigorate efforts to change the behaviour of teachers through programmes of pre-service and in-service professional development. In effect, what Lockheed is saying is that competition between countries in not necessarily bad, for the "positive jealousy" can in fact be impetus for institutional changes necessary for better learning outcomes.
- Goldstein hypothesizes that the international numerical achievement targets/assessments neglect complex, often culture specific and political nature of education as a social practice. He particularly raises concerns about the comparability and reliability of the data, and the methodological and operational differences between the various countries arising from these international assessments. Wagner on the other hand feels that technical parameters of a 'good' assessment such as sample sizes, alpha coefficients, test-retest reliability, predictive and content validity are non-issues and that these generally have substantial agreement among test-makers the world over.
- Goldstein strongly recommends that each educational system should consider "developing different criteria for assessing quality, enrolment, etc. instead of monitoring progress towards essentially artificial targets set by EFA (….) The emphasis should be on the local context and culture, within which those with local knowledge can construct own aims rather than rely upon common yardsticks implemented from a global perspective". On the other hand, Lockheed thinks that this micro approach will cause countries to miss-out on benefits that accrue from the participation, ostensibly because these assessments create a learning environment in which national assessment specialists can improve their technical skills and related performance. "A country's participation in international large-scale assessments reinforces national technical and managerial capacity for assessment, however, through both training and hands-on experience. It exposes participants to international quality standards in testing and measurement; it provides participants' experience with the technical fields of sampling, test development, questionnaire development, data management and quality control; it builds participants' management capacity for undertaking large research endeavours; it helps education officials prepare reports for policy-makers" (ibid Pg. 515)

7. Why the Divergence?

As already demonstrated, from a conceptual point of view**,** Goldstein and Barret argue that learner achievement target is illogical. Meanwhile, Wagner argues in favour, citing the public good of the tests. The question therefore is; why are the writers so divergent in their views?

From the onset, the writers hold two opposing epistemological (Note 8) views: positivism and interpretivism (Note 9). Goldstein adopts more of interpretivism approach and Wagner is positivist. For instance, Goldstein (2004) hypothesizes that the international numerical achievement targets/assessments neglect complex, often culture specific and political nature of education as a social practice. He particularly raises concerns about the comparability and reliability of the data, and the methodological and operational differences between the various countries arising from these international assessments. Wagner on the other hand feels that technical parameters of a 'good' assessment such as sample sizes, alpha coefficients, test-retest reliability, predictive and content validity are non-issues and generally have substantial agreement among test-makers the world over. Goldstein further recommends that each educational system should consider "developing different criteria for assessing quality instead of monitoring progress towards an essentially artificial target set by EFA. In refuting this, Lockheed in Wagner et. al (2012) contends that this micro approach will cause countries to miss-out on the many benefits that accrue from participation in international learner assessments. In other words, the writers differ in their viewpoint because they are adopting different philosophical positions; Goldstein posits that achievement can-not be universalized while Wagner thinks learner outcomes are comparable since the tests are standardized.

Divergence between the writers also emerges in the way they perceive dependency i.e. they take opposing sides as far as the dependency theory (Note 10) is concerned (read Noah and. Eckstein, 1988). Reading Goldstein's article, one feels that the writer is suspicious about 'who evaluates' and 'for what purposes'. He assumes that there are hegemonic penchants by western multinational/international institutions to abuse the intention and consequences of these assessments.

As already highlighted, the writers differ on the possible unintended side-effects of adhering to these assessments. Goldstein fears that they may lead to impoverished curriculum and put learners and teachers under pressure to maximize scores. On the contrary, Wagner feels that these assessments are neither targeted for improving the individual performance of students nor the individual effectiveness of teachers or schools because they are typically sample based – for content domains as well as by students, teachers and schools and pressure to individual students and teachers should not arise. Looking at the content of these discussions, it is apparent that the writers differ on two points: why the tests are conducted and the link between these assessments and the individual learner, teacher and school. We find this divergence purely emanating from a conceptual understanding, which is also exacerbated by the opposing philosophical positions adopted by the writers.

The final reason for the divergence could be due to scanty empirical evidence to support the notion that indeed international assessments actually impact on education policy development amongst the participating countries, thus giving Goldstein ammunition. It is likely that some of the contentious issues would have been resolved if adequate scientific evidence were available.

8. Discussion

Goldstein (2004) and Barret (2009) are justified in suspecting hegemonic tendencies by multinational/international institutions and how these bodies 'abuse' privilege by imposing control over developing countries in the guise of development aid (read Rodney and Pogge (Note 11) in Milner, 2005 (Note 12). It is important that these issues are flagged every so often so as to deter these tendencies. However, we find their argument about the possibility by institutions such as the World Bank or aid agencies with the support of global testing corporations to control individual national systems presumptuous. The writers would perhaps have cited similar instances to support their argument. Moreover, donor institutions are justified to demand for results whenever states accept aid because 'the most effective evaluation practice balances accountability and learning' (Jackson, 2013) (Note 13). And this is not in any way related to the adoption of quantitative indicators such as learner achievements. Demanding for results from aid is a basic tenet of accountability (Note 14).

Having said that, we concur with Goldstein that such a target should not be set artificially by UNESCO or any other external body but by individual countries, unless the said body is funding the anticipated change. However, I wish Goldstein would have strengthened his argument by asking the extent to which these international learning assessments have achieved their objective in helping countries design their own assessments models

(UNESCO/UNICEF, 1994) and like Lockheed in Wagner (2012), ask for the evidence to support the argument that these international large-scale assessments have indeed benefited participating countries i.e. have they led to better educational policies and outcomes? Have they inspired a culture of evaluation and monitoring at the grassroots level where EFA is delivered (Note 15)? In other words, do the benefits outweigh the theorized side-effects? These are issues we anticipated Goldstein and Barret to have discussed exhaustively. In any case, learning assessments are essentially tools. And, like any tool, the learning assessments may not be perfect (Wagner, 2012). Therefore unlike Goldstein, Lockheed while supporting the international large-scale assessments as a public good cites evidence from several systematic evaluations i.e. four evaluations of IEA's assessments across low- and middle-income countries (Elley 2002; Gilmore 2005; Aggarwala 2004; Lockheed 2010) as well as three evaluations of SAQMEC in sub-Saharan Africa (Murimba 2005; Ercikan et al. 2008; Nzomo and Makuwa 2006 which have established that international assessments have indeed influenced regulatory activities around curriculum content, performance standards, behavioural activities around classroom instruction and teacher professional development. But we are not suggesting that these studies are adequate to assuage Goldstein's fears, but it is the right step in the right direction and many more empirical studies should be conducted.

We posit that these assessments should be used to judge performance and also be regarded as a tool for policymakers and the general public to interrogate the values and purposes that their societies place on education-(see Feuer, 2012). As it is now, politics and attention is focused on the former with little attention given to the latter. In other words, there is minimal evidence, especially in developing countries to support the impact on policy that these international large assessments have had (ibid). This being the case, one wonders: what are the power aspects of assessments as in, whose interests do they serve, and why the gap between the systems of evaluation and student learning/outcomes in respective countries? we therefore postulate that as it stands now, Goldstein is right to suspect hegemonic undertones in the way the international assessments are run and that engaging a few national officials in test development does not necessarily mean policy makers will utilise the findings to review/formulate policy and that EFA should find better means to engage education planners to first understand what these results imply and show case how the findings should influence policy especially in resource allocation, teacher training, continuous assessment e.t.c. But more importantly, EFA should also empower national governments with requisite skills to develop and implement own learner outcome assessments especially literacy in early childhood grades (Note 16). As mentioned elsewhere in this paper, rigorous policy studies should be done to empirically document the role these assessments are playing to influence education policies across the globe.

It is indeed true that the validity and reliability of the international test instruments has been a major contentious issue in education. As researchers in literacy, we agree with Tierney (1998 cited in Hess et al, 2006) that though in most instances tools within an assessment package as applied by international/regional assessment frameworks are mostly *standardised* (Note 17), there are situations (cultural differences, learning differences, etc.) that will make uniform standardisation difficult to achieve. "There will always be tension between the need for uniformity and the need for measures that are sensitive to differences" (Hess, 2006:30). In other words, outcome tests should be context specific. "However, the challenge is that "if literacies are defined by their use in context, then assessing how literate a community has become will differ from one community to another. (…) Locally relevant assessment procedures will give the best picture of the growing use of literacy, but their context-specific nature will make it less meaningful and more difficult to generalise across provinces/states, let alone arrive at national estimates (UNESCO, 2005:25). In any case, all the countries participating in this international achievement tests have their own national assessments frameworks (but may not necessarily be measuring achievement in literacy, science and numeracy) which Goldstein and Barret seem to ignore.

Moreover, having each country set their own assessment framework is not disputed by EFA. Perhaps the two writers should have instead asked the extent to which the primary objective of learning assessments has been met. i.e. According to UNESCO/UNICEF (1994), the primary objective of the EFA learning achievement project is to help countries design their own assessments models, appropriate and flexible and country specific (taking into account the different educational needs. In addition, UNESCO (2005) argues that the aim is to achieve tables of literacy rates which are more reliable, which mean something in terms of the national reality they represent and which enable a sensible comparison for the purposes of allocating resources and effort. This implies data which are comparable across countries and regions of the world (Wagner et al, 2012). It is also important to note that UNESCO is aware of these issues as raised by Goldstein and Barret. While acknowledging this challenge, UNESCO postulates that the need for comparability may leave open the question of what is to be assessed in different contexts and whether the results will be comparable within and between countries. The EFA Global Monitoring Report of 2002 also drew attention to this, as well as to the question of how far local (i.e. sub-national) realities should shape assessment

processes (UNESCO, 2005:25). Analyzing these debates would have been a better endeavor for Goldstein (2004) and Barret (2009). For debate's sake, Goldstein should seek to answer questions such as: what benefits will countries miss by opting for his recommendation? And does participation in international qualitative assessments imply that a country cannot have another customized country-based criteria cut to fit? And what implication will this have on countries given globalization (Note 18)?

The argument by Barret (2009) that outcome assessment may have unintended effect of putting teachers and learners under pressure in order to maximize scores also need further interrogation. It is indeed true that learner assessments especially at national levels (micro) have been largely abused- just as Barret puts it (Note 19). However, Barret seem to miss out on an important element of these tests, which Wagner (2012) does well to highlight; that international large-scale assessments are neither targeted for improving the individual performance of students nor the individual effectiveness of teachers or schools because they are typically sample based – for content domains as well as by students, teachers and schools. We concur with Wagner's conclusion that, pressure to individual students and teachers should not arise since these tests rarely provide information about all students, teachers or schools in a country.

As already mentioned, Lockhead in Wagner et.al (2012) contends that international assessments can document the poor performance of a country relative to other countries at similar levels of economic development, which in turn could motivate a country to alter its policy and investments in human capital development. But the writer is quick to point out that concrete evidence of the impact of international assessments in developing countries is relatively slim and largely based on non-scientific surveys and interviews. However, in essence, what Lockheed is saying is that competition between countries in not necessarily bad but necessary for institutional changes necessary for better learning outcomes. Meanwhile, Goldstein (2004) and Barret (2009) argue that these assessments may lead to demoralization of poor performing countries. This is an important argument. Perhaps it explains why these international achievement tests have been politicized (Note 20). This paper will offer two recommendations in attempt to reconcile this argument: First, the paper will adopt Szekley's (2013) framework, who like Goldstein, posits that recognition of high performers inevitably goes hand in hand with discrediting and exposure of low performing countries. But citing examples of The World Bank's *World Development Indicators* and The IMF's *International Financial Statistics* publications (both make information accessible for further analysis-giving one opportunity to derive conclusions from the data without opportunity for ranking) Szekley offers the possibility to systematize international assessments without necessarily producing country rankings or offering value judgment derived from the indicators. While doing this, like the two publication which collect data on a wide array of sources and issues and make information accessible for further analysis (Szekley, 2013), EFA should also consider doing likewise and without giving unbalanced emphasis/attention to learning competencies,- but rather treat other indicators in similar breadth. In any case, the indicators measured by international assessments relate to basic learning competencies -- literacy, numeracy and life skills and other factors which influence learning-- from personal characteristics (of the students, parents and teachers), home and school environments, to issues of access and equity (UNESCO-UNICEF, 1994). In other words, we contend that the measuring of "learning" should take a "whole school approach (Note 21)", where evaluators look not only at learning outcomes, but the extent at which the education provided meets the socio-economic, religious aspirations of a people, and that such an education is provided without coercion and in a conducive learning environment, and the community participates in the delivery while the teachers are well trained and equipped to facilitate knowledge acquisition.

Even if we must rank or offer value judgments, the raking should be relative to a baseline and the resultant judgment made relative to this baseline (value). In essence, from an M&E perspective, "poor" performance needs to be clarified. A concrete baseline (consider having a counterfactual (Note 22)) will enable participating countries to demonstrate change in a scientific and easy way. Follow-up assessments should therefore demonstrate the change based on the baseline. See figure 1 for this illustration. As the figure illustrates, 'performance' should be construed in relative terms. Countries participating in the international large-scale assessments should be evaluated- and if must be ranked- should be on the basis of their performance relative to the baseline i.e. improvement. But as earlier said and in support of Goldstein, such target setting should be done by individual countries.

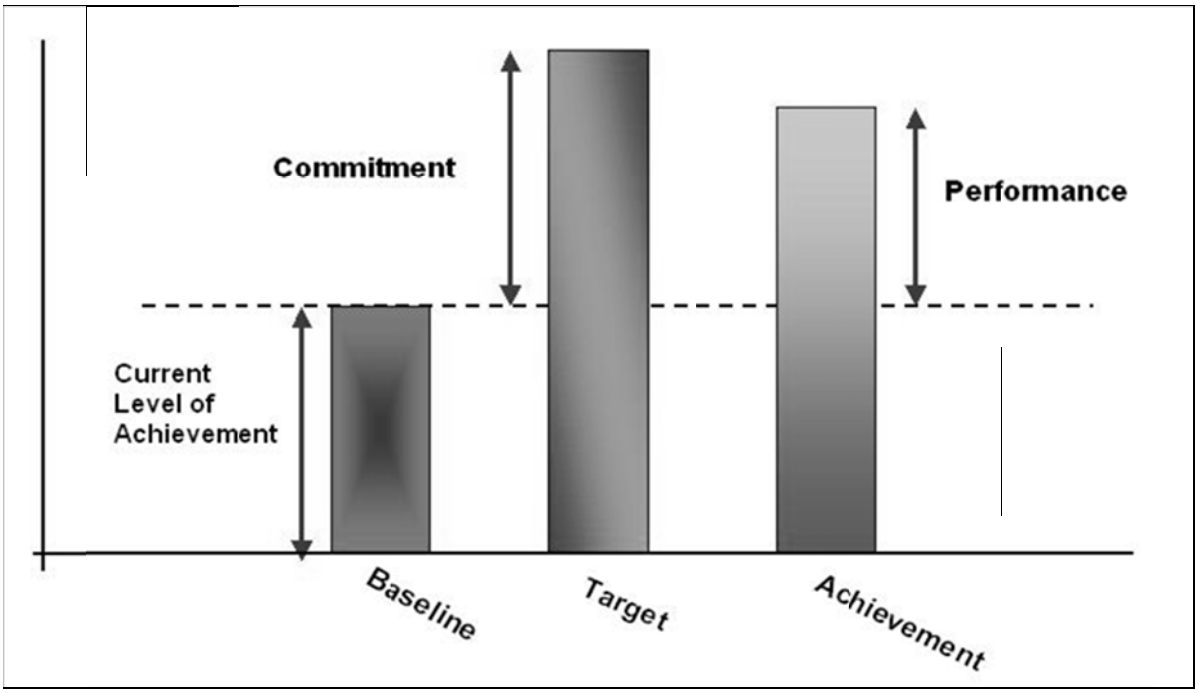


Figure 1. Understanding Project "Performance". Adapted from Mulongo in Stather. et.al (2014)

Finally, Goldstein (2004) raises some issues related to the technical side of the assessments that EFA ought to address. For instance, he finds the EFA framework deficient in a clearly set-out description of what form the relevant assessments might take. He does this by citing the goal related to obtaining a 50% improvement in adult literacy as an example. Moreover, Goldstein raises issue with the definition of the targets/indicators as adopted by EFA. He particularly finds the series of five 'levels', from basic to advanced as used in IALS complicated. He concludes that "the failure of EFA to recognize and articulate this issue suggests that the stated aim has more in common with a political slogan than with a scientifically based aspiration" (ibid, pg. 10). Goldstein's sentiments are echoed by Wagner (2003) who identifies: (a) scales of literacy achievement (from dichotomous, to five levels, to be too many); (b) difficult with determination of when a 'level' is achieved (e.g., in International literacy tests (ILT), is it adequate to say that a level is achieved if and only if 80% of the items in a level are completed successfully?; see Levine, 1998, cited in Hess et al, 2006); (c) what is included in the operational definition of literacy; (d) effectiveness of the use of proxy measures (Lavy et al., 1995; Murray, Kirsch, & Jenkins, 1998 cited in Hess et al, 2006). All these questions are relevant and are in essence questioning the strength of EFA's monitoring framework. In other words, EFA should clearly define the indicators as used, provide a simple operational definition of 'literacy' and consider reducing the dichotomous scales/levels of measurement from the current five to four; Runo (2010) and Mulongo (2013) adopted this with profound success. Table 1 shows a framework we would recommend to achieve the foregoing. We believe that if this framework is adopted, most of the technical contentious issues raised by the writers will be addressed. This framework also gives individual countries opportunity to set own targets and can also track progress/performance over time.

Table 1. Recommended Monitoring Framework

Indicator	Indicator definition	Baseline	Target	Means of verification	Method of data collection	Responsibility
List your indicators in this column(Note 22). E.g. % of learners achieving grade level competencies in reading Literacy.	'Set the boundaries' (define) for the indicator as clearly as possible. Also indicate level of disaggregation here. E.g. The proportion of learners who are sampled and tested using standardized reading test who achieve grade level competencies (level 2 and above(Note 22)) **Computation:** This will be computed by taking the number of learners who garner grade level competencies on standardized reading test over the overall sample of learners tested. **Disaggregation:** Grade, Gender and Region etc.	Indicate the baseline values here (see figure 1 above) ***e.g.2013: Values:*** 45.5% and 34.1% learners achieved grade level competencies at the primary and secondary levels respectively	Indicate a target here (see figure 1 above) 60.6% primary, 51.2% in secondary by 2015	Indicate evidence stakeholders will rely on to ascertain that the target has/is being achieved? e.g.PISA Achievement test reports, EFA monitoring report etc.	Indicate the method you will rely on to collect data for the indicator e.g. Annual school survey, census etc.	Indicate whose responsibility it is to manage technical issues related to this indicator (survey design, execution, data analysis, report writing) e.g. UNESCO Statistics Department

9. Conclusion

This review essay has looked at three publications discussing the contentious issue of evaluating education quality by learner outcomes as a proxy indicator. The essay has explored the debates looking at points of convergence and disagreements between Goldstein (2004) and Barret (2009) as opposers of quantitative international learner assessments and Wagner et.al. (2012) as proposers. The essay has also offered a few recommendations to reconcile the debates or bridge some gaps for better practice: that 'performance' on these international assessments should be construed in relative terms. Countries participating in the international large-scale assessments should not be ranked and if must be evaluated or ranked, it should be on the basis of performance relative to a national baseline and that EFA framework should clearly define its indicators, provide a simple operational definition of 'literacy' and consider reducing the dichotomous scales/levels of measurement from the current five to four as adopted by Runo (2010) and Mulongo (2013). The paper concludes that the international learner assessments may not be perfect but that there is tremendous power in measuring performance

References

Barret A. (2009). *The education Millennium Development Goal beyond 2015: Prospects for quality and learners.* EdQual, Department for International Development, UK

Dakar Framework: http://unesdoc.unesco.org/images/0012/001211/121147e.pdf

EFA monitoring report 2005: http://www.unesco.org/education/gmr_download/en_summary.pdf

EFA monitoring report 2009: Retrieved from http://download.ei-ie.org/docs/IRISDocuments/Education/Education%20For%20All/Global%20Monitoring%20Report%202009/2009-00090-01-E.pdf. Accessed on 16/02/2013

Feuer J.M. (2012). *No Country Left Behind: Rhetoric and Reality of International Large-Scale Assessment.* Research & Development Center for Research on Human Capital and Education, Princeton. Retrieved from http://www.ets.org/Media/Research/pdf/PICANG13.pdf (accessed on 4/6/2013)

Filmer, D., Hasan, A., & Pritchett, L. (2006). *A Millennium Learning Goal: Measuring Real Progress in Education*, Working Papers 97, Center for Global Development.

Goldstein H. (2004). Education for all: the globalization of learning targets, *Comparative Education*, *40*(1), 7-14. Retrieved from http://www.tandfonline.com.elibrary.ioe.ac.uk/toc/cced20/40/1 (accessed on 16/03/2013)

Hess (2006). Retrieved from http://www.schools.utah.gov/eval/documents/Lit_EarlyLiteracyReview.pdf. Accessed on 16/02/2013

Jackson E.T. (2013). Interrogating the theory of change: evaluating impact investing where it matters most, *Journal of Sustainable Finance & Investment.* Retrived from http://www.tandfonline.com/doi/pdf/10.1080/20430795.2013.776257 (accessed on 26/04/2013)

Kusek J., & Rist R. (2004). *Ten Steps to a Results-Based Monitoring and Evaluation System*. The World Bank, Washington. http://dx.doi.org/10.1596/0-8213-5823-5

Milner H. (2003). Globalization, Development, and International Institutions: Normative and Positive Perspectives. *Perspectives on Politics.* December 2005 | Vol. 3/No. 4

Mulongo G. (2013). *Preschool as a Predictor of Literacy Competency of Learners in Kenya: The Case of Pupils Enrolled in EMACK Affiliated Schools.* LAP LAMBERT: Academic Publishing

Noah H.J., & Eckstein M. (1988). Dependency Theory in Comparative Education. In Jürgen Schriewer and Brian Holmes, eds., *Theories and Methods in Comparative Education* (Frankfurt am Main: Peter Lang, 1988), pp. 165-192. Reprinted by permission of Peter Lang Publishers.

O'Donoghue, T. (2007). *Planning Your Qualitative Research Project: An introduction to interpretivist research in education*. Routledge. Abingdon, Oxon.

Pogge T. (2002). *World Poverty and Human Rights.* Cambridge, UK.

Runo M. (2010). *Identification of Reading Disabilities and Teacher-Oriented Challenges in Teaching Reading to Standard Five Learners in Nyeri and Nairobi Districts, Kenya.* Unpublished PhD Thesis, Kenyatta University, Kenya.

Stathers, T., Low, J., Mulongo, G., & Mbabu, A. (2013). Volume 6: Topic 12 - *Monitoring of Orange-fleshed sweetpotato dissemination and uptake*. In: Stathers, T., Low., J., Mwanga, R., Carey, T., David., S., Gibson, R., Namanda, S., McEwan, M., Bechoff., A., Malinga, J., Benjamin, M., Katcher, H., Blakenship, J., Andrade, M., Agili, S., Njoku, J., Sindi, K., Mulongo, G., Tumwegamire, S., Njoku, A., Abidin, E., Mbabu, A. (2013). *Everything You Ever Wanted to Know about Sweetpotato*: *Reaching Agents of Change ToT Manual*. Nairobi: International Potato Center. pp. 287-310.

Wagner D. et.al. (2012). The debate on learning assessments in developing countries. *Compare: A Journal of Comparative and International Education, 42*(3), 509-545.

World Bank (2004). *Assessing Student Learning in Africa.* The World Bank. Washington D.C.

Notes

Note 1. Read Tikly, 2011) for more details. In a summary, Tikly posits that a good quality education arises from interactions between three overlapping environments, namely the policy, the school and the home/community environments. Elsewhere in this paper, I have argued "quality education" should be regarded from a "whole school approach" point of view; that learning outcomes are just one face of a multifaceted concept, a concept that should include the extent to which the education provided meets the socio-economic, religious aspirations of a people, and such an education is provided without coercion and in a conducive learning environment with proactive participation by the community in its delivery and that teachers are well trained and equipped to facilitate knowledge acquisition. The graduates of such an education system should be dynamic, creative, adaptive and have unwavering loyalty to their local ideals but are globally focused and competitive. More articles "Understanding education quality" by UNESCO available at http://www.unesco.org/education/gmr_download/chapter1.pdf. Accessed on 16/02/2013

Note 2. According to UNESCO-UNICEF (1994), the indicators being measured relate to basic learning competencies -- literacy, numeracy and life skills. Other factors which influence learning-- from personal characteristics (of the students, parents and teachers), home and school environments, to issues of access and equity are also measured through a variety of tests, questionnaires and during these survey.

Note 3. The World Declaration on Education-For-All, adopted at Jomtien in March 1990, pointed to the need "*to define acceptable levels of learning acquisition for educational programmed and to improve and apply systems of assessing learning achievement"*. It is widely recognized that merely improving the supply of education -- quantity -- is not enough, and an improvement in quality is considered vital.

The joint UNESCO/UNICEF project on Monitoring Education-For-All Goals: Focusing on Learning Achievement, begun in September 1992, was designed to answer this need for a new type of monitoring. Because the primary objective is to help countries design their own models -- and not to hand over ready-made instructions on "how-to-do-it" -- the focus has been to assist countries in developing tools which will work for them. Unlike traditional forms of evaluation, based largely on rigid, normative and standardised procedures, the (UNESCO-UNICEF, 1994).

Note 4. According to Suter, International comparisons of student achievement involve assessing the knowledge of elementary and secondary school students in subjects such as mathematics, science, reading, civics, and technology. The comparisons use test items that have been standardized and agreed upon by participating countries According to Larry Sutter, these tests are administered to a sample of students in 100 to 200 schools, which are selected to represent all students in the country. An international referee monitors the school selection process to insure that all countries follow correct sampling procedures. The test items are scored according to internationally agreed-upon procedures and are analyzed at an international center to insure cross-national comparability. Countries that do not meet high standards of participation are not included in the comparisons (Un-dated online article available at http://education.stateuniversity.com/pages/2115/International-Assessments.html (accessed on 1/05/2013).

Note 5. Read more: International Assessments - International Association For Educational Assessment, International Association For The Evaluation Of Educational Achievement, Iea And Oecd Studies Of Reading Literacy - OVERVIEW - StateUniversity.com http://education.stateuniversity.com/pages/2115/International-Assessments.html#ixzz2S45ffR8j

Note 6. Un-dated online article available at http://education.stateuniversity.com/pages/2115/International-Assessments.html (accessed on 1/05/2013)

Note 7. Fareed Zakaria writing in *Time Magazine says that test-scores are only one measure of student's achievement and that other qualities must be taken into account as well. (Time, The thin envelop crisis vol. 18, No. 14 April 15 2013*

Note 8. Epistemology is "the study of how knowledge is generated and accepted as valid" (O'Donoghue, pp. 9)

Note 9. Positivism is linked to objectivism. Positivists believe that knowledge about society can be gained via rigorous scientific methods that revolve around prediction and control (Bryman, 2008; Cohen, 2007; O'Donoghue, 2007). An interpretivist presupposes that knowledge is subjective and dependant on specific situations (non-generalizable) (Bryman, 2008; Cohen, 2007; O'Donoghue, 2007). For an interpretivist, knowledge is fluid and changes. Interpretivism revolves around understanding different people's viewpoint and experience and interpreting them.

Note 10. Dependency theory argues that the worlds present state is an outcome of domination by nations who have over the have-nots and, within nations, by the domination of classes who have over have-nots. The world as it is supposed to exist today is explained by the concepts of center-periphery, hegemony, and reproduction. The unilateral exercise of power by the center on the periphery, the hegemonic of the dependent through the systematic reproduction in the periphery of the values of the center. (read more about this available at http://www.scribd.com/doc/51793069/Dependency-theory-in-Comparative-Education (accessed on 30/06/2013)

Note 11. Thomas Pogge in Milner (2003) employing normative and empirical analysis argue that developed countries and international institutions are harming the poor countries and have therefore an obligation to stop such harmful behavior. He postulates that in the interdependent world we live in, the advanced industrial countries support an international system that makes coups, civil war and corruption in LDCs not only possible but likely. Pogge opines that by upholding a government's privileges to borrow and assign rights for domestic resources—no matter how bad the government is, the international institutions and rich countries encourage a free for all for control of developing countries. "Failure to recognize the rich nations' role in harming the poor countries depends on the "explanatory nationalism" that dominates current research and thinking" Pogge in Milner (2003:4)

Note 12. Available at http://www9.georgetown.edu/faculty/jrv24/milner_05.pdf (accessed on 23/04/2013)

Note 13. Available at http://www.tandfonline.com/doi/pdf/10.1080/20430795.2013.776257 (accessed on 26/04/2013)

Note 14. "Whether it is calls for greater accountability and transparency, enhanced effectiveness of development programs in exchange for foreign aid, or real results of political promises made, governments and organizations must be increasingly responsive to internal and external stakeholders to demonstrate tangible results" Kusek and Rist

(2004:1)

Note 15. There is weak evidence that learner assessments through the EFA framework has inculcated the culture of monitoring of learner outcomes especially in developing countries. I post that though the regional and international learner assessments have endeavored to incorporate test experts from participating countries (the intention is to build capacity and for item standardization) in designing the test instruments (see Lockheed in Wagner, 2012), this has not automatically resulted into trickledown effect of the culture of evaluation in these countries. Perhaps it is time we changed this strategy. In this paper, we suggest the need to work through credible regional and national civil organizations (see 'Literacy Watch" article available at http://www.accu.or.jp/litdbase/literacy/nrc_nfe/eng_bul/BUL21.pdf-accessed on 7/6/2013 and Dowd (2012) in Wagner et.al, 2012) to help build the capacity of education ministries of developing countries so that they are not only able to design and implement national assessments (especially numeracy and literacy), but also able to teach (see Runo 2010; Chall, 1983) . As the situation is, it is difficult for most developing countries to design and implement national learner outcome assessments let alone sustaining the culture of evaluation and monitoring as hegemonically enforced by the regional and international learner assessment frameworks.

Note 16. Studies show that where proper foundation in reading is not given to learners, such learners continue experiencing problems in upper primary and probably in their lifetime (Runo, 2010; Chall, 1983). Chall says that learners pass through reading developmental stages just like child development and therefore teachers, curriculum developers parents and other stakeholders need to put in place the necessary tools and measures to enable children to learn how to read. According to research, if children have difficulties in literacy by age 8 or just about class 3, these children will continue to experience difficulties in the same and school subjects if nothing is done to mitigate the situation (Chall 1983; Morrow 2005; Runo 2010)

Note 17. Emphasis is mine

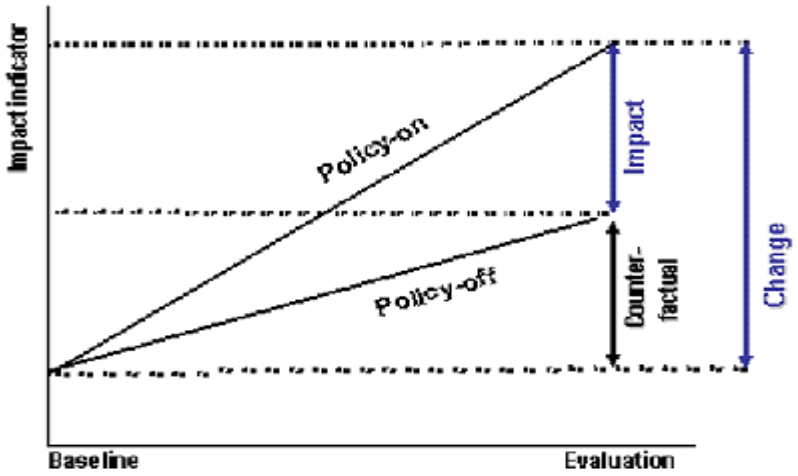
Note 18. According to Milner (2005), the increasing integration of national economies into a global one means that all countries are increasingly affected by what goes on in the others. Milner asserts that we are now "one world" and gone are the days of the Westphalian system of individual states. She argues that states are not separate, self-contained units that can implement autonomously programs without regard to the global agenda because their internal situation is affected by international factors. "In the contemporary world, human lives are profoundly affected by non-domestic social institutions—by global rules of governance, trade and diplomacy" (Pg, 13)

Note 19. Read Betebenne and Linn (2009) who used empirical data to unpack issues related to student growth by situating the discussion within three larger, intersecting topics: Measurement, longitudinal data analysis, and accountability. Available at http://www.k12center.org/rsc/pdf/BetebennerandLinnPresenterSession1.pdf. Accessed on 1/05/2013

Note 20. Andy hargreaves in an online article assert that there is widespread international interest on the relative performance of countries following each test, leading to debates and discussions. Hargreaves cites an example following the results of PISA 2009 in which Shanghai, participating in the study for the first time, became the top performing country in the reading, mathematics, and science literacy scales, causing intense debates in the USA, ranging from speculation on the reasons Shanghai was able to top the PISA ranks to whether certain interest groups were using the USA's performance to launch a panic attack in their efforts to further a particular political agenda for education. Other pundits were more blasé and less apprehensive about the 'threat' posed by the Shanghainese students; they attributed the surprising performance of Shanghainese children to an education system that conducts relentless test preparation under a highly centralized, highly pressurized education system. (available at: http://andyhargreaves.weebly.com/1/post/2011/08/international-achievement-tests-and-educational-change.html. Accessed on 1/05/2013)

Note 21. Read more about this at https://classroomconnections.eq.edu.au/topics/Pages/2013/issue-6/whole-school-approach.aspx (accessed on 4/06/2013)

Note 22. According to EuropeAID (available at http://ec.europa.eu/europeaid/evaluation/methodology/methods/mth_att_en.htm. Accessed on 29/04/13), a the counterfactual can be summarised as in the diagram below:



The "policy-on" line shows the observed change, measured with an impact indicator, between the beginning of the evaluated period (baseline) and the date of the evaluation. For instance: local employment has increased, as has literacy. The impact accounts for only the share of this change that is attributable to the intervention. The "policy-off" line, also called the counterfactual, is an estimate of what would have happened without the intervention. It can be obtained with appropriate approaches like comparison groups or modelling techniques. Impact is assessed by subtracting the policy-off estimate from the observed policy-on indicator. The assessed impact, derived from an estimate of the counterfactual, is itself an estimate. In other words, impacts cannot be directly measured. They can simply be derived from an analysis of impact indicators. Only a counterfactual allows for a quantitative impact estimate. When successful, this approach therefore has a high potential for learning and feedback. It is nevertheless relatively demanding in terms of data and human resources, which makes it somewhat unusual in evaluation practice in developing countries.