Universal Assessment and the Bottom of the Pyramid

Luis Crouch
RTI International

Contents
Introduction and background.................................................................................................................. 2
The bottom of the pyramid: not just an equity concern.......................................................................... 4
What kind of assessment will most reliably help countries bring up the bottom of the pyramid? .......... 10
  Current and historical usage: what has been best practice?................................................................. 11
  Ask the experts.................................................................................................................................. 11
  A set of practical approaches.............................................................................................................. 12
  “Bottomization” with rigor, “rigorizing” the bottom, “even faster and smaller,” and other options .... 13
  “Useful for the bottom” adjustments to classically rigorous measurements...................................... 13
    Some TIMSS and PIRLS data........................................................................................................... 13
    Some PISA data ............................................................................................................................ 17
  Rigorization of measurements designed for utility at the bottom of the pyramid............................ 19
    Exploration of “rigorization” of EGRA and similar tools .............................................................. 19
    “Rigorization” of ASER and “citizen-led” assessments .................................................................. 23
    The experience of LAMP................................................................................................................ 25
  Even smaller, quicker approaches ..................................................................................................... 26
  Hybrid approaches............................................................................................................................... 26
    Linking........................................................................................................................................... 27
    Common scales and quality assurance procedures ........................................................................ 29
Concluding comments, next steps ....................................................................................................... 31

---

1 This is an unpublished background paper for a conference presentation given at the Penn Invitational Working Conference on “Learning at the Bottom of the Pyramid”, University of Pennsylvania, March 2-3, 2017, Philadelphia, USA.
Introduction and background

While the SDGs now officially call for global reporting on learning outcomes, many institutions and scholars had noticed, at least since the mid-2000s, that many children were not learning much, and were starting to respond by, as a first step, advocating and developing assessments that, they felt, was perhaps more appropriate to learning at the bottom of the pyramid, or at the left end of the cognitive distribution. The GMR sounded a clarion with their estimate that there are some 250 million children in the world hardly learning. Researchers such as Filmer, Hasan and Pritchett (2006) had called for a “Millenium Learning Goal.” Scholars such as Dan Wagner at the University of Pennsylvania (Wagner 2003) had called for more appropriate, “smaller, quicker, cheaper” assessments to supplement the large-scale international ones.

Partly in response, and partly because many others were simply noticing the same things these researchers were noting, there has been a massive development of actual assessments, and assessment ideas and proposals, since then. Among the international assessments, PIRLS has spawned PIRLS Literacy (Mullis and Martin 2015), TIMSS has led to TIMSS Numeracy (Mullis and Martin 2013), and PISA for development now exists.2 LaNA, also intended for the bottom of the pyramid, is a relative newcomer in the IEA family,3 Citizen-led assessments such as ASER,4 UWEZO,5 and several others have come into being. ÉGRA, EGMA, and others were also a response.6 LAMP, in the meantime, had identified the fact that in poor populations “reading,” if interpreted as truly using print to get information, was difficult to measure, and that one needed to measure precursor or “component” skills—in a manner somewhat similar to ÉGRA.7 As of 2016, UNICEF is proposing using a very short literacy and numeracy “dipstick” that is the “smallest, quickest and cheapest” (one presumes) of all the options available since 2003 (Cardoso 2016). Finally, UIS, working with ACER in Australia, is proposing the development of a global learning scale—not a new assessment per se, but a tool against which assessments could benchmark.8 And, one should not forget that the DHS and MICS surveys have, for a relatively long time, asked respondents to read a single sentence on a card (Smith-Greenway 2015)—and the results of such surveys were being used for data analysis and policy recommendations as early as the mid-2000s, noting that in many countries quite a few years (4 or 5) of education were insufficient to guarantee universal literacy—even using the very low bar of reading a sentence.9 Of the existing three the regional assessments systems, two (PASEC and LLECE) reach quite far down into the grade structure so that they can be considered “early grades” (which tends to be more suitable to the bottom of the pyramid, given that it is in the early grades that problems start). In addition, the Center for Global Development has called for the creation of

---

4 http://www.asercentre.org/#9mm8y.
5 http://www.uwezo.net/
something that could be described as an early-grades PISA: something age-based, with 3 cognitive domains, applicable around age 9. This exists only as a very general concept at this time. Finally, new regional assessments are either incipient or being discussed. A quick count of the options listed here (and I am probably missing some) suggests some twelve real or potential options (though some are limited only to certain regions): a lot for countries and development agencies to deal with.

“Smaller, Quicker, Cheaper” therefore now exists—in droves, in a perhaps perfect illustration of the saying “be careful of what you wish for.” Moreover, “Larger, Slower, and Costlier” not only co-exist, but in fact, the interest in, and number of developing countries applying such assessments is growing—by 75% by my (conservative—I left out all former Soviet and Eastern European countries, as well as oil-rich countries) count, since the inception of each. So the landscape now offers more choices and is more complex even if, as other forms of assessment are increasingly applied, the growth in relative terms of the “Larger, Slower, Costlier” option has not been as high as it would be in absolute terms.

At the same time, the concern with learning outcomes has been reinforced since the early 2000s. The SDGs, as is well-known and in contradiction to the MDGs, call for increases in learning. The Education Commission (International Commission on Financing Global Education Opportunity 2016, p. 56) calls for a universal learning indicator. While the SDGs are presented as universal, much of the concern is, rightly, with the fact that the poorest children in the world, those who are at the bottom of the pyramid, are learning so little. And, it should be obvious, the concern is not just with measurement, but with improvement.

This paper addresses an issue that can be put simply but is extremely hard to answer: “Is the array of assessments emerging helping researchers, policy-makers, and implementers get a more accurate sense of how much or how little the poorest children in the world know, and is it helpful in remedying the situation?”

The paper specifically is not addressed at the question of global reporting—although it does touch upon the issue. The problem of interest here is what is most useful for countries to generate movement along the bottom of the pyramid.

The paper’s hypothesized answer is that while we are much better off than, say, 10 years ago, we could do much better still, in at least two or three ways. First, we could be doing a better job of really understanding the comparative advantages and uses of all the options, and offering countries and development agencies a well-annotated menu, particularly in light of the fact that the community has responded so multitudinously (by design or serendipity) to “Smaller, Quicker, Cheaper.” Second, we could be doing a better job of addressing not just how or how well these options measure, but how well they measure if the job is improving learning, not just knowing, reporting, and comparing. Thirdly, in cases where the assessments support learning well, we could be actually using the assessments to improve learning—but that’s perhaps another topic.

The rest of the paper is organized into four sections. A first section is a reminder of why the bottom of the pyramid matters, and that the issue is not just one of equity but of the importance of the bottom of the pyramid to the “average” or “middle” of the pyramid. A second section lays out some criteria, or sources of knowledge, for answering the question as to whether the tools are doing a good job measuring, and also helping improve the situation. A third section will document some of the main the technical issues encountered in trying to respond to measurement at the bottom of the pyramid. A concluding section will argue that, because all attempts to measure at the bottom of the pyramid have different sorts of problems, and different sorts of advantages that may be impossible to sort out with a single assessment for each key node (early grades, end of primary, etc.), we may need many of the different alternatives at the same time.

As a way to limit our scope, we have restricted the paper to look only at primary and lower secondary assessment and improvements, and focuses on cognitive skills. One can generalize (a little) down to early childhood and “up” to workforce preparedness, though the measurement problems there are much harder to deal with, particularly at the bottom of the pyramid. One has to start somewhere.

The bottom of the pyramid: not just an equity concern

The education SDGs rightly emphasize equity and equality issues. And the international community rightly focuses on the absolute numbers (250 million or so) of children in school but not learning the basics. This section will illustrate that the depth of the problem is more severe that the absolute numbers suggest but will also suggest that even if somehow one lived in a universe where policy makers did not care about equity (or, did not care about equity and dared say so), it is very difficult for countries to improve average performance without “taking care” of the bottom of the pyramid.

Various authors have provided estimates of the number of children in poor countries learning at very low levels. The absolute number (250 million) that is most often cited is valuable because absolute numbers draw policy-makers’ attention. But it may be useful to remember just how dire the situation is. I use two graphics to show this, one produced by myself and one by Justin Sandefur (2016).

Figure 1 shows, on the horizontal axis, the various percentiles of performance and on the vertical axis the scores in PISA 2015 reading, for each percentile. The lines represent groups of countries: the ten best overall performers, and the ten worst. Reading off the horizontal axis at the median (50th percentile), finding where that reading intercepts the line for the “Worst 10”, and finding where that reflection intersects the line for the “Best 10,” and, finally, reflecting back to the horizontal axis, we see that about half of the children in these countries perform at about the 5th percentile of children in the Best 10 countries. But this picture is way too rosy, because the Worst 10 in PISA are actually quite developed countries: they are solidly middle-income countries. A different perspective is needed.

---

11 The original expression of this now famous estimate is at http://unesdoc.unesco.org/images/0021/002193/219349E.pdf.
12 Macao (China), Norway, Japan, Korea, Estonia, Ireland, Finland, Canada, Hong Kong (China), Singapore, on the one hand, and Kosovo, Lebanon, Algeria, FYROM, Republic, Tunisia, Indonesia, Peru, Georgia, Qatar on the other.
Fortunately, Sandefur (2016) provides some data linking SACMEQ to TIMSS, thus extending down the range of comparison. The exercise is fraught with methodological issues which Sandefur is the first to address, especially if the purpose is to try to rank, with some accuracy, individual countries that did not participate in TIMSS, to those that did. But I believe it is justified to use his results, *a grosso modo*, for a broader use: to compare groups, and to compare large differences between groups at the extremes. For that purpose, his recent work I think is more useful than anything I have seen. The graphic produced by one of his equating attempts is reproduced here in Figure 2.
This graphic, using the linked circles on the right (Crouch’s addition), selects the levels of performance of the top performers in the (equated, approximately) extended “Worst 10” countries on the TIMSS scale (but which did not participate TIMSS), and compares that performance to that of the “Best 10.” It is clear that even the top 5% of performers in these countries are only at about the median, if that, of the better-off countries. But the linked circles on the left, dealing with the bottom of the pyramid, are more concerning. This shows the medians in the worst-off countries, linked upward to the best-off countries. It turns out that these medians are not in fact at the 5th percentile. These medians are, in essence” off the curve, or “off the chart.” Visually, and knowing that the distribution is of course not uniform or even normal, one would guess that the median is at about the 1st to 3rd percentile. Similar calculations have
been made in other papers (Filmer, Hassan, and Pritchett 2016, Crouch and Gove 2011). These papers don’t always depend on assumptions about normality: one can get insights into where children in one distribution would likely be placed in other distributions, even knowing nothing about the distribution, using Tchebysheff’s inequality.

All this, so far, gives new meaning to the notion that 250 million children are in school but not learning much. “Not learning much” means learning at approximately the level of children that would be considered as having with fairly serious cognitive disabilities. Clearly those children, in well-off countries, are not part of international evaluations. But we know they constitute somewhere around 5% (depending on the country) of the population. The next 5% or so would be the bottom of the distribution in the better-off countries. And it is these children that the median in the poorer countries match.

Now, this is a pure equity or humanitarian concern, in a sense. This is not to gainsay the concern, just to note it.

But another argument is that it seems as if the path towards improving the average tends to mean working up from the bottom. That is, even if all policy-makers cared about was a country’s average performance, the “normal” progress of countries suggests that the bottom of the pyramid has to be addressed, and tends to be addressed in countries that progress in terms of their averages.

The following two charts, from TIMSS 2015 and PISA 2015 make this point.

In both charts, the horizontal axes represent proficiency levels, from lowest proficiency on the left to highest on the right. The vertical axes represent the percentage of children or youth achieving those levels. The percentages are not cumulative. The three groups of countries are ranked in terms of the average scores on TIMSS or PISA.

Two important caveats need to be kept in mind when looking at these graphics. First, while the horizontal axes are a metric in some sense, it is not altogether clear to me that moving between proficiency levels implies the same conceptual leaps. Also, I have explicitly included, for the first level in TIMSS, those below the level (e.g., if only x% of students are at or above the lowest benchmark, I have presented 100-x% as being below it). (Note: PISA already presents those scoring below the lowest level.) Second, these data are a cross-section snapshot, not a dynamic path over time for specific countries. It would be wrong, though again tempting, to say that the “Best” countries in the graph (Singapore, Hong Kong, Korea) had a profile similar to the “Mid” countries today. More could be done to see whether there was some path like this, by looking at time series of TIMSS and PISA data, but that would have taken more work than was in our scope. In spite of these caveats, both graphs are strongly suggestive.

In TIMSS, it is clear that going from Worst to at least Middle involves a vast reduction (about 2/3) in the percentage of children at the bottom of the pyramid, a sizable improvement in the middle levels, and only a very small improvement in the performance at the top end. In going
from Mid to Best, though, numbers at the bottom of the pyramid are further reduced (to almost nothing), while numbers at the top end increase dramatically.\textsuperscript{13}

The same story is evident with the PISA data, though perhaps not as dramatically, perhaps because of the use of more levels of proficiency. (One could re-aggregate those to see if the pattern is similar.) In going from “Worst” to “Mid,” there is a vast reduction in the numbers at the bottom of the pyramid, hardly any increase in numbers at the top end, and a modest increase in the numbers in the middle range of performance. No increase at the top end is visible in the PISA data in going from “Worst” countries to “Mid.” This could be because the PISA assessment does not top out as easily as TIMSS or because the higher proficiency levels are defined as relatively more difficult.

\[13\] Note: the “Mid” countries were not chosen because they are at the middle of the performance distribution, e.g., countries near the median of average score. Had I done so, the results would have been more dramatic, but would have included, in that middle, countries that are upper income. (Just because participation in the international assessments is skewed towards higher-income countries.) Instead, by “Mid” I mean countries whose proportion of children at the lowest proficiency level is roughly a third of the proportion of children at the lowest proficiency level in the countries with lowest overall performance. This does not game my results, since the point is not how big the “drive” down is the proportion of students at this lowest proficiency level, but how big it is compared to the drive up in the high-proficiency levels. Choosing as “Mid” the countries at the median of the distribution would make the “drive” down in the low proficiency levels even bigger.
Figure 3. Performance at different proficiency levels in TIMSS, three groups of countries

Source: Graphed by the author from TIMSS Grade 4 mathematics data.
This section has shown that the “normal” way for countries to improve performance is to bring up the bottom of the performance distribution. This is subject to the caveats mentioned.

What kind of assessment will most reliably help countries bring up the bottom of the pyramid?

It seems to me that there are three ways to get reliable information on this, but typically only one way is being thoroughly used or discussed. In this sense, this section would like to propose a research agenda.

Two points need to be noted. First, whichever one considers to be the “best” kind of assessment at helping countries move at the bottom of the pyramid, has to fulfill at least two criteria of optimality: first, there has to be evidence that the assessment is actually useful, is being used or, as a last resort, there are practical or theoretical reasons that it could be useful; second, it has to meet certain technical criteria, though those criteria have to be sensitive to the type of use.

It also has to be noted that perhaps the best assessments for helping countries move are not the best assessments for, say, reporting on movement in a reliable manner. And, this latter is a
worthwhile goal. So, coming to technical judgment on some of these things involves tradeoffs, and these should be explicitly acknowledged.

These preambles out of the way, let us look at the possible sources of information.

**Current and historical usage: what has been best practice?**

If the purpose is to help countries move the bottom of the pyramid, then a simple and logical question to ask is: what kinds of assessments were used by the countries that are now in the middle of the range, and (presumably) have improved their performance over the last couple of decades (precisely by moving at the bottom of the pyramid), and how did they use (or are using) these assessments to underpin their improvements? Alternatively, to the degree that cross-sectional information is indicative of historical change for those who are now at the top of the pyramid, one could ask what they were using when they were transitioning to levels of skill currently characterizing the “Middle” countries. Note that in all this it is perfectly possible that the research would uncover that there is no pattern to the kind of assessment that was used. Or that, worse still, it is unclear that assessment even played a major role. Honesty would compel one to admit those possibilities from the beginning.

While this would seem logical, it seems to me that not enough of this has been done—at least, such work does not circulate widely in the current debates.

**Ask the experts**

Another logical step is to rely on expert knowledge on what constitutes “good” assessments. There are standard sources, such as the American Educational Research Association’s Standards for Educational and Psychological Testing (AERA 2014).

But there is a problem here. These standards are extensive, and very sophisticated. They do often take into account the purpose of the testing or assessment. But, they typically do so in the context of what one might call “normal science” in the Kuhnian sense. I assume readers are sufficiently familiar with the concept so that I need not explain it. And, I am aware that I am stretching the meaning, since here I am not talking about working within a scientific paradigm, I am talking more about whether those working at “normal” science reflect on the origins of their science (or the particular paradigm they are working under) and how their work might or might not be validated by the original purpose of the paradigm they are working under. The issue here is that working at “normal science” is productive, is efficient, and is what living, productive scientists actually do. (Of course, there are critiques of these notions from, e.g., Popper and others.) It would be inefficient if those practicing at normal science had to constantly re-create the origins of their science (or the origins, at least, of the particular paradigm they work under) and re-think the fundamental purpose of what they do. It would be paralyzing.

A useful analogy can perhaps be drawn to the field of economics but more specifically to the study of National Income Accounts (the equivalent of the psychometrics of assessment is probably closer to National Income Accounting and measurement, not econometrics as such, as is often thought). The most authoritative and comprehensive guideline and set of standards on
National Income Accounts is the UN Statistical Division/European Commission / IMF / OECD / World Bank’s “System of National Accounts 2008.” (UN, 2009). This 722-page (!) document goes on at exquisite length about every conceivable aspect of calculating all the key macroeconomic aggregates that are useful in tracking and managing economies. But nowhere does the document address the notion of how the accounts help countries grow, or the notion that the system of accounts originated, precisely, in helping countries get out of a “bottom of the pyramid” issue or set of issues in the Great Depression, or in competition with the Soviet Union as a way to quantify the relative position of Western economies vis-à-vis the Soviet Union. Even the 1947 (first undertaken by the new UN) system had only one or two brief and very “thin” paragraphs situating the methods in context (UN 1947). One would have to go back to the scholar/technocrats who originated the science, such as Simon Kuznets, for a clear explanation of what the accounts were “for,” that is, speaking loosely, what started the paradigm shift that later became the “normal science” of making tweaks on the national income accounts systems: “The economic changes that occurred in this country during recent years are sufficiently striking to be apparent to any observer without the assistance of statistical measurement. There is considerable value, however, in checking the unarmed observation of even a careful student the light of a quantitative picture of our economy. How extensive was the contraction in the volume of economic activity, year by year, from the peak attained in 1929? What was the impact of the current depression upon the various industrial branches of the economic system?” (Kuznets 1934, p. 1). Scholars who study the history of measurement in economics are clearer, but they are not “the” experts on the methods: “Modern national income accounting was designed in the early 20th century for the purpose of providing improved indicators about the performance of the economy so that government policy makers could better control the economy” (Holcombe, 2004).

As in economics and national income accounting, so it would seem, in psychometrics and standards on testing: the science is “innocent” to some degree of its historical purpose, even if very aware of its current purpose, and even if “meta” scholars do focus on the purpose: psychometricians tend to “just do it” while educators tend to worry about purpose. This is normal, but it does not satisfy if the aim is to get moving at the bottom of the pyramid.

A set of practical approaches

In practice, then, what seems to happen in response to the need to measure at the bottom of the pyramid, so one can move the bottom of the pyramid?

How can we be practical, and focused on purpose, while at the same time paying respect to rigor?

It seems that, at this point, assessment experts or providers are playing a particular maximization game: maximize practicality, subject to rigor constraints, as opposed to maximizing rigor subject to practicality constraints.

But assessment experts have approached it from two angles: a) some, typically associated with the large international assessments such as TIMSS, PIRLS and PISA, have taken an approach of walking “down” their assessments towards more usefulness at the bottom of the pyramid,
sacrificing as little rigor as possible, and b) some, typically associated with efforts such as ASER, EGRA, and others, have taken the approach of doing something that they see as very practical, and have been gradually building more rigor into their approaches.

And they both do so by consulting the experts, but not simply following the experts uncritically, given the experts’ normal concern with “normal science.”

Relevant questions to be answered in the next section include:

1. How precisely are the classical assessments adjusting? What is the evidence they have used, based on the “normal science” of the psychometric profession?
2. How are, on the other hand, the newer assessments, which started out as useful to the bottom of the pyramid, trying to increase their rigor?
3. What about mixed approaches such as the proposed Global Metric?
4. Finally, what else needs to happen before the research and technocratic community can offer the countries at the bottom of the pyramid tools that can help them move?

“Bottomization” with rigor, “rigorizing” the bottom, “even faster and smaller,” and other options

“Useful for the bottom” adjustments to classically rigorous measurements

The evidence that the classically rigorous assessments have been making adjustments, and that these adjustments are done with reference to strong “ask the expert” influence, is considerable. One also suspects that there are limits and tradeoffs involved that may in the end require supplementation, rather than simply making these classical assessments fit broader and broader needs. But more importantly one suspects that addressing the limits in one direction, might make the assessments less robust in another, or less robust overall (unless time—number of items—were to be increased?). Perhaps these are the concerns of a trespassing economist, and the psychometricians can show my concerns to be misplaced.

My review is not exhaustive for any of the international assessments, and it does not include the regional assessments, for lack of time.

Some TIMSS and PIRLS data

In discussing results of PIRLS 2009, and looking towards the future, IEA-related experts (Martin et. al. 2014) clearly stated that: “… for countries with lower performance levels, below 30 percent correct, information produced by the scaling process indicates that the PIRLS is not a good fit for many countries, and that the achievement results produced by the PIRLS scaling may not be reliable indicators of the average level of reading in the country. In the 2006 PIRLS, there were 4 countries with less than 30 percent correct scores, on average: South Africa, Morocco, Kuwait, and Qatar. But these are hardly at the bottom of the pyramid. If PIRLS is inadequate for these countries, it would be all the more so for the countries really at the bottom of the pyramid.

The strongest form of evidence has to do with the capacity of the scaling approach to pinpoint performance of students at the lower end of the proficiency scale, in comparison to students at
the middle of the scale. The results can be most dramatically illustrated comparing Russia (average score 565) and South Africa (average score 302). Figure 5 shows, on the horizontal axis, students’ skill as measured by the “first plausible value” of each students’ proficiency score and, on the vertical axis, the standard deviation of a sampling of students’ “explained” proficiency after taking into account their responses on the items and their background. In some sense, it is a measure of the ability of the model to explain students’ proficiency based on the first plausible estimate of the proficiency and their background characteristics. As can be seen comparing Russia and South Africa, the model performs quite badly in South Africa, particularly at the lower ends of proficiency (that is, to the left on the horizontal axis). It is also clear that the accuracy of estimation is also worse, in both countries, to the right. That is, the model tends to do better at the middle of the distribution within countries. Note that in the figure it is the height on the vertical axis that matters (the size of the standard deviation), not the spread of the points (that is, one is not interpreting the standard deviation from the height of the scatter at any given point on the horizontal axis).

Figure 5. Accuracy of PIRLS Scale in two countries, and at different levels within each country

Source: Martin et. al. 2014, Figs. 1 and 2.

An interesting case comparing the Test Information Function in TIMSS countries to SACMEQ countries (using only one of his linking methods) is provided by Sandefur (2016). The information content of TIMSS, if applied to SACMEQ countries, would be very poor indeed.
But, there is another important issue. Martin et. al. say that concerns arise when the percent correct is below 30%. Yet the countries in the PIRLS database that they studied contains only four or so countries in that range, and they hardly typify the true bottom of the pyramid. I do not know whether that 30% rule applies to TIMSS. I will assume it does. I then correlated the percent correct in TIMSS 2015 with GDP PC, and extrapolated to the many countries that did not partake in TIMSS 2015. Naturally, this is a mere extrapolation, with all the dangers that entails, and so on, but it is strongly suggestive and not illogical. (Note: I removed the low-results, high-wealth oil-producing countries from the sample as they are clearly not part of a pattern.) The results are shown in Figure 7.
So, one wonders: the TIMSS Numeracy and Literacy adaptations seem to have been done “in the spirit of” analyses done with the four or five countries that tend to under-perform, based on a cut-off of about 30%. But there are a lot of other countries. After the more recent rounds of PIRLS and TIMSS, that include easier items, it will be interesting to re-do similar analyses with the PIRLS Literacy and TIMSS numeracy data, to see whether with easier items but poorer (or more poor) countries participating (e.g., Botswana and Colombia in 2011, Trinidad and Tobago in 2016), the problem has, in some sense, simply been displaced down.

Further, how feasible is it to extend it down to countries that might be expected to perform only at the level of 20% to 25% correct on average?

One other item is important to note. If the idea is to extend the accuracy of the scaling down by including easier items in PIRLS Literacy, then will this further increase the inaccuracy at the upper end? The figure makes it clear that the standard deviation increases at the upper end of the scale even in non pre-PIRLS (or non PIRLS Literacy) in 2006, particularly in South Africa (which shows a much greater range of proficiency than Russia). If one injects easier items, thus necessarily lowering the number (or proportion only?) of more difficult items, will this compromise the accuracy at the upper end both within countries and between countries? It would seem so, but perhaps that is a matter for those more expert than I do decide. However, if I am
right, then the question arises as to whether there are limits to the idea of more and more global measurement, and whether, therefore, a variety of approaches is needed.

Some PISA data

PISA seems to have undergone similar, perhaps even deeper introspection, in moving towards PISA for Development. And, the reach into the bottom of the pyramid seems to have been somewhat deeper, as it included analysis of regional results from India.

The data are very clear that the traditional PISA analyses were not very accurate for the worst-performing countries. Adams and Cresswell (2016) show the performance on PISA 2009 items for the OECD average and for a state in India. The data show the percentage correct for a variety of items in the two geographies. If the data were gathered around the diagonal line, and gathered equally well around that line, it should show not only that the results are about the same in both geographies, but much more importantly that the data are equally reliable in both geographies. The comparison here suggests that items that are relatively easy for OECD children have a huge variance of difficulty for children in this Himachal Pradesh. That is, the variance for this state, at 80% correct for OECD, is huge, and that for items where OECD scores about 60% correct, the percent correct in Himachal Pradesh seems to average about 20% correct. If the reasoning used by PIRLS can be applied, it suggests that for these sorts of items the measurement in Himachal Pradesh would be very inaccurate. So, not only is the average performance in Himachal Pradesh much lower than in the OECD, but the relationship is extremely heteroskedastic, perhaps placing into question the possibility of drawing firm comparisons about the levels of performance in both places.

Figure 8. PISA 2009 results in OECD and Himachal Pradesh, India

They also show that for the less well-performing countries, the amount of information about the bottom of the pyramid, in those countries, is very weak and that, if one were to focus on the easiest items, the amount of information about the bottom of the pyramid increases substantially. This can be seen in Figure 9, where the proficiency distributions and the information distributions are graphed together.

Figure 9. Distribution of latent ability and distribution of information, for all of PISA released items and for the easier items, Kyrgyzstan


The fit for the easier items is clearly better. The recommendation then could then be to focus on easier items. However, even so, Adams and Cresswell observe: “[Even—my editorializing] If an easy subset of items that approximates the framework is selected from the secure pool it will remain more difficult than is psychometrically ideal for many developing countries – i.e. with the smallest possible measurement error. In other words the test will be mis-targeted.” And this is with Kyrgyzstan, hardly at the bottom of the world’s pyramid.

In the case of PISA, a lot of thought has also gone into how to deal with children not in school and in adapting the background information for the bottom of the pyramid. We don’t go into that for lack of space.

In conclusion, it seems clear that the traditionally highly rigorous international assessments have thought hard about the issues. However, one has to ask whether, if one is really talking about the bottom of the pyramid, the adjustments go far enough. Or, more importantly, whether they really could go far enough, without hurting the accuracy of global measurement. It could also be that at the very bottom of the scale, nothing one could do would provide much information. In the discussion of PIRLS noted above, Martin et. al. estimate that the percent correct in PIRLS that is achievable with just guessing is around 15%. Yet many, many countries (using TIMSS data) are only a bit above that. Can performance within those countries, or for comparisons between those countries, be measured at all using the sorts of techniques used by TIMSS and PIRLS, not to mention the risk of reducing accuracy at the other end of the spectrum?
Rigorization of measurements designed for utility at the bottom of the pyramid

Tools such as EGRA and ASER were explicitly designed for the bottom of the pyramid. LAMP was designed to at least be sensitive at the bottom of the pyramid, by including in the assessment sections or modules that focused on the “components” of reading (or precursors or pre-reading, call them how one will). Over time, as their popularity increased and as development agencies came to see some value in these tools, an interest in their psychometric characteristics seems to have increased.

The tools are typically also applied (except in the case of LAMP) with a greater focus on earlier grades. So, even though they are designed with the bottom of the pyramid in mind, and are therefore much easier, they are also applied typically in grades 2 and 3, and therefore in some sense have to be even easier.

However, it is clear that they are not easy enough, in many cases. The number of cases where children cannot read any words at all can range as high as 90%. Average fluency levels in reading connected text can be as low as 1, 2, 3 or 4 words per minute, or, naturally, zero. One would therefore expect many of the same problems that, say, PIRLS has with accuracy when there are just too many incorrect answers. It is hard to meaningfully distinguish between degrees of incorrectness when incorrectness is so pervasive. (And, one has to ask, would one really want to? What exactly is the point?)

Exploration of “rigorization” of EGRA and similar tools

Be that as it may, efforts such as EGRA that focused initially (or were inspired by) on the pedagogics of reading (RTI International 2016) and ways to measure particular components, have more recently been subjected to the standard sorts of psychometric analyses. Before summarizing some of those findings and their uses, however, one also has to wonder who appropriate this is, at all. In EGRA, or ASER, for instance, what is an “item.” I would argue that in EGRA, conceiving of particular items such as specific letters or words as “items” and applying techniques such as Item Characteristic Curves (ICC), is, at the very least, odd. I would claim that, compared to items in traditional assessments, EGRA items are “degenerate.” It is hard to think of what could be the meaning of an ICC for the letter “g.” Or the word “dog.” Is one really expecting that the ability to read such things is related, in any meaningful way, to a useful construct of overall, latent proficiency in reading, and that this relationship can be characterized by an ICC? I sense that there is an inappropriate application of technique: if what one does for a living is IRT analysis, one may as well do it to everything, without stopping to think about fundamental appropriateness. The point here is not whether one can “technically” or mathematically, indeed, fit, say, ICCs using EGRA “items.” The point is, what’s the meaning?

As an example: in typical IRT analysis comparing the proficiency distribution and the item difficulty distribution, part of the purpose is to decide which items are redundant, because they are too “stacked up” at certain points in the proficiency distribution. A typical case in point would be found in the analysis of EGRA results in letter identification in the Wolaytatto language in Ethiopia as shown in Figure 10, panel A.
Figure 10. Wright map for letter recognition in Wolaytatto language, Ethiopia 2014; simpler presentation (panel A), and simple histogram of letter recognition skills (panel B).


As an example, then I would question the (explicit or implicit) conclusion that some items are redundant simply because many are stacked around the same latent ability level. If the purpose,
the ultimate aim, is to understand fluency in letter recognition, then in fact one needs many items—that’s the very essence. The point is not (or not so much) to try to understand the relation between a fairly sophisticated notion of latent proficiency (and, of overall reading proficiency or proficiency in letter recognition?) and the recognition of particular letters. With many of these pre-reading skills the response is a rapid yes or no: not something that requires much thought, in the sense that a comprehension item might. That is the utility of the notion of proficiency in recognition of particular letters is somewhat suspect, since what matters is the proficiency and automaticity in the recognition of lots of letters (100 in the typical EGRA application), some of which may be repeated in the assessment (in fact, they are, in this assessment, as they are in the examples given in the basic EGRA toolkit—in the latter, the letter “a” is repeated 12 times: clearly, there is no sense in which reading the letter a more than once carries information about the ability to read the letter a). “Proficiency in letter recognition” is not the same thing as “Proficiency in recognition of particular letters.” In my view, this is a “category mistake.” In that sense, the “items” in assessments such as EGRA may be not usefully construed as particular letters or words, but as the overall sub-task of, say, fluency in letter recognition. And for this, IRT analysis seems misplaced. This is not to gainsay that some useful information about trouble in dealing with certain letters may not be derived from analysis; but it is to note that classical analyses of the percent of students who are having trouble with particular letters may be enough; and it is also to say that overall analysis of the letter sub-task, of a more “classical” nature, is probably more appropriate. The simple histogram in Panel B of Figure 10 shows the distribution. And the fact that children in Grade 2 here are about as fluent as Grade 1 children are expected to be according to some developed country benchmarks (but not more), suggests that the sub-task overall is not necessarily too easy, as the Wright map might lead one to think (since so many specific letters are “too easy”).

It may be more appropriate to consider sub-tasks such as fluency in letter recognition, fluency in non-word recognition, and to ascertain the internal correlation between the overall performance in each of these sub-tasks (that is, fluency over 100 words). This is done (among many examples which could be cited) in Jimenez, Gove, Crouch, and Rodriguez (2014) and parts of that analysis are shown in Figure 11.

---

14 The tendency to view each specific letter or word in an EGRA sub-task such as letter or non-word decoding (which may include 100 letters or non-words) may be what is behind the oft-repeated notion that Cronbach’s alpha may not be an appropriate measure of reliability. That may well be true for calculating the reliability of particular sub-tasks, where each letter or word is treated as an “item.” But we have noted that this may be the wrong perspective. If that’s the case, then the reliability of the overall EGRA may well be characterized by a measure, such as Cronbach’s alpha, based on multiple correlation and the correlation structure of the various sub-tasks.

15 See http://terpconnect.umd.edu/~dlspeece/cbmreading/norms.html.
The analysis reveals that most of the tasks (true items!) are internally highly correlated, while some are not. Were one to want to seek parsimony, then one could get rid of the (few) sub-tasks that are not very correlated with the others (such as oral comprehension). Furthermore, it is clear from the component loading that some tasks don’t load as well as others. Again, this could be used to increase assessment parsimony. The first component explains 65% of the variance and, for the factors that matter, the loadings are high and relatively uniform. All this suggests that there is an underlying or latent construct that could be called early (or precursor) reading ability, and also suggests ways to measure this even more efficiently. The paper does not carry out a Cronbach-type analysis of all the sub-tasks (or items) together, which we have noted above may be the most appropriate thing to do. In most EGRA applications this tends to be around 0.8 or better (RTI International 2016).

One factor that characterizes the classical assessments is the possibility of sorting out proficiency levels that are themselves based on the results of the testing (the estimated latent proficiency levels), and understanding of the items in the framework, and of the performance of the students against those items. This has not been done (that I know of) for EGRA. Instead, EGRA relies on the estimated proficiency in fluency needed to reach certain comprehension levels, or in comparison (and adjustment) of normative levels reached by children in other countries, but it is not clear that either of these are the best approach. More work in this area may be good. One

### Table 1

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Knowledge of letter sounds</td>
<td>1</td>
<td>.577(***)</td>
<td>.514(***)</td>
<td>.554(***)</td>
<td>.583(***)</td>
<td>.220(***)</td>
<td>.999(***)</td>
<td>.551(***)</td>
</tr>
<tr>
<td>2. Reading of familiar words</td>
<td>1</td>
<td>.929(***)</td>
<td>.956(***)</td>
<td>.700(***)</td>
<td>.217(***)</td>
<td>.842(***)</td>
<td>.812(***)</td>
<td></td>
</tr>
<tr>
<td>3. Reading of pseudo-words</td>
<td>1</td>
<td>.905(***)</td>
<td>.655(***)</td>
<td>.201(***)</td>
<td>.777(***)</td>
<td>.766(***)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Oral reading fluency</td>
<td>1</td>
<td>.686(***)</td>
<td>.205(***)</td>
<td>.869(***)</td>
<td>.815(***)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Reading comprehension</td>
<td>1</td>
<td>.575(***)</td>
<td>.195(***)</td>
<td>.751(***)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Phonological awareness</td>
<td>1</td>
<td>.220(***)</td>
<td>.627(***)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Oral comprehension</td>
<td>1</td>
<td>.226(***)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Dictation</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Component</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading of familiar words</td>
<td>.961</td>
<td>.032</td>
</tr>
<tr>
<td>Oral reading fluency</td>
<td>.059</td>
<td>.027</td>
</tr>
<tr>
<td>Reading of pseudo-words</td>
<td>.017</td>
<td>.004</td>
</tr>
<tr>
<td>Reading comprehension</td>
<td>.093</td>
<td>.119</td>
</tr>
<tr>
<td>Dictation</td>
<td>.057</td>
<td>.087</td>
</tr>
<tr>
<td>Phonological awareness</td>
<td>.703</td>
<td>.318</td>
</tr>
<tr>
<td>Knowledge of letter sounds</td>
<td>.589</td>
<td>.255</td>
</tr>
<tr>
<td>Oral comprehension</td>
<td>.060</td>
<td>.564</td>
</tr>
</tbody>
</table>

Source: Jimenez et al. (2014).
could take inspiration from either the LAMP process, which uses Latent Class Analysis (among other inputs) to figure out proficiency levels when only “Components” of reading are assessed, or from ASER, which uses explicit proficiency levels based on assignment by the enumerator into “Paragraph,” “Story,” “Sentence,” and other levels. It would seem that this kind of ability to classify into proficiency levels would be important for purposes of international reporting, more than for purposes of country-by-country movement at the bottom of the pyramid. Furthermore, even for global reporting, more work would be done on the accuracy (and global validity) of classification in these proficiency levels, since they are not based on IRT and, we would argue, IRT is probably an inappropriate paradigm to use for these “Components” assessments. (And noting that LAMP, in its assessment of the proficiency levels in the “Components” of reading, does not use proficiency levels that are comparable across countries, though it is not clear that, with some effort, this would be impossible. The real question is how meaningful it would be, given the characteristics of the different languages and that, at the “Components” level, these characteristics affect the Proficiency as much as policy or pedagogy does, if not more.)

Since its inception, EGRA has also attempted to increase rigor in many other ways, such as working on, and reporting, Inter-Rater Reliability (as have ASER and other citizen-led assessments). Such measures are quite high (Results for Development Institute 2015).

Whether such attempts to increase rigor helps the countries where EGRA has been applied move their “bottom of the pyramid” children is a separate matter. It seems that EGRA, and similar tools such as ASER (see below), or ASER-like tools for placement of children for remedial programs, has been helpful in improving learning levels, when the assessment is clearly linked to a policy or pedagogical response (see Kim et. al. 2016), though thus far the evidence is more about pilot programs than about scaled-up ones. Evidence about the utility of this kind of measurement will tend to be forthcoming as scaled-up efforts start to be implemented more widely, as in the Tusome effort in Kenya. (But using citizen-led assessment merely for reporting and awareness raising seems to have limited impact—see below.)

“Rigorization” of ASER and “citizen-led” assessments

ASER (Annual Status of Education Report), created by Pratham in India, was originally created for two purposes. First, as an attempt to catalyze attention and action about severe problems of learning outcomes at the bottom of the pyramid, and less as tool for rigorous measurement. Second, ASER-like tools are used to classify children simply, in order to create ability rather than age- or grade-groupings, for purposes of remedial intervention of the “teaching at the level where the child is” type—a type of intervention gaining global recognition as quite promising (see David Evans blog on the subject16). These “creation stories” are important, since they are very close to a “move, not just measure, the bottom of the pyramid” motivation. As with EGRA, perhaps, as the tools acquired a reputation as doing fairly well what they were designed to do, various commentators, and Pratham itself, began to wonder about its sampling, psychometric, and other properties. Furthermore, as the use of such tools began to spread and spawn “sister” assessments such as Uwezo in East Africa, the natural curiosity about the level of rigor, and the desire to increase the level of rigor, but in a manner consistent with original purpose, increased.

Attempts have been made by ASER itself to understand ASER by comparing it to fluency assessments. Vagh (2013) and Vagh (2009) estimated reliability and concurrent validity in various ways. The results were generally quite good. Test re-test correlations in the order of 0.9 to 0.95 were found. Inter-rater reliability in some of these earlier measurements was also estimated as high, suggesting that “even” volunteers, using the ASER assessment, could reach high level of reliability. Later evaluations (Results for Development Institute 2015) summarize reliability and validity of citizen-led assessments such as ASER itself, as well as Uwezo. The findings suggest that concurrent validity between ASER and Uwezo with EGRA and EGMA (or with other assessments deemed to be reliable) is generally quite good, but the reliability of Uwezo was generally not as high as for EGRA or EGMA (Results for Development Institute 2015, pp. 29 and 30). Inter-rater reliability is also found to be high in many of these assessments, showing that volunteers can work the tools appropriately (Results for Development Institute 2015, p. 31).

While rigor appears fairly fit-for-purpose, there are useful remaining questions.

First, do citizen-led assessments, by themselves, help move things at the bottom of the pyramid? The most in-depth evaluation of the impact of these assessments (Results for Development Institute 2015) clearly indicates that the answer is a strong “No.” Comparisons are made, for example, between Uwezo, which did not have a remedial, intervention, or reaction response, and Pratham’s intervention model(s) which use the awareness created by ASER, as well as ASER-like tools to group children by ability in “teach where the children are” (Results for Development Institute 2015, p. 40). Or, between two treatment arms in the EGRA-Plus intervention in Liberia, one of which simply provided information and had almost no impact, and one of which provided lots of pedagogical support, and did have impacts (Results for Development Institute 2015, p. 41).

Second, there are many extant suggestions for improving the rigor of these citizen assessments (at least some of which would perhaps also apply to other assessment initiatives that started as bottom of the pyramid efforts, such as EGRA), such as those coming from Mendelovits, Munro-Smith, Murphy, and Waters (2014). These include ideas worthy of consideration, such as formalizing comparability of results over time (which has not been seriously and regularly done, it appears, for citizen-led assessments, as far as I have been able to determine, and which is undeniably an absolutely essential component for tracking movement at the bottom of the pyramid), standardizing and reporting on population definition and coverage, improving sampling procedures, improving data management, and so on. However, sometimes these evaluations seem de-linked from original purpose: what is most useful for moving things at the bottom of the pyramid? In that sense they seem more like what we have characterized as an “ask the expert” approach, where such experts are devoted to “normal science” (efficiently and professionally so, but working within the boundaries of a particular paradigm). More discussion could perhaps take place around which of these sorts of recommendations increase rigor at little cost in terms of resources or time (thus keeping things quick and cheap, relative to purpose), and which might actually increase fitness-for-purpose, and which others might not. As an example, consider the use of volunteers. It seems fairly logical for experts practicing at normal science, within a given paradigm, to question whether volunteers should be used. The research reported
by the Results for Development Institute (2015) draws attention not only to the fact that volunteer use may not require a reduction in rigor, but may be a fairly essential part of the effort, given its fundamental, “non-normal” purpose of mobilizing and galvanizing action.

The experience of LAMP

LAMP (Literacy Assessment and Monitoring Programme) was an initiative of the UNESCO Institute for Statistics to provide richer descriptors of literacy than a simple binomial “Yes/No” as is traditionally derivable from either self-declaration in household surveys or census forms, or from simple evaluations of whether a respondent can or cannot read a sentence, typically in household surveys (such as DHS).

What can LAMP tell us about measurement that is responsive to the “bottom of the pyramid” issues, and for foundational grades? LAMP was aimed at adults (respondents 15 and older) who, presumably, had coursed through as many school grades as it was common to course, in their countries, when the now-adults were of school age. Thus, it is not a grade-oriented assessment, and much less is it oriented only at the early grades. In that sense it would appear not to have much to say. However, because LAMP was interested in a range of literacy abilities, and because some of the respondents’ skills were too low for them to be assessed using more traditional methods that emphasize comprehension (and that could also be subjected to IRT analysis and the development of standardized proficiency levels), LAMP added a “Components” section. These are skills relatively similar to those assessed in tools such as EGRA: they emphasize letter, word, and sentence-level skills. They do not allow for assessment of detailed levels of comprehension, with the possibility of developing fine-grained proficiency levels for comprehension. So, in that sense, because these skills matter, particularly in the foundational years and particularly in poor countries or in the poorer sections of those countries, LAMP’s experience may carry insights relevant to this paper. A few lessons seem to be derivable from LAMP’s efforts on “Components,” that might be appropriate to bottom of the pyramid measurement and movement, and to global comparisons.

First, LAMP concluded that global comparisons on the “Components” sections were not possible or were, at any rate, too difficult to create and might not have been reliable. Proficiency scales were not common to the four countries where the “Components” were reported.

Second, LAMP also concluded that the skills in the “Components” module could not be placed on the same proficiency scale as the more advanced comprehension skills the other reading modules (whose assessment uses tasks similar to those used by PIRLS or PISA, and are more easily susceptible of IRT analysis and the creation of rigorous and comparable proficiency scales). Whether this is an inherent problem due to the lack of uni-dimensionality in the “Components” skills, or because of the difficulty in considering the “Components” skills as even having true “items” (see discussion of EGRA above) is not clear and may need further research. But, in any case, if it is true, then using these “Components” proficiency levels for global comparisons may be difficult. However, this should not compromise the ability to use these proficiency levels, or the idea of “Components” for national policy and policy tracking. (Assuming that the proficiency levels can be rigorously established, i.e., are reliable.)
Third, if LAMP, for its “Components” section, is able to establish ex-post proficiency levels taking advantage of Latent Class Analysis techniques and other information (and ASER doing it ex-ante), perhaps EGRA can do so. Again, the purpose may not be so much international comparability, but helping countries move at the bottom of the pyramid.

Even smaller, quicker approaches

Current UNICEF plans (Cardoso, 2016) to include some measure of learning in their standard Multiple Indicators Cluster Surveys (MICS) (undertaken every several years in many countries—a venerated household survey ongoing now for several decades, with, often, several rounds in each country) are the “smallest, quickest” that we know of, and that have some promise of comparability, as well as measuring at the bottom of the pyramid, since the measurement is based on household surveys and children are included whether in school or not. The age range is an unusual factor: ages 7 to 14. By including such a broad age range (and presumably also asking about the grade of the child), the surveys will allow one to determine grade profiles for learning over a large number of grades. This would promise to be extremely useful in policy dialogue and awareness-raising. Given that MICS is a very long and complex survey, a great deal of emphasis is being placed on “getting in and out quickly” with each child: the aim is for 17 minutes of interaction with the child. The “items” assessed are inspired on ASER/UWEZO, Save the Children’s Literacy Boost, as well as EGRA and EGMA. These are all assessments whose original intent was to measure at the bottom of the pyramid. They are, importantly, I would argue, not based on “items” in the classical TIMSS, PIRLS, or PISA sense, and hence one hopes that the eventual analysis is sensitive to this fact. The items are planned as including reading skills and number skills. Among the former: oral reading accuracy (story, 70 words), literal comprehension (3 questions), and inferential comprehension (2 questions). Among the latter: number reading (6 items), number (size) discrimination (5 items), addition (5 items), and pattern recognition & completion (5 items) (Cardoso, 2016). Early indications are that reliability, even for such as “small, quick” assessment, as measured with classical statistics, will be reasonably high—high enough for purpose.

Hybrid approaches

In developing measurement that is useful for global reporting, but that is also sensitive to the bottom of the pyramid measurement issues, several options have been put forth. While the idea of recommending a single assessment has been sometimes put on the table, it is unlikely that international agencies would want to take on the task of selecting a current assessment (or a modification of one) as “the” standard for reporting. The Center for Global Development has proposed a new assessment for global comparison purposes as well as for national policy dialogue purposes, similar to PISA, but at age 9, a notion that has been picked up by the Education Commission’s (International Commission on Financing Global Education Opportunity 2016, p. 56) report on The Learning Generation, where it is proposed that “The Commission recommends that the international community agree on a lead global learning indicator [though, presumably, not necessarily a single assessment, though the elision is often easily made—author’s editorializing] to focus national and global efforts on learning and not just participation.

---

The indicator should be based upon the learning and skills expected by a given age, rather than by school grade.” It is hard to forecast whether this will be acceptable to all important stakeholders. But partial efforts that could either lay the foundation for a single indicator (based or not on a single assessment) or simple serve as reference points for various indicators and assessments, are worth looking into.

**Linking**

One possibility is to link existing assessments via common items, with a focus on the earliest grades available in such assessments (because of the relevance, for the bottom of the pyramid, of learning in the foundation years). This was proposed (probably not for the first time) as a particular option requiring common items for the next round of assessments for these actors, at a meeting between GPE, UIS, IEA, and the three extant regional assessments (LLECE, SACMEQ, and PASEC) in 2012. This particular incarnation of the idea did not progress, but the idea of creating linkages did not die. Recent work by Hastedt and Desa (2015), however, suggest that this is more easily talked about than done. They studied the linking error by taking released items from TIMSS 2011, and embedding them in simulated national assessments for Botswana, Honduras, and Tunisia. What were, to me, the most telling results are reproduced in Figure 12. The results are particularly important for debates about the bottom of the pyramid. The errors in the proportions of students in the “High” and “Advanced” proficiency categories are not very large, in part because there are so few students in those categories. But the classification errors at the “Low” proficiency level are huge and highly statistically significant (that is, not produced by sampling error but by the nature of the linkages themselves). In Honduras the “true” percentage at the “Low” level is 49 (based on TIMSS itself). But even a 30-item linkage between a national assessment and TIMSS would have produced difference of 12 percentage points (but roughly 24%--that is 12 over 49). In Tunisia, the difference was about 20 points (but roughly 57%, or 20 over 35!).
A recent attempt by Sandefur (2016) to link across two assessments, namely TIMSS and SACMEQ, using a set of common items, finds contradictory results depending on the linking method used. As far as I can tell the paper, unlike the paper by Hastedt and Desa (2015) does not produce confidence intervals on the scores created by “projecting” the SACMEQ countries onto the TIMSS scale. But the paper nevertheless comes to conclusions that linking, at least of this type, is unlikely to produce reliable estimate for countries that are close to each other. Using one of the methods leads the author to conclude that “In Botswana, a SACMEQ score of 400 is just below the eighth percentile, which is equivalent to a 279 on the TIMSS scale. In South Africa, the same SACMEQ score would be in the twentieth percentile, which is equivalent to a score between 180 and 190 on TIMSS. So the predicted TIMSS score for children with a given SACMEQ score differs by roughly a full standard deviation, depending on which country is used to make the link” (p. 12-14). And the linking seems to be worst for the lowest-performing students, as can be seen in Error! Reference source not found. Figure 12. Classification errors in linked tests, based on TIMSS analysis.

| Table 6. Percentages (standard errors) of students reaching the international benchmarks |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|
|                                | Low (1.6)       | Intermediate (1.7) | High (1.1)     | Advanced (0.1)  |
| BWA (TIMSS)                   | 60              | 29              | 7               | 0               |
| BWA (10 Items)                | 68              | 35              | 11              | 2               |
| BWA (20 Items)                | 55              | 22              | 4               | 1               |
| BWA (30 Items)                | 54              | 21              | 4               | 1               |
| HND (TIMSS)                   | 49 (2.5)        | 29 (2.1)        | 7 (0.8)         | 0 (0.1)         |
| HND (10 Items)                | 64 (1.5)        | 32 (1.4)        | 9 (0.9)         | 2 (0.4)         |
| HND 20 Items                  | 36 (1.5)        | 9 (0.9)         | 1 (0.3)         | 0 (0.0)         |
| HND (30 Items)                | 37 (1.5)        | 10 (0.9)        | 1 (0.3)         | 0 (0.1)         |
| TUN (TIMSS)                   | 35 (1.8)        | 11 (1)          | 2 (0.3)         | 0 (0.0)         |
| TUN (10 Items)                | 70 (1.3)        | 38 (1.4)        | 13 (0.9)        | 3 (0.4)         |
| TUN (20 Items)                | 39 (1.4)        | 10 (0.8)        | 1 (0.3)         | 0 (0.1)         |
| TUN (30 Items)                | 54 (1.4)        | 20 (1.1)        | 4 (0.5)         | 0 (0.2)         |

Source: Hastedt and Desa 2015, p. 9.
While the linking methods seem to incur considerable error in allowing comparisons at the bottom of the scale, the data presented in the two papers suggest that comparisons between the worst-performing and the best-performing countries are valid. But this may not be good enough.

**Common scales and quality assurance procedures**

We seem to have arrived at the following conclusions so far:

No one has the stomach (for now) for accepting propositions for a single global assessment (new or by selecting an existing one). Even though there have been discussions around this area (see International Commission on Financing Global Education Opportunity 2016, p. 56), there seems to be no headlong rush to accept this proposition or even seriously explore its contours, at least not thus far.

Really rigorous linking between assessments seems impossible or would seem to require as much work as creating a whole new assessment, as it is to be feared from the linking discussion.

An assessment that is capable of discriminating and being useful for global comparisons both at the upper end of the proficiency distribution and at the bottom of the pyramid is unlikely at least within reason (as is strongly by the discussions of TIMSS and PISA above).

If these are true, then what is the alternative?

But, before addressing that last question, it may be worthwhile to review why these issues might matter for bottom of the pyramid considerations. Perhaps, after all, they do not. Perhaps it is good enough for each country to have its own scale and its own measurement, and to report globally on its own proficiency levels defined as national priorities suggest. The advocacy and accountability for improving is, after all, a matter of national responsibility, and a matter of dialogue and pressure between civil society, Ministries of Education, Ministries of Finance, etc. Yet, it seems that international comparisons are useful. Those of us who have worked on reading assessment in various countries are constantly asked: but what is a reasonable level of performance, what do other countries do? Furthermore, it is clear that most curricula are fairly globalized, with common expectations by certain grades. There is also the danger that completely unmonitored and un-standardized reporting by countries could be a race to the bottom. Finally, there is evidence that global statistics and benchmarking help civil society make demands on governments (e.g., movements around eliminating fees so as to increase coverage), and that having powerful indicators has helped in other sectors (e.g., the Under-Five Mortality Rate).

It also seems as though rigor is useful even if there were no reason to make global comparisons. Certainly, to measure progress, for instance, it is useful to have rigor around inter-temporal comparisons, and this imposes certain technical requirements on sample size, constancy in the nature of the sampling and population frames of the sampling, difficulty of assessment items, and so on. Technical rigor, in particular, in detecting movement among the poor, or at the bottom of the pyramid, is particularly useful, if the desire is to detect and track movement at that level.

Given that reasonable rigor is desirable even when the main aim is to improve the bottom of the pyramid, but that ultimate rigor is either undesirable or in fact impossible, what is a good compromise?

It would seem that even if countries were to report on their own assessments using their own proficiency levels, some kind of benchmarking of both proficiency levels and assessment quality control is a good idea. In this manner, countries can self-assess, in a minimally-interventionist model against globally comparable benchmarks for assessment administration and for proficiency. Or, in an approach that is a bit more interventionist, international agencies can use these tools so as to have instruments for deciding which reporting to accept or not, and for directing technical assistance. These tools can exist together with tools such as SABER Assessment, a World Bank tool for evaluating assessment systems.18

At a very minimum (that is, even for the least interventionist option) it seems one needs two tools, as noted: one for benchmarking against proficiency levels on some kind of common, global metric, and one for benchmarking processes of assessment management.

18 See http://saber.worldbank.org/index.cfm?index=8&pd=5&sub=0.
Based on discussions that have taken place in with and around UIS and ACER, a process that seems to make sense would proceed something like as follows:  

1. Establish a global framework of approximately sequential skills that can span the very beginnings of skills, all the way up to relatively advanced ones (say, up to the lower-secondary level: a level that is convenient because it is the limit of compulsory education for so many countries, and because it is a level after which knowledge branches out much more, especially in mathematics). This framework could be in terms of some posited universal sequencing of how knowledge is acquired, based on psychological principles or deeply pedagogical ones. But it seems that such principles are unlikely to be universal. A simpler expedient would be to base oneself on what are reasonably common curricular expectations around the world. If the issue is to generate movement at the bottom of the pyramid, ensure that proper attention is paid to the most fundamental skills even if these are “terminal” skills (no need—and no room—to continue to improve letter recognition skill once that skill is mastered!).

2. Find suitable examples of items from a variety of assessments that are already in use, and place them on the continuum. Ensure that the items are of high quality by assessing their performance in those assessments. Ensure that they are of approximately equal difficulty or placement relative to the “rung” in the continuum where they would purport to measure. Most likely, find a variety of items for each “rung” that purport to measure the same construct or skill.

3. Validate empirically in a variety of countries, including the poorer ones, so as to make sure that bottom of the pyramid issues are covered.

4. Set examples of benchmarks on the proficiency scale that countries could use, but without creating universal expectations.

It seems that such a tool would get away from the problem of having to have an assessment that is all things to all skill levels and all countries, which might be really difficult to agree on, politically, and might be technically impossible anyway. The tool, along with tools that can help evaluate the quality of assessment administration (including technical issues such as sampling), could be used by countries to self-assess, and/or by international organizations to decide which reporting from which countries to accept, and/or to provide advice to countries on how to benchmark themselves.

**Concluding comments, next steps**

It seems, in the end, that no single assessment approach or tool will do.

At the bottom of the pyramid, and for the foundation grades, one may need tools that are good at assessing the “components” of literacy, but without forgetting comprehension. But such tools would not be as useful for higher grades and/or countries where instruction is so good that children move beyond these basics almost effortlessly.

---

19 Most elements of this discussion can be found at the GAML 2nd meeting notes at: http://uis.openplus.ca/gaml/meeting-2.html.
Furthermore, there simply are certain analytical tasks that are much easier to do when assessments have the fine-grained ability that the classically rigorous assessments possess. The ability to distinguish, and do so rigorously, between many levels of proficiency on a single scale (that is, having a measurement of skill that is a continuous variable) and that can be related to both pedagogical and social factors, is useful. Yet, such tools tend to be unsuitable for very low levels of skill. Thus the inescapable conclusion that one needs a variety of tools.

But it also seems that the more this variety of tools can be placed within a coherent proficiency continuum, with examples of items, and placed within benchmarks, the better, as suggested by the Global Metric work taking place in GAML.

Research therefore could fruitfully continue on:

1. Creating or improving global tools for assessment administration evaluation and self-evaluation.
2. Creating and improving a global metric that is useful at the bottom of the pyramid.
3. Improving the rigor of tools that were originally designed for the bottom-of-the-pyramid tools, such as by adding more comprehension questions, and improving sampling and improving on time-invariance of difficulty.

But these are all either “let’s start by asking the experts” or “let’s do something practical and then ask the experts” approaches. I have argued that there may be real value on learning from the assessment practices of countries that have in fact “eliminated” the bottom of their pyramids in the past but in living memory (e.g., Korea), or are doing so. It may be that one will find little to generalize. Or it may be that assessment did not play that much of a role, in which case we will all have to be more humble.
References


