II. Quality, learning, and cultural comparisons: Trade-offs in educational policy development

Daniel A. Wagner

With the advent of the United Nations *Education First* initiative, and considering the continued efforts to focus on the quality of education in low-income countries, there has been a renewed interest in the improvement of learning (as distinct from school attendance) in poor and marginalized populations (Wagner, Murphy, and de Korne, 2012).[1] There is a large and diverse empirical research base in the area of human learning. Yet much of the available research is substantially limited by *boundary constraints* of various kinds. Most prominent among them is the limited ability to generalize from findings in one population context to other distinct population contexts. Similarly, research methods may vary greatly between one set of studies and another, making it difficult to discern whether the findings vary due to the methods or to other factors. These are classic problems in the social sciences, and inevitably lead to substantive trade-offs in how policy development takes place in education.

Skills and population sampling

If a learning assessment needs to be representative of an entire population of a country, and for multiple countries in a comparative framework, then time and money is likely to expand significantly. Up to the present, time and cost have been controlled by delimiting the range of *skills* that would be assessed (the *skills sample*), and by constraining the *population* that would be included (the *population sample*). These two forms of sampling need to be understood in terms of technical and statistical requirements, as well as policy requirements and outputs.

1. Parts of this paper were derived from the Brookings Report.

It is widely accepted that humans learn by sampling their environment, beginning with built-in senses from birth onwards. Clearly, no infant, child, or adult could possibly survive by taking in the totality of information available in the environment. In other words, human systems are designed to discriminate in order to sample for information that will be effective in handling learning challenges. Indeed, parenting and socialization that effectively prepare a young child to adapt, learn, and survive, involve exposing the child to the range of situations they will encounter in their lives. Not all these learning environments may be positive, but exposure to them will be important. When it comes to scientific research in general, and learning research in particular, humans also sample their informational environment, whether in educational institutions or via word of mouth or, increasingly, via internet search engines such as Google. The relevance of this relatively simple observation should not be underestimated, since one of the most vexing problems in learning research and evaluation is how to generalize from one sample population to another or, just as importantly, from one research study to another.

All research on learning depends on the sampling of a finite set of skills, and knowledge of the contextual situations in which they occur. Skills sampling can be done in the traditional paper and pencil fashion, increasingly through online methods (e.g. the OECD Programme for the International Assessment for Adult Competencies [PIAAC], or orally between the child and a testing enumerator (as in Early Grade Reading Assessments [EGRA]). In designing learning research and evaluation strategies, the choice of contextual and demographic variables (e.g. age, year of schooling, gender, supplemental educational services [SES]), the selection of skills to be assessed, and the type of research methodology are highly complex decisions. Each option is tied to a set of assumptions and compromises, and the selections included in the final research design will influence the validity, reliability, and practical feasibility of the chosen approach (see Braun and Kanjee, 2006; Wagner, 2011*a*). Furthermore, research designs need to be responsive to dynamic changes over time, and as expectations of literacy, numeracy, and higher-order

skills adapt to changes in social and economic environments, the measurement methods must also adapt to align with evolving educational goals.

Population sampling also matters. For example, roughly 95 per cent of the world population today resides outside the United States, while nearly 95 per cent of scientific publications on psychological development are based on American population samples (Arnett, 2008). Other studies have shown that, in the USA, more than 80 per cent of research on psychological development is based on 'majority' ethnic groups (European origin), while this population is only about 50 per cent of the current USA population (Arnett, 2008). These are not unique occurrences. Global research on learning parallels the above findings, since much of the research reviewed here is constrained in important ways by scientific datasets and research studies drawn from population samples living mainly within mid- to high-level income countries.

The area of population exclusions is more problematic. Gender has been a leading factor in school non-participation in low-income countries, although significant progress has been made in recent decades. Nonetheless, in the poorest countries, girls continue to be less present in school than boys, at the point of both primary and post-primary school entry. Systematic exclusion of girls in poor low-income countries usually results in lower participation in schooling among adolescent girls, as well as depressed scores relative to boys in national assessments.[2] Similar trends show differences in national assessments when comparing rural and urban areas in low-income countries. In some low-income countries, the difficulty of literally tracking down nomadic children can make their inclusion onerous for authorities (UNESCO, 2010).

Language variation from one ethnic group to another exists in nearly all countries. Many of these groups – sometimes termed 'ethno-linguistic minorities' – are well integrated into a national mix (as in Switzerland), but at

2. In the SACMEQ regional assessment in Grade 6, undertaken in 2007, Saito (2011) found that boys generally outperformed girls in mathematics, when averaged over 15 African countries, while girls outperformed boys in reading. However, national differences in gender disparities varied widely in both reading and mathematics.

other times they may contribute to civil strife. Often, social and political forces help to resolve differences, and usually include policy decisions resulting in a hierarchy of acceptable languages to be used in schools and governance structures. In such situations, whether in OECD countries or low-income countries, it is not unusual for children who speak minority languages to be excluded from learning research and assessments. This may be particularly problematic in regions in which civil conflict or economic distress leads to substantial cross-border migration, or in which immigrant groups (and their children) are treated as transients, and children are provided with little or no schooling. As noted earlier, differences in language, and increasing multilingualism, are among the most challenging aspects for improving learning in schools.

In sum, both skills and population samples vary, as do the learning processes (structured and informal) that individuals deploy, and the contexts (formal and non-formal) in which they take place.[3]

Methodological credibility

Research that can be converted into policy depends on its credibility, which means that well-trained scientists and experts can achieve consensus on the merits of a particular set of findings, even if they might disagree with the interpretation of such findings. The two most oft-cited components of learning science are validity and reliability.

The validity of any learning measurement tool or test is determined by the degree to which skills can be credibly linked to the conceptual rationale for the test. For example, do questions in a multiple-choice test really relate to a child's ability to read, or to the ability to remember what he or she has

3. There are also those stakeholders who *do* the sampling. Whether policy-makers, psychometricians, or local teachers, all come to the task of sampling skills and populations with their own experiences and points of view. Choices about which skills to sample among which populations, languages, and in which contexts, also add potential bias to an already complex set of sampling issues. In order to address such biases, researchers can use methods such as tailored sampling and subsample designs, matching samples, oversampling of marginalized populations, and mixed methods designs.

read earlier? Validity can vary significantly with context and with population, since a test that might be valid in London may have little validity in Lahore. A reading test used effectively for one language group of mother-tongue speakers may be quite inappropriate for children who are second-language speakers of the same language. With respect to international large-scale educational assessments, there have been a number of critiques of content validity around the choice and appropriateness of test items, given their application to local cultures and school systems (Sjoberg, 2007; Howie and Hughes, 2000).[4] While much learning research takes the form of quantitative testing, qualitative and ethnographic methods can also contribute, particularly with respect to cultural variation. Indeed, culturally sensitive research often requires qualitative approaches, given the uncertainty about learning processes in diverse contexts and the need to observe transitions between contexts.

Reliability is often measured in two quantitative ways. Generically, reliability refers to the degree to which an individual's results in a test are consistently related to additional times that the individual takes the same (or equivalent) test. High reliability usually means that the rank ordering of individuals taking a given test would, on a second occasion, produce a very similar rank ordering. A second and easier way to measure reliability is in terms of the internal function of the test items – do the items in each part of an assessment have a strong association with one another?[5] Of course, reliability implies little about the validity of the instrument, wherein agreement must be reached concerning the relevance of the instrument for educational outcomes. Considered in a qualitative perspective, reliability would be achieved when context-sensitive ethnographers, for example, agree on a set of observations of learning processes that they have independently gathered in a particular

4. Sjoberg (2007) claimed that some test items deviated substantially from the stated PISA goal of evaluating competencies for the workforce. Howie and Hughes (2000) found that the TIMSS covered only a very small fraction (18 per cent) of the curriculum of science in Grade 7 in South Africa, but as much as 50 per cent in Grade 8.
5. This is inter-item reliability (measured by Cronbach's *alpha* statistic).

context.[6] Considering that learning occurs in non-formal areas as well as formal ones, learning research cannot be limited to the sophisticated psychometric methods developed for formal learning sites, such as schools. Similarly, highly structured learning processes (guided by teachers) may be relatively easy to observe and monitor in the classroom, while informal (less structured) learning may be more difficult to determine and measure.[7]

Comparability of learning outcomes across contexts

Comparability is central to global education data collection, such as the large-scale data collection carried out by UIS. Nonetheless, if comparability is the primary goal, less attention is paid to the local and cultural validity of the definitions and classifications of learning, and therefore the data may become less meaningful and potentially less applicable at the ground level. This is a natural and essential tension between universalistic *etic*

6. 'Team ethnography' has become increasingly used in education research in the USA and Europe (see Blackledge and Creese, 2010; Bartlett and García, 2011).
7. The use of randomized control trials (RCT) is seen as one important way of increasing the credibility of research findings, by comparing interventions with control groups. Recent reviews by Kremer and Holla (2009), Banerjee and Duflo (2011) and Bruns, Filmer, and Patrinos (2011) assert the importance of this methodology for improving research designs in international development work. Other work (e.g. Burde, 2012; Castillo and Wagner, 2014) has begun to describe the limitations of the RCT approach in such settings.

   Another credibility issue is what constitutes a 'sizeable' impact. Traditional statistics emphasize, through inferential statistics, the notion of a 'significant' difference. In international development interventions, some prefer the use of 'effect size' as a way of measuring impact, since 'effect size' is a way of quantifying the size of the difference between two groups. For example, with work on EGRA reading assessments, the effect size (moving from 1–5 words per minute on an oral reading fluency test to approximately 30 words per minute), is not only significant, but may also have a very large effect size, indicating a large difference in mean scores. However, the credibility of this large impact also depends on the nature of the assessment itself. EGRA's use of words per minute seems to be a very malleable score, especially since many children in poor communities do so poorly at the outset when this measure is used. With other measures, such as reading comprehension, the research evidence suggests a much longer gradient to achieve a high effect size. See Paris and Paris (2006) for an overview of skill measurement trajectories. A related critique of EGRA concerns the prevalence of 'floor effects' on statistical results, especially on correlations between key variables; see Hoffman (2012) who also provides a broad-based critique of EGRA's use in low-income countries.

and context-sensitive *emic* approaches to measurement, and is particularly relevant to marginalized populations.[8]

Can both comparability and context sensitivity be appropriately balanced in learning research? Should countries with low average scores be tested on the same scales with countries that have much higher average scores? If there are countries (or groups of students) at the 'floor' of a scale, some would say that the solution is to drop the scale to a lower level of difficulty. Others might say that the scale itself is flawed, and that there are different types of skills that could be better assessed, especially if the variables are evidently caused by race, ethnicity, language, and related variables that lead one to question the test as much as the group that is tested. Yet having different scales for different groups (or nations) seems to some to be an unacceptable compromise on overall standards.

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

8. 'Emic' approaches are those that are consciously focused on local cultural relevance, such as local words or descriptors for an 'intelligent' person. 'Etic' approaches are those that define 'intelligence' as a universal concept, and try to measure individuals across cultures on that single concept or definition. Some also see this as one way to think of the boundary between the disciplines of anthropology (emic) versus psychology (etic). See Harris (1976).

been a matter of serious discussion and debate over the years (Stevenson and Stigler, 1982).[9]

The key issue here is the degree to which it is necessary to have full comparability in learning outcomes, with all individuals and all groups on the same measurement scale. Or, if a choice is made not to 'force' the compromises needed for a single unified scale, what are the gains and losses in terms of comparability? Can international goals (and statistics) be maintained as stable and reliable if localized approaches are chosen over international comparability?[10] The way this question has been answered has led to situations in which some low-income countries, while tempted to participate in international learning assessments, nevertheless hesitate due to the appearance of very low results, or the feeling that the expense of participation is not worth the value added to decision-making at the national level.[11]

In the end, global research on learning requires some form of comparability, but not necessarily in identical ways. For example, international and regional assessments are aimed specifically at cross-national comparability, while hybrid assessments are more focused on local contexts and increased validity. Hybrids offer some kinds of comparability that large-scale assessments do not, such as that relating to marginalized populations or younger children. Which types of comparability are most important depends on the policy goals desired, as well as timing and cost considerations. As in comparative education more

9. In a more recent example, closer to present purposes, senior officials in Botswana were interested in knowing how Singapore came to be first in mathematics (Gilmore, 2005).
10. Translation of international large-scale educational assessments (LSEAs) remains a problem, as it is often uncertain whether an equivalent translated item will have the same statistical properties as an indigenous word chosen independently. See Hambleton and Kanjee (1995) for a discussion on translation issues in international assessments.
11. See Greaney and Kellaghan (1996) for a useful review of this issue. Others may participate because they do not want to be viewed as having 'inferior' benchmarks to those used in OECD countries. It should be noted that donor agencies often play a role in this decision-making by supporting certain assessments as part of a 'package' of support for evaluation capacity building.

generally, cultural context will determine whether and when research findings are deemed credible.[12]

Evidence uptake

Policy-makers, ministers of education, community leaders in rural villages, teachers, parents, and educational specialists should be held to account for what and how children learn. Until now, educational specialists and statisticians in most countries (and especially in low-income countries) have been the primary 'guardians' of learning processes and their importance for school and economic success. This restricted access to knowledge about learning is due, at least in part, to the complexities of the science of learning. But it is also due to insufficient knowledge – and at times erroneous beliefs – among both parents and children about the importance (or lack of importance) of learning and schooling on life's chances.[13]

Today, it is more important than ever before to involve multiple stakeholders in education decision-making and in learning. Public interest in children's learning and school achievement has grown in many countries, due in part to globalization, but also to the influence of international agencies, the efforts of NGOs, greater community activism, and parental interest. Some of the recent Pratham and EGRA field studies have involved strong community engagement that has led to significant government take-up of empirical findings.[14]

12. See Steiner-Khamsi (2010) for a discussion on comparability in comparative education.
13. Much evidence from many societies suggests that poor communities underestimate the value of learning and schooling. See Stevenson and Stigler (1982) for a comparison of parental beliefs in the USA, China, and Japan.
14. See Bhattacharjea, Wadhwa, and Banerji (2011) on India, and Piper and Korda (2009) on Liberia. Though solid research is lacking to date, several African countries have devoted considerable attention to the UWEZO initiative, which has adapted a version of Pratham's community mobilization and accountability approach. See: www.uwezo.net/index.php?c=38 (downloaded 16 September 2012), and Pratham (2011), http://pratham.org/file/Pratham%20Annual%20Report.pdf (downloaded 2 November 2012).

This type of multilevel information exchange is another way of speaking about accountability and expectation. Whose problem is it if a child, teacher, school, district, or nation is not performing at a given level of learning? Indeed, how are such expectations even built? Whose expectations should be taken into account? Knowledge about the importance of learning – and how it can be achieved in formal and non-formal settings, and in structured and informal ways – has the potential of breaking new ground in policy development, community and family participation, and local ownership.

Choosing a research approach

Research can take many forms and have multiple approaches. This is not just a matter of methodological choice (e.g. quantitative vs qualitative) or disciplinary training (e.g. economics vs anthropology), though these two dimensions often get the most attention. Rather, in trying to address how research can improve learning, it is also important to understand three broad (and sometimes overlapping) approaches that continue to channel researchers' efforts, each of which has been utilized extensively in the study of education and development (see Wagner, 1986):

- *Knowledge-driven research.* This approach is most commonly seen in doctoral dissertations, in which the researcher usually follows in the footsteps of previous scientists in order to elaborate on a particular theory, hypothesis, or knowledge unit. Hence, knowledge-driven research is of the sort that is found in many scientific journals seeking to build up the knowledge base around particular topics. A good example from the present review is the role of phonics in reading, in which much of the research has been undertaken in OECD countries and in laboratories that explore the psychometrics of reading skill acquisition.
- *Decision-driven research.* Many implementation projects in development set aside some funds (or find external funding) for 'what works' research. Thus a project such as a pre-school intervention programme would seek to know, for example, whether the programme itself was implemented properly (classrooms available, teachers and children present, etc.), and

whether (say) learning outcomes tracked the instructional inputs provided (such as use of a national language in the classroom).

- *Context-driven research.* In holistic culture-specific work, researchers (especially ethnographers) focus on the special characteristics of particular contexts. The goal is to understand the unique relationships between factors that occur in a particular cultural context, rather than the sampling of common elements that might occur between contexts or ethnographic settings.

Conclusion

In sum, multidisciplinary and multi-method approaches to improving learning in low-income countries and marginalized communities are not *scientifically* more difficult than similar research done in wealthier communities. However, given where most scientific (human and fiscal) resources are located, it can be much less convenient for those with the advanced training needed to do the work. That fact, among others, is why so much remains to be known about learning in low-income countries. Multiple methodologies will need to be brought into play and debated. Limits (or boundary constraints) will be invoked to account for why one or another generalization can or cannot be made.

The challenge to policy development of working on learning and the quality of education in highly diverse cultural contexts is serious, especially for international agencies whose bias is towards international comparability. The main implication of this argument is that such comparability may be seen as a trade-off with validity in local contexts. The more that comparability is required, the less likely it is for results to be applicable to diverse settings.

References

Arnett, J. 2008. 'The neglected 95%: Why American psychology needs to become less American'. In: *American Psychologist, 63*(7), 602–614.

Banerjee, A.V.; Duflo, E. 2011. *Poor economics: A radical rethinking of the way to fight poverty.* New York: Public Affairs.

Bartlett, L.; Garcia, O. 2011. *Additive schooling in subtractive times: Bilingual education and Dominican immigrant youth in the Heights.* Nashville, Tenn.: Vanderbilt University Press.

Bhattacharjea, S., Wadhwa, W.; Banerji., R. 2011. *Inside primary schools: A study of teaching and learning in rural India.* New Delhi: ASER. http://images2.asercentre.org/homepage/tl_study_print_ready_version_oct_7_2011.pdf

Blackledge, A.; Creese, A. 2010. *Multilingualism: A critical perspective.* London: Continuum.

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

Bruns, B.; Filmer, D.; Patrinos, H.A. 2011. *Making schools work: New evidence on accountability reforms.* Washington, DC: World Bank.

Burde, D. 2012. 'Assessing impact and bridging methodological divides: randomized trials in countries affected by conflict'. In: *Comparative Education Review, 56*(3), 448–473.

Castillo, N.M.; Wagner, D.A. 2014. 'Gold standard? The use of randomized controlled trials for international educational policy'. In: *Comparative Education Review, 58*(1), 166–173.

Gilmore, A. 2005. *The impact of PIRLS (2001) and TIMSS (2003) in low- and middle-income countries: an evaluation of the value of World Bank support for international surveys of reading literacy (PIRLS) and mathematics and science (TIMSS).* New Zealand: IEA.

Greaney, V.; Kellaghan, T. 1996. *Monitoring the learning outcomes.* Washington, DC: World Bank.

Greaney, V.; Khandker, S.R.; Alam, M. 1999. *Bangladesh: Assessing basic learning skills.* Washington, DC: World Bank.

Hambleton, R.K.; Kanjee, A. 1995. 'Increasing the validity of cross-cultural assessments: use of improved methods for test adaptation'. In: *European Journal of Psychological Assessment, 11*(3), 147–157.

Harris, M. 1976. 'History and significance of the emic/etic distinction'. In: *Annual Review of Anthropology, 5*, 329–350.

Hoffman, J. V. 2012. 'Why EGRA – a clone of DIBELS – will fail to improve literacy in Africa'. In: *Research in the Teaching of English, 46*(4), 340–357.

Howie, S.; Hughes, C. 2000. 'South Africa'. In: D. Robitaille, A. Beaton, and T. Plomb (Eds), *The impact of TIMSS on the teaching and learning of mathematics and science.* Vancouver, Canada: Pacific Educational Press.

IIEP-UNESCO. 2010. *SACMEQ III Project Results: Pupil achievement levels in reading and mathematics* (Authors: N. Hungi, D. Makuwa, M. Saito, S. Dolata, F. van Cappelle, L. Paviot, and J. Vellien). Paris: IIEP-UNESCO.

Kremer, M.; Holla, A. 2009. 'Improving education in the developing world: What have we learned from randomized evaluations?' In: *Annual Review of Economics, 1*, 513–542.

Paris, S.G.; Paris, A.H. 2006. 'The influence of developmental skill trajectories on assessments of children's early reading'. In: W. Damon, R. Lerner, K.A. Renninger, and I.E. Siegel (Eds), *Handbook of child psychology: Vol. 4. Child psychology in practice* (6th edition). Hoboken, N.J.: John Wiley & Sons.

Piper, B.; Korda, M. 2009. 'EGRA Plus: Liberia – data analytic report'. Unpublished technical report. Washington: RTI and Liberian Education Trust.

Piper, B.; Mugenda, A. 2012. *The Primary Math and Reading (PRIMR) Initiative Baseline Report.* Washington, DC: RTI.

Pratham. 2012. *Pratham India Education Initiative: Annual Report 2011.* New Delhi: Pratham Resource Centre. http://pratham.org/images/Aser-2011-report.pdf

Saito, M. 2011. *Trends in the magnitude and direction of gender differences in learning outcomes.* Paris: IIEP-UNESCO.

Sjoberg, S. 2007. 'PISA and "real life challenges": Mission impossible?' In: S.T. Hopmann, G. Brinek, and M. Retzl (Eds), Vi*PISA according to PISA: Does PISA Keep What It Promises?* enna: LIT Verlag. http://folk.uio.no/sveinsj/Sjoberg-PISA-book-2007.pdf.

Steiner-Khamsi, G. 2010. 'The politics and economics of comparison'. In: *Comparative Education Review, 54*(3), 323–342.

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

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

—— 2011*a*. *Smaller, quicker, cheaper: Improving learning assessments in developing countries.* Paris and Washington, DC: IIEP-UNESCO and EFA Fast Track Initiative of Global Partnership for Education.

——. 2011*b*. 'What happened to literacy? Historical and conceptual perspectives on literacy in UNESCO'. In: *International Journal of Educational Development, 31*, 319–323.

Wagner, D.A.; Murphy, K.M.; de Korne, H. (2012). *Learning first: A research agenda for improving learning in low-income countries.* Center for Universal Education Working Paper. Washington, DC: Brookings Institution.