CICE Hiroshima University, *Journal of International Cooperation in Education*, Vol.20-2/21-2 (2018) pp.85 ～ 105

Evidence, SDG 4, Targets and Indicators: Summative assessments of systems vs. formative assessments of learners?

Angela W Little
University College London and Hiroshima University

Abstract

In 2015 the United Nations advanced 17 global goals for 'sustainable development', accompanied by 169 Targets and 304 Indicators. Goal 4 aims to 'Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all', accompanied by 10 targets, and 43 indicators. This paper explores the technical and political challenges that face the creation of internationally comparable assessment evidence for Indicator 4.1.1 – the proportion of children and young people: (a) in grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex. It also questions whether the search for internationally comparable evidence of system-performance is the most effective way of improving learning. Drawing on the international literature on types of assessment and distinctions between summative vs. formative assessment, as well as a consideration of the unit of analysis (whole systems vs. the individual learner), it suggests that the intensive practice of class-based formative assessment deserves more attention if learning levels are to be improved. (177)

Introduction

In 2015 the United Nations advanced 17 global goals for 'sustainable development'. They came to be known as the Sustainable Development Goals, and replaced the previous eight Millennium Development Goals. They were adopted by 193 countries of the United Nations General Assembly as the 2030 Development agenda 'Transforming our world: the 2030 agenda for Sustainable Development' (United Nations, 2015).

The concept of Sustainable Development has been advanced at least since 1987 by the World Commission on Environment and Development as development 'that meets the needs of the present without compromising the ability of future generations to meet their own needs' (WCED, 1987). The more complete definition contained two key elements: (i) the needs referred to were the essential needs of the world's poor, and (ii) the state of technology and social organisation imposed limits on the ability of the environment to meet both present and future needs (Pearce, 2007). In 1992 the world's governments adopted Agenda 21 at the Earth Summit held in Rio de Janeiro. *Inter alia*, the agenda introduced the idea of 'sustainable consumption' and the call to people in rich countries to change their consumption patterns if sustainable development could be achieved (UNCED, 1992). This would involve investment in education.

A second World Summit on Sustainable Development was convened in Johannesburg in 2002. This recognised that education would have to play a major role in the future realisation of a 'vision of sustainability that links economic well being with respect for cultural diversity, the Earth and its resources' (UNESCO, 2007, p.6). As a consequence of this, the United Nations General Assembly adopted Resolution 57/254 and declared the Decade for Education for Sustainable Development (DESD) 2005-2014. The overall goal of DESD was to

> Integrate values, activities and principles that are inherently linked to sustainable development into all forms of education and learning and help usher in a change in attitudes, behaviours and values to ensure a more sustainable future in social and environmental and economic terms' (UNESCO, 2007, p.5)

Sustainable Development lay in three spheres – environment (including water and waste), society (including employment, human rights, gender equity, peace and human security), and economy (including poverty reduction, corporate responsibility and accountability). This conceptualisation would come to influence the SDG discourse. So too did the United Nations Conference on Sustainable Development, also known as Rio 2012 or Rio +20 (UNCSD, 2012), which aimed to reconcile the economic and environmental goals of society. The 2030 agenda, declared in 2015, reinforces the challenges in three areas of development – environmental, economic and social. Most recently, the 24th conference of the parties to the UN Framework Committee on Climate Change, held in Katowice, Poland, committed to limit global temperature rises to well below two degrees Centigrade.

In parallel with the discourse about tensions between environmental sustainability and economic growth and the need for education to address these issues, which may be

termed Education for Sustainable Development (ESD) and essentially a curriculum or education content issue, was a rather separate discourse about the stages of education in which investment was required and universal participation to be encouraged (Little & Green, 2009; Lewin, 2015a). At an international level these may be traced to the regional UNESCO conferences on Universal Primary Education, held in Karachi in 1960, Addis Ababa in 1961, Santiago in 1962, and Tripoli in 1966. The first education conference held on a world scale was the World Conference on Education for All: meeting basic learning needs, held in Jomtien, Thailand in 1990. Here six Education for All (EFA) goals declared. A decade later, the World Education Forum produced the Dakar Framework for Education in Dakar in 2000. This reworked and re-ordered the Jomtien goals but were very similar (Table 1).

Table 1. Goals and targets for EFA affirmed in Jomtien (1990) and Dakar (2000)

Jomtien Framework	Dakar Framework
1. Expansion of early childhood care and developmental activities, including family and community interventions, especially for poor, disadvantaged and disabled children	1. Expanding and improving comprehensive early childhood care and education, especially for the most vulnerable and disadvantaged children
2. Universal access to and completion of primary education (or whatever higher level of education is considered as "basic") by the year 2000	2. Ensuring that by the year 2015 all children, particularly by girls, children in difficult circumstances and those belonging to ethnic minorities, have access to and complete free and compulsory primary education of good quality
3. Improvement of learning achievement so that an agreed percentage of an appropriate age cohort (e.g., 80% of 14-year-olds) attains or surpasses a defined level of necessary learning achievement	3. Improving all aspects of the quality of education and ensuring excellence of all so that recognized and measurable learning outcomes are achieved by all, especially in literacy, numeracy and essential life skills
4. Reduction of the adult illiteracy rate to, say, one half of its 1990 level by the year 2000, with sufficient emphasis on female literacy to reduce significantly the current disparity between male and female illiteracy rates	4. Achieving a 50% improvement in levels of adult literacy by 2015, especially for women, and equitable access to basic and continuing education for all adults
5. Expansion of the provision of basic education and training in other essential skills required by youth and adults, with program effectiveness assessed in terms of behavioral change and impact on health, employment and productivity	5. Ensuring that the learning needs of all young people and adults are met through equitable access to appropriate learning and life skills programs
6. Increased acquisition by individuals and families of the knowledge, skills and values required for better living and sound and sustainable development, made available through all education channels	6. Eliminating gender disparities in primary and secondary education by 2005, and achieving gender equality in education by 2015, with a focus on ensuring girls' full and equal access to and achievement in basic education of good quality

Source: Jomtien Framework for Action to Meet Basic Learning Needs (1990); Dakar Framework for Action for EFA: Meeting our Collective Commitments (2000)

Contrary to much of the current discourse on EFA and the SDGs, the declarations at Jomtien and Dakar both emphasised the need for improvements in learning outcomes. While the Jomtien and Dakar conferences and declarations focused on goals for education, the United Nations Millennium Summit, held in New York in 2000, positioned goals for education alongside seven Millennium Development Goals (MDGs) from across all development sectors. This had the effect of restricting the goals for education and the associated targets and indicators.

The 17 Sustainable Development Goals declared in 2015 are accompanied by 169 targets and 304 indicators. The main education goal, listed as Goal 4, aims to 'Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all'. SDG has 10 targets, seven of which may be viewed as primary targets and three are 'enabling' targets (Table 1). Alongside these 10 targets are listed no fewer than 43 indicators (IAEG-SDGs, 2016), though as we shall see later, some of the indicators involve a cluster of sub-indicators, if they are intended to be measurable. A comparison of the MDGs with the SDGs suggests a proliferation of targets and indicators of education, a proliferation of information and social media surrounding their determination and monitoring of progress towards them, and a proliferation of agencies with interests in them, especially national and international non-governmental organisations and international private sector interests (Lewin, 2015b). The interdependence and synergistic potential of the SDGs are recognised rather more than had been the case in the discourse surrounding the MDGs (Waage et al., 2008). Education can be expected to contribute to Sustainable Development both directly and indirectly - directly, through its achievement of improvements in learning outcomes, and indirectly, through its contribution to the other SDGs – such as to the reduction of poverty and hunger, the positive contribution to health and well being, gender equality and climate action.

We turn now to a more detailed consideration of SDG 4, Target 4.1 and Indicator 4.1.1. SDG 4's first (of ten) targets is to 'ensure that all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning'. The evidence required to assess progress towards the achievement of this target is the

> Proportion of children and young people: (a) in grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex (Indicator 4.1.1).

The Call for Evidence on Learning Outcomes

Most systems of education around the world already measure the academic achievements of its school children; some of these systems stretch back decades, if not centuries. Traditionally, assessments of learning achievement have been used to certify the completion of a stage of education and to select students for further education and/or employment. Some systems conduct systematic comparisons of assessment results over

time and this evidence sometimes feeds into national policy debates. Schools compare the performance of their students' performance in national examinations annually and proudly boast of their achievements.

But some elements of the international education community seek internationally comparable data. Indeed, since the establishment of the International Association for the Evaluation of Educational Achievement in the early 1960s, successive international and regional student assessments have been conducted. With each round, more countries have been invited to participate in these assessments (e.g. IEA, TIMMS, PISA). Lockheed (2015) reports that participation in PISA grew from one fifth of all countries in 2000 to one third of countries in 2015. Participation in PISA was higher for OECD member countries, countries in the Europe and Central Asia region, high- and upper-middle-income countries, and countries with previous national and international assessment experience. However, while participation may be thought to indicate willingness and ability of countries to participate, we do not know why countries chose to participate, who has used the assessment evidence generated and what impact the assessment has had on teachers and learners.

There has been a massive expansion of international comparisons and national learning assessments since 2000. During the post-Jomtien decade 70 countries conducted at least one national assessment while this number had increased to 142 during the post-Dakar, 2000-2013 period (UNESCO, 2014). Many of those in the developing world were funded by loans and grants from the World Bank, continuing their push since the 1990s for the creation of national assessment systems. The growth has not been confined to developing countries. Among developed countries only five national assessments appear to have been conducted in 1990, compared with 36 in 2013. Among developing countries eight national assessments were conducted in 1990, compared with 64 in 2013 (UNESCO, 2014). Note here that national assessments refer to surveys of learning outcomes designed to monitor the performance of a system as a whole, rather than national systems of public examination used for certification and selection purposes. This has been driven by a growing interest on the part of powerful blocs of the international community in finding ways of measuring comparable educational quality. For example, the EFA-Fast Track Initiative (EFA-FTI) (later to become the Global Education Partnership (GPE)) included the monitoring of learning outcomes as a criterion in the endorsement of funding to countries from the EFA-FTI and the need for objectively verifiable indicators.

International and bilateral donors have shown a greater degree of interest in the measurement of learning outcomes since the World Education Conference in Dakar in 2000. The United States of America State Department and USAID now require countries to demonstrate increases in the proportions of children who attain minimum grade level proficiency in reading at the end of Grade 2 primary and on completion of the full primary stage. As Bruns (2018) has commented 'with an $800 million international basic education budget on the line, there are high stakes around how 'minimum grade-level proficiency' is defined and measured'. While some donors focus their attention on basic

literacy and numeracy skills in primary education, others focus on the skills needed for economic growth. Evidence provided by The International Association for the Evaluation of Educational Achievement's surveys of trends in Mathematics and Science (TIMMS) and the OECD's Programme for International Student Assessment (PISA) is used by economists, politicians and aid donor groups to assert causal links between test scores and the economic growth of a country (e.g. Hanushek & Woessmann, 2008). Others are concerned with the global competition for mobile skilled labour and emerging knowledge societies (e.g. OECD, 2012.

Concerns with system-level achievement and assessment were also fueled by 'evidence' presented in reports on 'high performing systems' produced by McKinsey through the 2000s. In *How the world's most improved school systems keep getting better*, Barber, Chijoke, & Mourshed (2010) analyse successful education reforms in 20 education systems. These are described as 'sustained improvers' and 'promising starts'. In some countries, several 'episodes' of reform are studied, yielding 34 'reform journey' cases are studied. Not all of these system-wide reforms. Some are undertaken in a state within a country, while others are city-based or even small school networks within cities. Systems are further divided into four categories, those that have improved from 'poor to fair', 'fair to good', 'good to great' and 'great to excellent', these categories being based on levels of achievement. Eight main lessons are drawn: (i) A system can make significant gains from wherever it starts; these gains can be achieved in six years or less. (ii) There is too little focus on "process" in the debate today. Improving system performance ultimately comes down to improving the learning experience of students in their classrooms. (iii) Each particular stage of the school system improvement journey is associated with a unique set of interventions. (iv) A system's context might not determine what needs to be done, but it does determine how it is done. (v) Interventions occurring equally at every performance stage for all systems include building the instructional skills for teachers and management skills of principals, assessing students, improving data systems. (vi) Systems further along the journey sustain improvement by balancing school autonomy with consistent teaching practice. (vii) Leaders take advantage of changed circumstances to ignite reforms, and (viii) Leadership continuity is essential.

While the evidence in the McKinsey report is presented in an extremely upbeat and persuasive style, its underlying design and presentation of evidence is flawed in at least one major respect. Because it focuses only on improving systems, it omits from its analysis of evidence systems in which student performance was stable or in decline. Many of the interventions mentioned by the leaders of the improvers are not uniquely associated with improving systems. Only through a comparison with systems deemed to have stayed in one place or gone into decline could the authors assert with any confidence that they have identified the most important reform drivers. The report acknowledges this when it says: "the systems that have been unsuccessful in trying to improve may carry out the same types of intervention that successful system undertake", but it goes on to assert "but there appears to be one crucial difference, that they are not consistent, either in

carrying out the critical mass of interventions appropriate to their performance stage, or in pursuing them with sufficient rigour and discipline" (Barber, Chijoke, & Mourshed, 2010, p20). Since the report offers no evidence from 'unsuccessful systems' it is difficult, if not impossible, to judge the veracity of the evidence and the validity of the inferences drawn.

Back to the Indicator

Nonetheless, the call for more and more evidence on student achievement and learning outcomes has gathered momentum, so let us focus on the suggested indicator and explore the challenges for its measurement. Indicator 4.1.1 reads

> Proportion of children and young people: (a) in grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex.

Even a cursory glance of this statement suggests immediately that this is not a single indicator If the proportions are to be disaggregated by gender, this Indicator involves the computation of evidence on 12 sub-indicators. A further question arises: are these proportions of children who are attending/enrolled in school, or all children, whether in school or not? If the proportions are intended to indicate the 'stock' of human capital available within a country, and if only children in school are to be assessed, then any performance 'scores' would need to be adjusted for the proportion of children enrolled in school. In some countries these proportion vary widely between Grades 2/3, primary completion and lower secondary completion. And there is wide variation between countries at each of these levels, raising questions about the interpretation of cross-national evidence.

These questions are arguably simpler to address than the much greater challenge of how this evidence is generated, who collects the evidence, by whom will the evidence be used, and for whose purposes? The current ambition of the United Nations SDG community is the involvement of all countries in internationally comparable assessments of learning outcomes.

A large number of international and regional assessment surveys have been conducted in recent years and these provide much experience from which the international community can draw. Trevino and Ordenes (2017) compare fifteen international and regional assessments in current use (Table 2).

Table 2. Student assessments reviewed

International assessments (non-regional)	Regional assessments	Assessments of foundational skills
ePIRLS: Progress in International Literacy Study (online reading) **LANA**: Literacy and Numeracy Assessment **PIRLS**: Progress in International Reading Literacy Study **PISA**: Programme for International Student Assessment **PISA D**: Programme for International Student Assessment for Development **TIMSS**: Trends in International Mathematics and Science Study	**LLECE**: Latin American Laboratory for Assessment of the Quality of Education **PASEC**: Programme for the Analysis of Education Systems **PILNA**: Pacific Islands Literacy and Numeracy Assessments **SEA-PLM**: Southeast Asia Primary Learning Metrics **SACMEQ**: Southern and Eastern Africa Consortium for Monitoring Educational Quality	**ASER**: Annual Status of Education Report **EGMA**: Early Grade Mathematics Assessment **EGRA**: Early Grade Reading Assessment **UWEZO**: Uwezo Annual Learning Assessment

Source: Trevino and Ordenes (2017)

The authors set our three main criteria by which these assessments need to be compared, and, in so doing, raise awareness of the technical challenges posed by the global ambition for a single test on which all countries can be compared. The first criterion is the design of an assessment which involves a range of technical decisions about the overarching purpose of the assessment, the intrinsic rationale of the instrument and the conceptual framework to be assessed. The design defines purpose 'as well as *what* to measure and *how* to measure it' (Trevino & Ordenes, 2017, p.6). *Inter alia* the design dimensions include purpose (e.g. system–monitoring, programme evaluation, base line definition, student population diagnosis); target populations (age groups, grade groups), what is being assessed (curriculum knowledge, competency-based), domains (specific knowledge and skills), inferences (the validity of inferences made from the assessments), sample (the subgroup within the targeted population included), and modes of assessment (e.g. paper vs. computer-based), site of test administration (school, household, test centre), and individual vs. group administration.

The second criterion is standard setting. This is important for Indicator 4.1.1 because it aims to identify the proportion of students who perform at or above a minimal level. This involves identifying cut-scores on a scale that define the threshold between different levels and writing of substantive descriptions of what students classified into different level can do.

The third criterion is statistical criteria, of which three are very important. The first is the scaling technique chosen to create the measure of achievement (Classic test theory, Rasch modelling and multi-parameter models). The second is the way individual achievement results are estimated, and the third is 'equating', the procedure used to make

assessments comparable.

The fifteen assessments listed in Table 2 vary in their purposes. While the majority aim to monitor the performance of a school system over time and/or in relation to other systems, some – EGMA, EGRA, ASER and UWEZO – conduct system diagnosis or programme evaluation at the national or sub-national level. These are not designed to produce cross-national comparisons. Some of the assessments are age-based; others are grade-based. PISA is aged-based and is designed to generate evidence across systems and/or over time. By contrast, ASER and UWEZO are age-range (5-16 years) assessments and are not designed to compare across systems and/or over time. In systems where many students enter school late and/or repeat grades age-based approaches will include some students who will not have had an opportunity to learn the same grade-related material as others.

While the distinction between content-based and competency-based assessments is not clear-cut the authors describe a content-based assessment as measuring 'the extent to which students know the contents or standards of a particular subject matter' and a competency-based assessment as measuring 'the extent to which children can apply competently the knowledge and skills they have learned in the education trajectories' (Trevino & Ordenes, 2017, p.9). This distinction is still obscure, especially if particular subject matter involves the application of knowledge, skills and principles to unfamiliar situations. Nonetheless the authors claim that the distinction is important for reasons of external validity and the feasibility of comparison. Five of the surveys are classified as content-based (including EGMA, EGRA, PISA) and ten as competency-based (including SACMEQ and TIMSS). In order to define commonality among different assessments, it is important to know whether they are assessing the same thing. For example, in Grade 2/3 tests of literacy (EGRA, LLECE and PASEC), the domains and sub-domains range from phonological awareness and reading fluency to text comprehension. In Grade 2/3 tests of numeracy, they vary from number identification, quantity discrimination, number patterns, addition and subtraction and word problems (EGMA) to proficiency in numbers, geometry, measurement, statistics and variation (LLECE).

Significant variation is also found on all other design, standard setting and statistical criteria. These variations in technical criteria are suggestive of the magnitude of the challenge of combining the results of these various surveys in order to attain global coverage.

The Political Economy of Assessment

As well as explaining the technical challenges, Trevino & Ordenes (2017, p.21) outline three political challenges to the development of internationally acceptable measures of performance. The first relates to how much of respective national curricula are represented in the definition of minimal level and in terms of items included in the test. A second challenge concerns the internal political and social pressures likely to be

exerted within low achieving countries, even if that countries system has been improving over time. A third challenge, arising from these two, would arise within the international community if countries, especially low-achieving countries, were to question the validity of the test, pointing to the low representation of respective national curricula in the composition of test items and content. The authors go on to identify four strategies for measuring Indicator 4.1.1 and their attendant implications for international comparability, costs, technical robustness, external validity, involvement of international agencies in the design, implementation and analysis of test results. The four are (i) short-run use of national assessments with adjustments using international assessments, (ii) medium-run equating among international and regional assessments, (iii) medium to longer run equating between different international evaluations aiming at similar school grades, and (iv) long run creation of a worldwide proficiency assessment of numeracy and literacy. On balance the authors favour the last, a specific instrument with a clear and agreed on minimum level of competency. This they argue would be psychometrically robust for purposes of comparison. But they acknowledge the low level of external validity in terms of the representation of respective national curricula and the difficulty of convincing countries to participate in the assessment. Finally, they note the need for the support and collaboration of a number of agencies specialised in international evaluation (e.g. OECD/PISA, ETS, ACER, UIS) in the form of a consortium to ensure technical quality and to add political legitimacy.

This last point is important for it draws attention to a broader political dimension, what we may term the political economy of educational assessment. A political economy approach to educational assessment involves the analysis of (i) the underlying drivers for change and (ii) the identification of a wide range of actors with interests in the change and (iii) the incentives for change for particular groups of actors. Changed practices can motivate and de-motivate changes in behaviour. While some groups of actors respond to incentives in ways that promote policy reform, others may also perceive reforms as threats to their interests and lead to resistance (Little, 2008).

A major underlying driver for change has been a push from the international community, linked in many cases with the promise of funding. A number of powerful groups have been involved in this. They include the Learning Metrics Task Force, a working group with representatives/members international organisations (e.g. UNESCO, UNICEF, World Bank, USAID) and the Education Commission (The International Commission on Financing Education Opportunity, 2016).

Early in 2018, the World Bank published its flagship report ‘Learning to realise education’s promise’ (World Bank, 2018). This was the first time that Bank group had devoted an entire World Development Report to learning and education. In his foreword, and drawing on the experience of his own country, Korea, the Bank’s President Jim Yong Kim wrote

> Delivered well, education – and the human capital it creates – has many benefits for economies, and for societies as a whole. For individuals, education promotes

> employment, earnings and health. It raises pride and opens new horizons. For societies, it drives long-term economic growth, reduces poverty, spurs innovation, strengthens institutions and fosters social cohesion (World Bank, 2018, p.v)

The overall themes of the Report are that the assessment of learning should be a serious goal and action needs to be based on evidence is required to make schools 'work for learning'. This indicates the importance with which the Bank accords to investments in learning and education and to assessment in particular, which in turn is likely to guide future Bank spending. Other major players are the international assessment agencies themselves, some of which are funded by the Bank and other aid agencies.

The OECD's PISA is a major driver of assessment surveys and has an interest in the conduct of more in the future. It encourages countries to participate, with their own financial resources, some of which are then sought for from the Bank. UNESCO's Institute of Statistics has had a major stake in creating Indicator 4.1.1 and holds a major stake in the creation of measures. All of the organisations and associations involved in the testing programmes listed in Table 1 have stakes in continuing their activities. And these stakes, some of which are financial and some political, may mean that the cooperation and sharing of technical expertise required for global tests of proficiency will be thwarted. As Lewin has pointed out the 'learning crisis' is really a financial crisis arising from persistent underfunding of education (Lewin, 2018).

A broader view of stake-holding in assessment systems indicates a very wide range of possible stakeholders, both inside and outside national boundaries, with a range of possible incentives and responses. The World Bank's report outlines a useful list of the multiple interests that govern the actions of education stakeholders in countries (World Bank, 2018, Table O.2, 2018) and distinguishes learning-aligned interests from competing interests. So, for example, teachers have interests in student learning and practice in line with a professional ethic, but they also have the competing interests of their employment, job security, salary and private tuition. Other stakeholders within countries include school principals, bureaucrats, politicians, parents and students, the judiciary, employers, nongovernment schools (e.g. religious schools, private for-profit schools) and suppliers of educational inputs, all of whom have some learning-aligned interests and all of whom have competing interests.

In terms of external actors, the Bank's report mentions only international donors – but there will be many more. There are many whose business is the creation of more and more assessment tests, analysis and reporting. They will be joined by those who represent teachers through international trade unions, international civil society organisations, international research consortia, international curriculum providers and textbook publishers, international computer software firms, international and national firms that seek to attract highly skilled labour from specific countries, creating skill deficits within those same countries etc. All of these groups will, variously, perceive opportunities for budgets, contracts, expansion of jobs and increased power. At the same time some may perceive threats – national governments, if their system is exposed internationally as a

low performer, teachers, if their class and their school is exposed as a low performer, national examination boards, if there is a move over time to substitute public selection examinations with scores from the performance of individuals on system-wide tests, trade unions if teachers are expected to increase their workloads and/or change their practices to increase national levels of achievement. Political responses are likely to range from an embrace of proposed assessment reforms to active resistance, sabotage and avoidance of implementation.

Assessment of systems or of student learning?

We turn now to the question of learning and the role of assessment in promoting it. The main concern here is the teacher and the student and how the teacher can support the student in their learning. We start with the World Bank's recent report on learning and its urgings to make assessment a serious goal, act on evidence to make schools work for learning and align actors to make entire systems work. The report opens with the following:

> Schooling is not the same as learning. In Kenya, Tanzania and Uganda, when grade 3 students were asked recently to read a sentence such as 'the name of the dog is Puppy', three quarters did not *understand* what it said. In rural India, just three quarters of students in grade 3 could not solve a two-digit subtraction such as 46-17. (World Bank, 2018, p.3, author's emphasis)

The report concludes by outlining a number of ways in which external actors can reinforce strategies for learning through, for funding assessments, spotlight challenges and catalyse domestic efforts for reform and promote results-based financing. Its main focus is on national and international assessment systems.

Unfortunately, for those who promote the power of evidence from large-scale assessment surveys to improve the quality of learning and the development of human capital, the two simple examples above expose the assessment challenge. While they are intended to capture readers' attention they exemplify the technical issue raised earlier about the validity of the item content. But here I address the issue from the perspective of teachers and students. Teachers and students want to know whether a test item makes sense, whether it bears any relation to what they have taught or learned to date, whether the test is presented to them in a way they recognise and/or understand, and whether or not they succeed or fail on the item.

The first assessment item above is curious to say the least. First, what sense does it convey to the student? Many students will know that a puppy is the name of a baby dog. A puppy is a subset of all dogs. Few dogs are named 'Puppy' by their owners (certainly not in North America https://dogtime.com/top-100-dog-names), even though a puppy named Bailey, Bella or Rover may occasionally be referred to as 'the puppy' in everyday conversation. Second, in which language and orthography was the item administered? If Swahili, then orthography is not an issue, but was the word 'Puppy' retained within the

sentence, or was it replaced by the Swahili word for a baby dog, *motto wa mbwa*? Third, was the item written first in English and translated to Swahili, or vice versa? Fourth, was the item 'backtranslated' and modified if necessary? I (author) can read out this sentence to an examiner, but I do not understand it or where it is coming from. Was there a comprehension passage that preceded the posing of a question about the name of a dog? These are questions that strike an educator between the eyes.

The second item, the subtraction item, appears to be a little more straightforward. But even here, an assessment expert would wish to ask how this subtraction item was presented to students? Was it preceded by the word 'Subtract?' If so, in whose language and whose orthography was it written? Was it presented in a horizontal format, rather than a, more familiar for some, vertical format? If presented in a vertical format, was the subtraction sign on the top of bottom line, on the right or the left? All students worldwide benefit from familiarity with different types of test item, including the way they are presented visually, on paper, a blackboard, a whiteboard or a screen. If the test is to be 'fair' in an 'international test', its visual representation would need to be familiar to all students being assessed,

These tiny examples reinforce the technical challenges outlined by Trevino & Ordenes (2017) above, but they also highlight the way in which technical challenges can become lost in an international discourse, produced largely by economists, that is promoting a massive increase in the business of assessment. The international discourse struggles with this level of enquiry because it employs a world view and top-down analysis which, while it has some value in itself, is of limited value if it cannot be analysed and interpreted in a diversity of languages and countries. The questions raised may seem trivial from the point of view of economists of education but they are critical if students' learning is to be assessed fairly and in relation to their opportunities to learn.

In section 1 above the various purposes and types of assessment were outlined. SDG 4 Target 4.1 and Indicator 4.1.1 focus on the assessment of nationally representative samples in order to assess whether this target is being reached. The main purpose of the evidence generated by the assessment is the monitoring of system performance. Although SDG4 is intended to improve the quality of education as measured by this target and indicator, there is no clear indication of what system designers should change to bring about improvements in the percentages of children achieving the desired result. The results chain contained within the SDGs consists of a goal, a target and an indicator. A little like the Outcomes-Based Curriculum (OBC) reforms in the 1990s there are expected results/outcomes but no prescription of how to reach the outcome. The "action plan" links in the chain are missing. Perhaps this was intentional, as it was with OBC. Set the goals and targets for the curriculum and teachers are granted the autonomy to work out for themselves how to reach them. As we saw with the OBC reforms in Australia and South Africa, highly experienced, qualified and motivated teachers enjoyed the freedom offered by OBC. Less experienced and less qualified teachers struggled.

Summative vs. Formative Assessment

We turn now to a very different type of assessment which, if employed widely, could, in principle, be used to improve student learning in the classroom – and contribute to the assessment of performance across a system. Central to this discussion is the distinction between summative and formative assessment of learning.

We are all familiar with summative assessment. The purpose of national examinations, end-of-year tests and end–of-term tests is to provide evidence of the level of student learning at the end of a course of study. Evidence is judged in relation to a set of criteria (criterion-based assessment) or the performance of others (norm-based assessment). When this evidence is used to select persons for future education and occupations or life chances in general and when these life chances are limited and highly sought after, the stakes are high. These may be referred to as ‘’high-stakes summative assessments’.

By contrast, the goal of formative assessment is the monitoring of student learning in order to provide *timely feedback* for teachers. The critical words here are *timely* and *feedback*. Feedback evidence can, in principle, be used by teachers to improve their teaching and by students to improve their learning. Formative assessments help teachers and students to identify strengths and weaknesses and content/skill areas that require extra and/or remedial work. They help both teachers and students to address and remediate problems *immediately* and help teachers identify students in need of extra support. Formative assessments generally occur during classroom teaching and include, *inter alia*, student answers to focussed teacher questions, quizzes, teacher diagnosis of learning errors and remediation on the same day or soon after and submission and marking of draft work. In contrast to summative assessments, formative assessments rarely carry high stakes. Formative assessments are designed to assess learning progress and provide evidence that enables teachers and students to diagnose and resolve learning difficulties in the immediate or short term. In principle, formative assessments can be carried out as part of a lesson on a daily basis. Students are encouraged to engage in self-assessment as part of this process. The evidence generated by a formative assessment goes no further than the classroom. Formative assessments are not intended to be used to rank order students, schools, provinces or nations.

The key difference between formative and summative assessment is the use made of the information generated by the assessment. Formative assessment is the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there. Whereas summative assessments are typically used to make one-off judgements of learning levels achieved and are often used to select and stream students into future education provision. Formative assessment provides teacher and learner with information about performance on learning tasks that is available immediately and can be used to modify the next steps of teaching and learning. Inter alia, formative assessment is part of effective planning for

teaching and learning, is central to classroom practice, focuses on the task and the steps needed to perform the task well. Formative assessment refers to activities that teachers and students undertake to gain evidence that can be used diagnostically to alter teaching and learning. These may include teacher observation, classroom discussion, quizzes, analysis of student work, including homework and tests. It can also be built into the lesson itself as an integral learning task (e.g. students demonstrate their understanding of a subtraction task in mathematics and other students are invited to comment; teacher deliberately offers a wrong answer to a problem and asks the students to point out where he/she has erred). Assessment becomes formative when the information generated is used to adapt teaching and learning to meet student needs (Boston, 2002; Black & Wiliam, 1998; Assessment Reform Group, 1999). Summative assessments are to prove learning while formative assessments are to improve learning; or summative assessments are assessments of learning while formative assessments are assessments for learning.

Formative assessment of students relies on the notion that the identification of errors in learning and their diagnosis provide valuable information for both student and teacher. There is a long tradition of education research that analyses the errors that children make in mathematics and the strategies that may be adopted to turn a learning failure into a learning success. We begin with two examples.

> *Example 1* The teacher shows Siri a picture in his workbook of 10 children and 20 sweets. She asks him to use his pencil to draw lines and give each of the children the same number of sweets. Siri gives each child one sweet. The teacher asks, 'What about the others'? Siri replies, 'They are for me'!

From the teacher's point of view Siri has made an error in tackling this task. In an assessment 'test,' Siri's answer would be marked 'wrong,' but from Siri's viewpoint he had done what he was asked to do – he gave each child the same number of sweets, i.e. one sweet each. But the teacher expected him to give out all the sweets to all the children, not including himself. Siri had not yet learned a basic 'rule' of this type of workbook task, that is, that the answer is contained within the information provided on the workbook page. He is expected to distance himself from the task in hand and is not expected to seek any reward for himself.

> *Example 2* The teacher gives Kumari the following information and question: Pradeep had only 50 cents left with him. His mother gave him another 10 rupees. How much money does he have altogether? Kumari's answer was 60 cents.

Through further questioning by the teacher Kumari was able to read out the question correctly and she knew the meaning of cents and rupees. She knew that she had to add the two amounts together so she added 50 + 10, disregarding the different units of money, rupees and cents.

The errors that children make in mathematics tell us as much about the processes of learning as do the correct answers. Research on children's errors in arithmetic stretches back to at least the 1920s in the United States of America, Germany and Russia. That research remained largely confined within national boundaries and was exchanged rarely,

due in part to differences in learning theory traditions, education politics, the structure of education and curricula (Radatz, 1979, p.163).

Since that time, research findings have been more widely shared across national boundaries (within the limitations posed by language and the availability of language translations) and include errors in place value, ordering, problem solving, decimals and percentages, ratio and proportion, shape and area. Emphasis has been placed on the diagnosis of the underlying causes of error and the specific actions that teachers can take to help children overcome error. For just a few examples, see the work of Hansen et al. (2014), Ryan & Williams (2007), and Carpmail et al. (n.d.) in England; Newman (1977, 1983) in New Zealand; and White (2005, 2009) in Australia; Nanayakkara (1992) in Sri Lanka; White & Clements (2005) in Brunei; Radatz (1979) in Germany; and Resnick et al. (1989) in the United States of America. Researchers classify the sources of errors in overlapping ways. Radatz (1979) distinguished errors due to language difficulties, obtaining spatial information, deficient mastery of prerequisite skills, facts and concepts, incorrect associations or rigidity of thinking and the application of irrelevant rules of strategies. Newman (1977, 1983) distinguished nine types of errors (Table 3).

Table 3. Nine sources of errors in assessments

Type of errors	Definition
Reading errors	The child cannot yet read key words and symbols, or reads incorrectly
Comprehension errors	The child does not understand the overall meaning or specific terms
Transformation errors	The child is unable to identify the operation or sequence of operations needed to solve the problem
Process errors	The child is unable to perform the mathematical operations correctly (numerical, spatial, logical)
Encoding errors	The child is unable to write the answer in an acceptable form (words, symbols)
Careless errors	These are usually committed by working too fast and not checking work
Motivational errors	The student is tired, bored, hungry and/or cannot see the point of the assessment
Question wording errors	The error lies not with the student but with the author of the item. The item is poorly worded and designed. The writing of good items is a sophisticated skill in its own right.
Miscellaneous errors	The student guesses the answer, the student copies (incorrectly) the answer from a fellow student; or the student does not attempt the item for lack of time to be completed.

Source: Newman (1977, 1983)

Teachers' lack of expertise in the subjects means that they also make errors in mathematics. A large-scale study of primary teacher first-year undergraduates in Australia and a smaller study of primary teacher second-year undergraduates in England identified

a number of errors made by trainee teachers (cited in Ryan & Williams, 2007). Teacher trainees were given a test based on a primary teacher curriculum. Items assessed number, measurement space and shape, chance and data, algebra and reasoning and proof. Many teacher trainees made errors in place value and the conversion of a fraction to its decimal notation. For example, 24% of the sample could not write 912 + 4/100 in decimal form. Out of three options, 12% of the teachers selected 912.004 and 6% 912.25. Errors were also apparent in the division of decimal numbers by 100, as for example, 300.62 ÷ 100. Two options given were 3.62 and 3.0062. 22% selected 3.62. Fractions, computation, chance and measurement also generated errors. It would surprise nobody if these errors were passed on to children. More recent examples of teacher knowledge from developing countries include Bold et al. (2017) and Cueto et al. (2017).

Training teachers and support staff in formative assessment for learning

In this final section, I describe an extensive in-service training programme in Sri Lanka designed to help teachers and in-service advisers identify, diagnose and remediate errors in mathematics. Between 2014 and 2016, a series of residential workshops for education officers, in-service advisers (ISAs) and teachers on the use of formative assessment to improve the quality of primary education were held in each of the nine provinces of the country. The initial idea for a series of workshops on ways to improve primary education quality came from one of the Provincial Directors, dissatisfied with the performance of his province in a national assessment on the one hand, and frustrated that the same assessment did not permit him to compare performance across schools, divisions and zones on the other. More was a concern that a school needed not only to know where it stood compared with others but how its teachers could improve the performance of students. The initial workshops were conducted over 2-3 days for different staff cadres separately. Through experience and evaluation, the workshop programme was developed as the training team moved from one province to the next. In its final form a single workshop of 5 days was mounted, with two components: the first (five days) for ADEs primary, ISAs primary and primary teachers, and the second (one day) for province/zonal directors and theme convenors of the provincial plan. The one-day programme for the senior staff, conducted towards the end of the parallel five-day programme, included inputs from the ADEs, ISAs and teachers, based on their five-day workshop experience. A significant feature of the five-day workshops for ADEs, ISAs and teachers was the two half-days spent working with individual students in schools. In some provinces, separate Sinhala and Tamil workshops were organised; in others, Sinhala and Tamil participants attended the same workshop, with separated language streams for specific sessions. Between 2014 and 2017, around 700 in-service advisors, education officers and teachers had participated in the workshops.

The content of the workshops included brief reviews of international research on ‘effective teachers’, and exploration of the purposes of formative and summative

assessment, the identification of common errors through an item analysis of test performance (both one's own (i.e. the workshop participant) performance on a Grade 5 test and that of Grade 5 students in 2-4 schools in each province prior to the start of the workshop), a review of common types of error, the development of interview question to diagnose errors, interviews with children to identify and diagnose a range of errors, review of relevant curriculum units/learning tasks and the development of remedial activities, the trial of remedial activities, the writing up and sharing of error, diagnosis, remediation note, and the development of future work plans. All teaching materials were prepared in Sinhala, Tamil and English and copies of all training materials shared digitally with provincial staff for their use and adaptation in the future by staff in the provinces.

Since the programme was designed to enhance the professional development of teachers and education officers, and was adapted and developed over time as a result of formative feedback, its effects have not been evaluated in a systematic fashion to date. Nonetheless, the programme is indicative of what can be done in Sri Lanka and elsewhere in the future, providing teachers are granted a degree of autonomy and principals and advisers/inspectors tolerate and support such an approach by teachers As well as providing a wealth of material for the continuous professional development of teachers in the future, the material has enormous potential for use in curriculum manuals for teachers and as a training module for teachers and teacher educators.

Conclusion

The development of internationally comparable tests of performance of students in primary and secondary education poses formidable technical challenges and considerable political challenges. But if national curricula are to continue to be valued and not be determined by the backwash effects of international assessments, why does the evidence garnered for SDG 4 Indicator 4.1.1 need to be *internationally comparable*? There is no doubt that countries should be striving for use evidence generated through national systems to improve curriculum materials, teacher education and various assessment practices within countries over time. Just as learners need to know whether they are learning and improving, so national policymakers need to know whether their national system is improving. Learners and policymakers need access to assessment and education experts who can assist them in the diagnosis of learning problems and the identification of ways of overcoming them. At the heart of SDG4 are learners. The framework of Goals, Targets and Indicators seems to leave them, their teachers, their teaching and learning processes, teaching and learning materials and daily learning assessment practices out of the loop and the chain of activities and processes required for learning.

It has been suggested that formative assessment inside classrooms make an important contribution to learning outcomes. Unfortunately it is unlikely to attract the attention of the SDG international community and international business interested mainly in international comparability. Formative assessment occurs in diverse classrooms with

diverse teachers and diverse learners facing myriad local learning and teaching challenges. Its tools and results cannot be homogenised. Nonetheless it is a key assessment tool in every classroom worldwide. Perhaps all governments could be persuaded to start reporting whether formative assessments are implemented with what frequency and at what stage, whether formative assessment is embedded within curriculum materials and in teacher education programmes and what actions might be taken in the future.

Acknowledgements

I am grateful for comments on an earlier version of this paper to Keith Lewin, Steve Packer and reviewers. I am also grateful to colleagues and students at the Center for the Study of International Cooperation in Education at the University of Hiroshima for comments on lectures delivered on aspects of this paper during 2018.

References

Assessment Reform Group (1999). *Assessment for Learning: beyond the black box*. Cambridge: University of Cambridge, School of Education.

Barber, M., Chijoke, C., & Mourshed, M. (2010). *How the world's most improved school systems keep getting better*. https://www.mckinsey.com/industries/social-sector/our-insights/how-the-worlds-most-improved-school-systems-keep-getting-better (accessed on December 18, 2018).

Black, P., & Wiliam, D. (1998, October). Inside the Black Box: raising standards through classroom assessment. *Phi, Delta, Kappa*, 1-13.

Bold, T., Filmer, D., Martin, G., Molina, E., Stacy, B., Rockmore, C., Svensson, J., & Wane, W. (2017). Enrollment without learning: Teacher effort, knowledge, and skill in primary schools in Africa. *Journal of Economic Perspectives*, *31*(4), 185-204.

Boston, C. (2002). The concept of formative assessment. *ERIC Digest ERIC* Clearinghouse on Assessment and Evaluation (pp.8). College Park, MD 20742.

Bruns, B. (2018). *Three years after SDG adoption: it's time for action on learning data*. https://www.riseprogramme.org/node/658 (accessed December 18, 2018) .

Carpmail, M., Burnett, L., Chapman, K., & Crowder, D. (n.d.) https://www.google.co.uk/#q=misconceptions+circular+carpmail (accessed 27 October 2016).

Cueto, S., Leon, J., Sorto, M.A., & Miranda, A. (2017). Teachers' pedagogical content knowledge and mathematics achievement of students in Peru. *Educational Studies in Mathematics*, *94*(3), 329-345.

Hansen, A. (Ed.) (2014). *Children's errors in mathematics: understanding common misconceptions in primary schools*. 3rd Edition, London: Sage.

Hanushek, E., & Woessmann, L. (2008). The role of cognitive skills in economic development. *Journal of Economic Literature*, *46*(3), 603-668

Inter-Agency and Expert Group on SDG Indicators (IAEG-SDGs). (2016). Annex IV- Final

list of proposed Sustainable Development Goal indicators. *Report of the Inter-Agency and Expert Group on Sustainable Development Goals Indicators*. (E/CN.3/2016/2/Rev.1). https://sustainabledevelopment.un.org/content/documents/11803Official-List-of-Proposed-SDG-Indicators.pdf

International Commission on Financing Global Education Opportunity. (2016). *The Learning Generation: investing in education for a changing world*. New York http://report.educationcommission.org/wp-content/uploads/2016/09/Learning_Generation on_Full_Report.pdf (accessed April 20, 2019)

Lewin, K. M. (2015a). Educational Access, Equity and Development: Planning to Make Rights Realities. *Fundamentals of Educational Planning Serial Number 97*. Paris: International Institute for Educational Planning, UNESCO.

Lewin, K. M. (2015b). *Goals and Indicators for Development: consolidating the architectures*. New York: Open Society Foundations. https://www.opensocietyfoundations.org/sites/default/files/lewin-goals-indicators-edu-dev-20150515.pdf (accessed 27 November 2018)

Lewin, K. M. (2018). *Learning Matters and the World Development Report 2018*. UK Forum for International Education and Training. London https://www.ukfiet.org/2018/learning-matters-and-the-world-development-report-2018/ (accessed January 30 2019)

Little, A.W. (2008). Education for All: politics, policies and progress. *CREATE Research Monograph* 13, 95. http://www.create-rpc.org/pdf_documents/PTA13.pdf.

Little, A. W., & Green, A. (2009). 'Successful Globalisation, Education and Sustainable Development.' *International Journal of Educational Development, 29*(2), 166-174.

Lockheed, M. (2015). Why Do Countries Participate in International Large-Scale Assessments? The Case of PISA. Policy. *Research Working Paper No. 7447*. Washington DC. World Bank. https://openknowledge.worldbank.org/handle/10986/22875 (accessed May 15 2019).

Nanayakkara, G.L.S. (1992). Assessment of pupil achievement in primary Mathematics with special reference to analysis of pupil errors. Sri Lanka. Unpublished D Phil Thesis, Falmer, University of Sussex.

Newman, M. A. (1977). An analysis of sixth-grade pupils' errors on written mathematical tasks. *Victorian Institute for Educational Research Bulletin*, *39*, 31-43.

Newman, M. A. (1983). *Strategies for diagnosis and remediation*. Sydney: Harcourt, Brace Jovanovich.

OECD. (2012). *Better skills, better jobs, better lives*. Paris: OECD https://www.oecd-ilibrary.org/content/publication/9789264177338-en (accessed December 18, 2018).

Pearce, D., (2007). Sustainable consumption. In D. A. Clark (Ed.), *The Elgar Companion to Development Studies*. Cheltenham, UK: Edward Elgar, pp.612-615.

Radatz, H. (1979). Error Analysis in Mathematics Education. *Journal for Research in Mathematics Education*, *10*(3), 163-172.

Resnick, L.B., Nesher, P., Leonard, F., Magone, M., Omanson, S., & Peled, I. (1989). Conceptual Bases of Arithmetic Errors: The Case of Decimal Fractions. *Journal for Research in Mathematics Education*, *20*(1), 8-27.

Ryan, J., & Williams, J. (2007). *Children's mathematics 4-15: learning from errors and misconceptions*. Maidenhead: Open University Press, McGraw-Hill Education

Trevino, E., & Ordenes, M. (2017). Exploring Commonalities and Differences in Regional and International Assessments. *Information Paper No. 48*. Montreal: UNESCO Institute for Statistics.

United Nations Conference on Environment and Development (UNCED). (1992). *Agenda 21*, Rio de Janeiro, June 3-14.

United Nations Conference on Sustainable Development (UNCSD) (2012). *The future we want*. https://www.un.org/ga/search/view_doc.asp?symbol=A/RES/66/288&Lang=E (accessed May 15 2019)

United Nations (2015). *Sustainable Development Goals, Targets and Indicators*. New York https://unstats.un.org/sdgs/indicators/Global%20Indicator%20Framework%20after%20refinement_Eng.pdf (accessed December 18, 2018).

UNESCO. (2007). *The UN Decade for Education for Sustainable Development (DESD 2005–2014): the first two years*. Paris: UNESCO.

UNESCO. (2014). *The Global Monitoring Report on Education: teaching and learning*, Paris: UNESCO https://en.unesco.org/gem-report/report/2014/teaching-and-learning-achieving-quality-all (Accessed December 18, 2018)

Waage, J., Banerji, R., Campbell, O., Chirwa, E., Collender, G., Dieltiens, V., Dorward, A., Godfrey-Faussett, P., Hanvoravongchai, P., Kingdon, G. Little, A., Mills, A., Mulholland, K., Mwinga, A., North, A., Patcharanarumol, W., Poulton, C., Tangcharoensathien, V., & Unterhalter, E. (2008). The Millennium Development Goals: a cross-sectoral analysis and principles for goal setting after 2015: Lancet and London International Development Centre Commission. *The Lancet*, *376*(9745), 991-1023. https://www.thelancet.com/commissions/mdgcommission (accessed December 18, 2018).

World Commission on Environment and Development (WCED). (1987). *Our Common Future*. Oxford University Press, Oxford.

White, A. L. (2009). *Counting on 2008: Final report*. Sydney: Curriculum K-12 Directorate, Department of Education and Training.

White, A. L. (2005). Active Mathematics in Classrooms: Finding Out Why Children Make Mistakes - And Then Doing Something to Help Them. *Square One*, *15*(4), 15-19.

White, A. L., & M. A. Clements (2005). 'Energising upper-primary mathematics classrooms in Brunei Darussalam: The Active Mathematics in Classrooms (AMIC) Project.' In H. S. Dhindsa, I. J. Kyeleve, O. Chukwu, & J.S.H.Q. Perera (Eds.). *Future directions in science, mathematics and technical education*. Proceedings of the Tenth International Conference. Brunei: University Brunei Darussalam, pp.151-160.

World Bank. (2018). *World Development Report. Learning to realise education's promise*. http://www.worldbank.org/en/publication/wdr2018 (accessed on December 18, 2018).