

Measuring learning quality in Ethiopia, India and Vietnam: from primary to secondary school effectiveness

Padmini Iyer and Rhiannon Moore

Department of International Development, University of Oxford, Oxford, United Kingdom

This paper examines the way in which learning quality has been conceptualised and measured in school effectiveness surveys conducted by Young Lives, a longitudinal study of child poverty. Primary school surveys were conducted in Vietnam in 2010-11 and Ethiopia in 2012-13, and surveys at upper primary and secondary level were conducted in Ethiopia, India and Vietnam in 2016-17. The paper discusses the design of cognitive tests to assess Maths and reading at primary level, and then focuses on the development of cognitive tests to assess Maths, functional English and transferable skills at upper primary and secondary level. In particular, the paper explores how learning quality can be conceptualised and measured in relation to ‘21st century skills’, which are increasingly seen as an important outcome of secondary education. The challenges of designing cognitive tests to measure and compare learning quality across three diverse country contexts are also explored.

Keywords: learning quality; school effectiveness; primary education; secondary education; 21st century skills

Introduction

‘Learning quality’ has been conceptualised and measured in numerous ways, from a focus on objective indicators of ‘inputs’ (financial and human resources invested in education) to ‘outcomes’ (school tests, international curriculum-based or skills-based tests such as TIMSS and PISA) (Scheerens, Lutyen & van Ravens 2011). The emphasis on quality learning outcomes in Sustainable Development Goal (SDG) 4 has also led to increased interest in how learning quality can be conceptualised, measured and compared across diverse contexts. In this paper, we consider learning quality in terms of the ‘value-added’ by schools, using a school effectiveness model. While measuring the

value-added by schools has certain limitations (for example, by focusing only on certain subjects), a school effectiveness approach goes beyond cross-sectional measures of learning quality used in large-scale assessments such as a PISA and TIMSS, and allows us to attribute learning progress to schools and teachers, after controlling for prior attainment and background effects (Rolleston 2013).

This paper examines the way in which learning quality has been conceptualised and measured in school effectiveness surveys conducted by Young Lives at primary level in Vietnam (2010-11) and Ethiopia (2012-13), and at upper primary and secondary level in Ethiopia, India (Andhra Pradesh and Telangana) and Vietnam in 2016-17. In the following sections, we briefly introduce the Young Lives study, and then provide an overview of the Young Lives primary school effectiveness research in Vietnam and Ethiopia. After a brief discussion of the design of cognitive tests to assess Maths and reading at primary level, the paper focuses on the development of cognitive tests to assess Maths, functional English and Transferable Skills at upper primary and secondary level in the 2016-17 school surveys. In particular, we discuss the ways in which we have conceptualised and measured learning quality in relation to ‘21st century skills’, which are increasingly seen as an important outcome of secondary education (OECD 2015), and some of the challenges of designing cognitive tests to measure and compare learning quality across three diverse country contexts.

School Effectiveness Research

School effectiveness research is based on the assumption that schools affect children’s development, and that there are observable regularities in schools that explain variation in student learning (Reynolds et al 2000; 2011). Students are tested at the beginning and end of the school year in school effectiveness research, which allows the progress they

make over one academic year to be measured. The test and re-test design provides a number of analytical advantages when assessing learning quality, particularly in low and middle-income countries where less is known about the educational context (Thomas et al 2016). For example, a second measure of children's achievement in the tested subjects improves the reliability and robustness of measurement, allowing student background and prior student achievement to be controlled for. Moreover, since the two tests are linked, they provide a measure of progress over the course of an academic year rather than a cross-sectional measure of students' learning levels (James 2013). Students' progress can also be used to examine the 'value-added' by schools and teachers, through the use of Item Response Theory (IRT) (Rolleston 2013). Background data on schools and teachers then make it possible to explore some of the factors which lead to more or less effective schools.

The Young Lives Study

Young Lives is an international study of childhood poverty which has followed the lives of 12,000 children in Ethiopia, India (Andhra Pradesh & Telangana), Peru and Vietnam since 2002. The study follows two groups of children in each country – the Younger Cohort born in 2001-2, and the Older Cohort born in 1994-95. This means that we can compare the same children at different ages to see how their lives are changing, as well as different children at the same age, to see how communities have changed over time. In all four countries, a sentinel-site sampling design is employed. The Young Lives sample is not nationally representative; in each country, 20 purposively-selected sites were chosen at the beginning of the study to represent national diversity, with a pro-poor bias (Rolleston et al 2013).

The household survey has been conducted with Young Lives children and their families every three years since 2002, with Round 5 of the survey (the latest round) conducted in 2016-17. Child questionnaires, household questionnaires and community questionnaires gather data on household composition, livelihood and assets, household expenditure, child health, access to basic services, and education.

Young Lives school effectiveness research

In 2010, a school component was introduced to explore Young Lives children's experiences of schooling and education in more depth. The Young Lives primary school survey in Vietnam was conducted in 2011-12, with 3,284 Grade 5 pupils (approximately 1,000 of whom were Young Lives children) in 92 school sites. The survey took place in the 20 Young Lives sites in Vietnam, which are located in five provinces: Ben Tre, Da Nang, Hung Yen, Lao Cai and Phu Yen (Rolleston et al 2013). The Ethiopia primary school survey was conducted in 2012-13, with 9,757 Grade 4 and 5 pupils (approximately 500 of whom were Young Lives children) in 94 schools. The survey took place in 20 Young Lives sites in five regions, Addis Ababa, Amhara, Oromia, SNNP, and Tigray; an additional ten sites in Somali and Afar were included in the school survey to further reflect the cultural and geographic diversity of the country (Aurino, James & Rolleston 2014)¹.

In the Young Lives school surveys, a school effectiveness design involves the administration of cognitive tests at the beginning and end of the school year (referred to as Wave 1 and Wave 2 of data collection respectively). As mentioned above, this design

¹ Young Lives also conducted primary school surveys in India in 2010 (Galab, Reddy & Reddy 2014) and in Peru in 2011 (Guerrero et al 2012), and a secondary school survey in Peru in 2017. As these surveys did not follow a school effectiveness design, they are not discussed in detail in this paper.

allows us to consider the progress made by students over one academic year. (James 2013).

Young Lives primary school surveys: Vietnam and Ethiopia, 2010-13

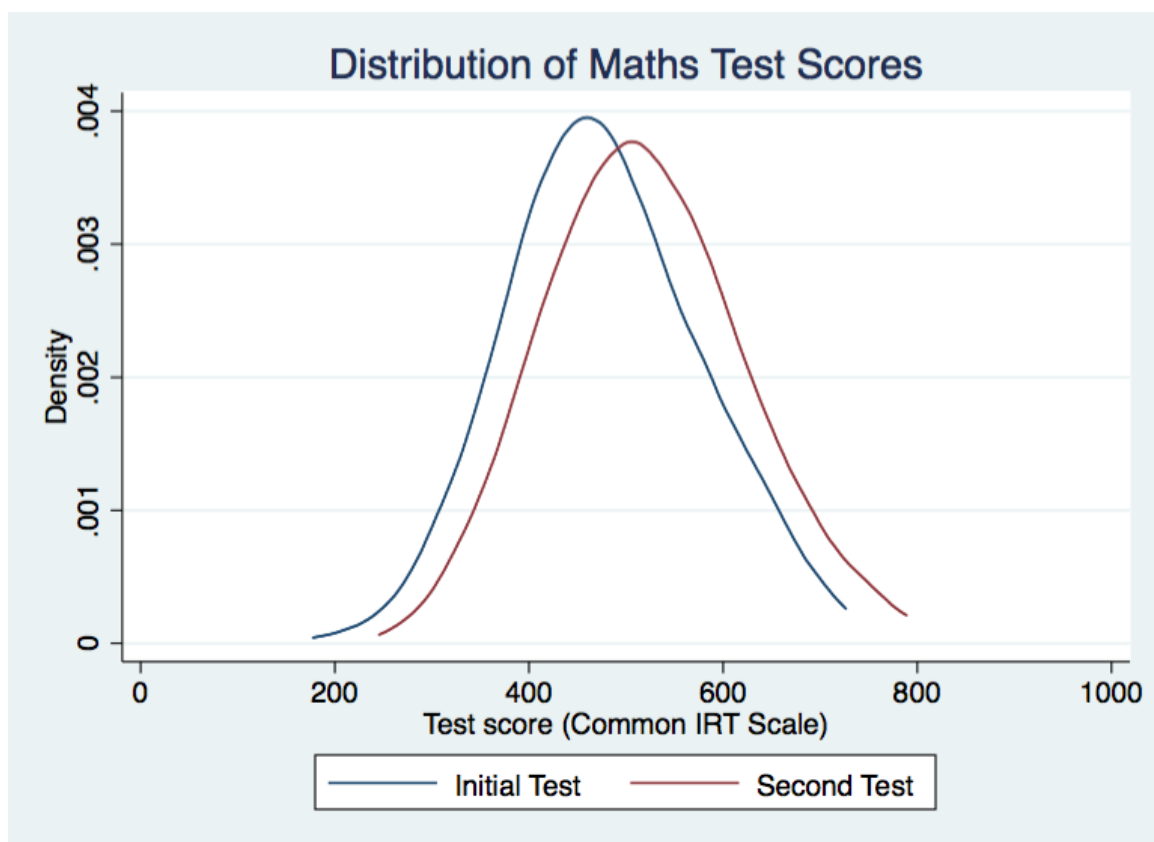
The Vietnam and Ethiopia primary school surveys provide important insights into learning quality both in terms of pupils' learning levels and pupils' learning progress over the course of one academic year. For example, in their analysis of Maths competency levels in Ethiopia², James & Rolleston (2015) found that 82% of Grade 4 and 5 pupils performed two to three grade levels lower than their expected Maths performance level; only 2.5% of pupils were at the level expected for their grade (James & Rolleston 2015). While similar competency level analysis has not been carried out for the Vietnam primary school survey data, 63% of Vietnamese pupils were able to answer Grade 4 items correctly at the start of Grade 5, while 52% of pupils were able to answer Grade 5 items correctly at the end of the school year (Rolleston et al 2013).

To consider learning quality in terms of progress and school effectiveness, an IRT scale was used to scale the tests in Ethiopia and Vietnam so that they had a mean of 500 and standard deviation of 100 (Rolleston 2013). It is important to note that the tests in the two countries did not have common items, and so progress cannot be compared directly across the countries – one point of progress in Vietnam is not equal to one point of progress in Ethiopia. When Maths progress was assessed on tests linked to country-specific curricula (as discussed in more detail below), pupils in Vietnam progressed by 40 points over the course of one grade (Rolleston et al 2013), and pupils in Ethiopia by 30 points (James & Rolleston 2015), both of which are statistically significant average

² James & Rolleston (2015) define Maths competency levels for Grades 4 and 5 using the actual item difficulty within the tests as well grade-level expectations from the curriculum. Pupils who scored correctly on two-thirds of items in a particular competency level were considered to be at that level, providing they also reached required competency levels below this.

‘learning gains’. While there is progress in Maths performance in both countries, it is important to consider these findings in light of the learning levels discussed above. Pupils in Ethiopia made significant progress over one year, but the majority were still two or three grades below their expected performance level at the end of the year (James & Rolleston 2015). By contrast, in Vietnam, over half of the pupils were performing at grade level by the end of Grade 5 (Rolleston et al 2013).

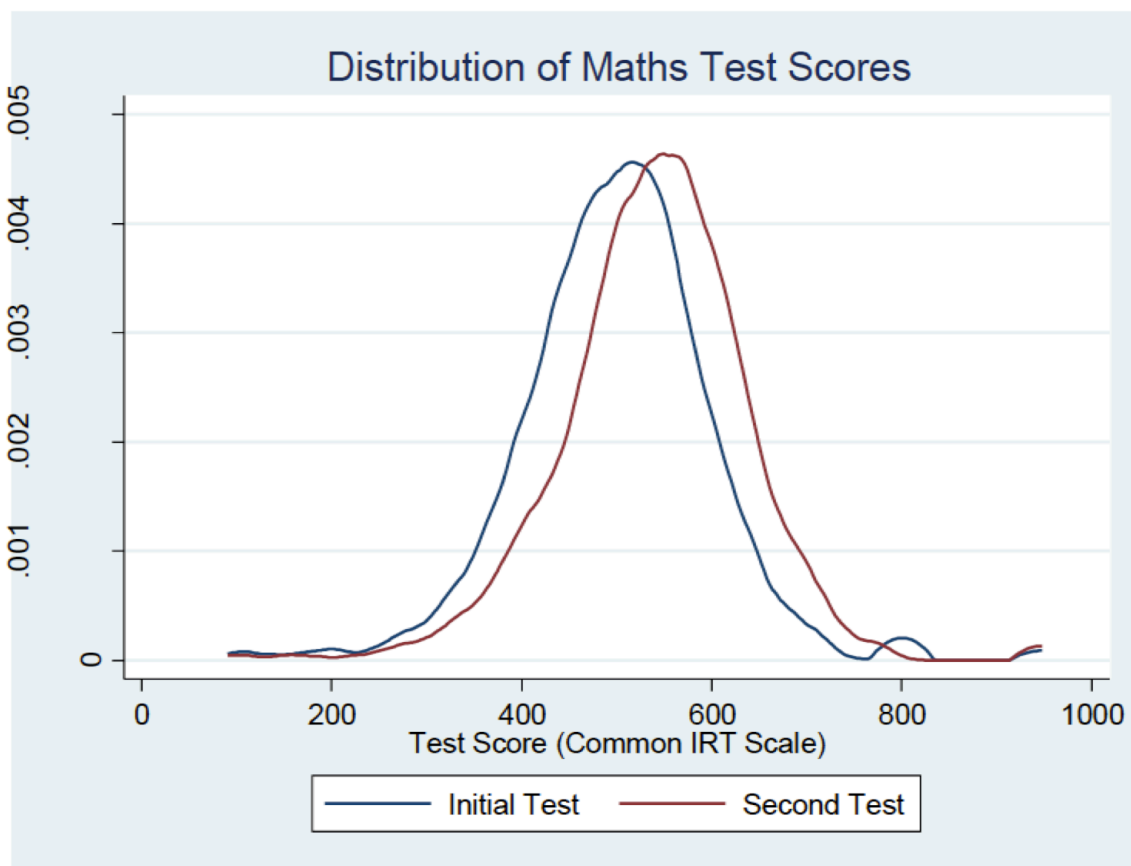
Figure 1: Distribution of Maths Test Scores, Vietnam Grade 5 (Rolleston 2013)



The progress on IRT scores can also be used to examine the ‘value-added’ by schools and teachers (Rolleston 2013). The value-added has been calculated from the difference between pupils’ actual end of year scores and their ‘expected scores’ based on the whole sample aggregated at the school level (Rolleston 2013). Results from the Ethiopia primary school survey indicate that, even when controlling for prior attainment and background effects, schools do add value in terms of pupil learning in Maths and

reading. Notable characteristics of high value-added schools included teachers who scored highly on the teacher Maths test (associated with high value-added for Maths), and teachers who had a university education (associated with high value-added for Maths and reading). Meanwhile, characteristics of low value-added schools included those which only taught shift classes (associated with low value-added for Maths and reading) (James & Rolleston 2015).

Figure 2: Distribution of Maths Test Scores, Ethiopia Grades 4-5 (James & Rolleston 2015)



In Vietnam, pupils from the most disadvantaged home backgrounds attended schools with only slightly lower than average pupil progress; this suggests that schools catering to disadvantaged pupils do add value. The schools which add the most value in terms of Maths and reading progress tend to have better facilities, for example separate classrooms for Grade 5, working electricity, and a higher proportion of teachers

qualified to degree level. These schools were also more likely to be selective schools, i.e. less likely to admit all pupils who apply (Rolleston et al 2013). In terms of class-level factors, higher value-added was associated with higher levels of class assets and facilities, fewer children who arrive late each day, and teachers with permanent rather than temporary contracts. Finally, teachers in high-performing classes tended to demonstrate a higher sense of their own ability to improve children's learning (Rolleston et al 2013).

Young Lives upper primary and secondary school surveys: Ethiopia, India and Vietnam, 2016-17

In 2016-17, Young Lives conducted further rounds of school effectiveness surveys in the Young Lives sites in Ethiopia and Vietnam, and additionally in the 20 Young Lives sites in Andhra Pradesh and Telangana, India. In each country, school surveys in 2016-17 focused on the level of schooling accessed by the majority of 15-year-olds, the age of Young Lives Younger Cohort children that academic year. In Ethiopia, the survey was therefore conducted with Grade 7 and 8 students (upper primary level); in India, with Grade 9 students (lower secondary level); and in Vietnam, with Grade 10 students (upper secondary level). The 2016-17 school survey sample reflects the educational context and structures in each country, and the Wave 1 sample is as follows:

- Ethiopia: 64 upper primary schools, approx. 12,000 students
- India: 205 secondary schools, approx. 10,000 students
- Vietnam: 52 upper secondary schools, approx. 9,000 students

In each country, the sample includes Young Lives Younger Cohort children and their classmates, a feature which allows us to link to the previous circumstances of the Young Lives children themselves, and additionally to examine classroom effects within

the survey (see Rossiter 2016, Moore 2016, and Iyer 2016 for more details on the Ethiopia, India and Vietnam school survey samples respectively).

The Young Lives primary and secondary school surveys in 2011-13 and 2016-17 broadly followed the school effectiveness design described above. Cognitive tests (Maths and reading comprehension at primary level; Maths and English at secondary level) were administered to students at the beginning and the end of the school year, while background instruments (principal, teacher and student questionnaires; school facilities observations) collected data to contextualise the learning progress made by students. Psychosocial measures for students and teachers, for example relating to academic self-concept and motivation, were included at the end of the year. In the secondary school surveys, a one-off Transferable Skills test was also administered at the end of the school year in India and Vietnam, and a functional Amharic test was administered in Ethiopia³.

While cognitive tests were linked over time in the Vietnam and Ethiopia primary school surveys, in the secondary school surveys the tests were linked via common items both over time and across all three countries. The use of these linked tests allow us to measure progress precisely through points on an interval scale. In the following sections, we discuss in more detail the conceptualisation and measurement of learning quality through these cognitive tests in the Young Lives school surveys.

³ In the secondary school surveys, cognitive tests were administered in the language(s) most suitable to the context in each country. In Vietnam, tests were administered in Vietnamese; in India, tests were administered in bilingual Telugu/English or Urdu/English format, according to the medium of instruction in each school. In Ethiopia, tests were administered in Af Oromo, Tigrigna, English, bilingual Amharic/English, or bilingual Af Somali/English, according to the medium of instruction in each region.

Conceptualising and measuring ‘learning quality’ at primary and secondary level

Learning quality at primary level

In the primary school surveys in Ethiopia and Vietnam, ‘learning quality’ was understood in terms of progress on the core curricular domains of Maths and reading comprehension. As a result, cognitive tests were designed with links to existing national assessments and Ministry of Education standards, and developed with experienced test consultants in both countries (James 2013; 2014). In Vietnam, reading comprehension tests were administered in Vietnamese, while the diversity of mediums of instruction in Ethiopia led to the development of reading comprehension tests in eight different languages in the Ethiopia school survey – Amharic, Oromiffa, Tigrigna, Sidama, Wolayta, Hadiyya, Afar and Somali (James 2014).

The suitability of items for all tests was determined through qualitative pre-piloting and larger scale piloting in both countries. Item functioning was established by considering the item difficulty (the percentage of children who got each item correct, and each item’s rank in terms of difficulty), and the ‘fit’ of the item as measured through Item Response Theory (IRT) analysis (James 2013). This led to the development of two multiple-choice Maths tests and two multiple-choice language tests for each country, administered at the beginning and end of the year, with ‘anchor’ items (i.e., items directly replicated) between the two tests. Anchor items in all tests were typically higher-level competency items, which supports the analysis of learning progress on items we would expect students to learn via the school curriculum during the academic year (James 2013; 2014).

Learning quality at secondary level

At primary level, it seems appropriate to consider learning quality in terms of the development of basic skills such as Maths and reading, and defining these domains according to the curriculum. However, defining learning quality at secondary level is more complex. Secondary education involves (or ideally should involve) more than the accumulation of curriculum knowledge and the development of basic skills; the stated intention of secondary education in many countries is to equip young people with skills for future labour market or higher education opportunities (World Bank 2009). When designing tests for the secondary school surveys, we therefore started by considering how to conceptualise ‘quality’ learning at secondary level.

A useful starting point was Bloom’s Revised Taxonomy. The six dimensions in Bloom’s Revised Taxonomy (originally developed in 1956 and revised during the 1990s) are ordered from simple (‘Remembering’) to abstract (‘Creating’), and accumulating each cognitive process is required in order to move up the hierarchy (see Kraftwohl 2002: 215-6). Drawing on Bloom’s Revised Taxonomy, Mayer’s (2002) definition of ‘meaningful’ learning is useful when considering what is meant by quality learning. Mayer (2002) defines meaningful learning as ‘not only acquiring knowledge but also being able to use knowledge in a variety of new situations’ (2002: 226). Mayer (2002) identifies ‘retention’ and ‘transfer’ as two essential components of meaningful learning, with retention defined as ‘the ability to remember material at some later time in much the same ways it was presented during instruction’, and transfer as ‘the ability to use what was learned to solve new problems, answer new questions, or facilitate new learning subject matter’ (Mayer 2002: 226).

Evidently, Mayer’s (2002) definition of ‘meaningful’ learning is relevant for quality learning at primary as well as secondary level. However, it was a useful starting

point when defining learning quality in the secondary school surveys for several reasons. Firstly, Mayer's (2002) definition highlights the importance of both curriculum knowledge and the ability to apply this knowledge in new contexts and to develop new skills; knowledge and application were therefore two key constructs for our assessment of learning quality at secondary level. Secondly, Bloom's Revised Taxonomy encouraged us to think about the 'higher order' cognitive processes that we might want to assess specifically at secondary level, which in turn led us to consider 'transferable skills'.

The World Bank (2014) defines 'transferable skills' to include critical thinking, problem solving, communication and teamwork; these skills have also been described as '21st century skills', which are increasingly seen as important outcomes of secondary education (OECD 2015). Transferable skills, which are learned in school and then 'transferred' to the workplace, are essential to quality learning at secondary level as defined by a human capital model – i.e., to ensure 'gains in productivity and competitiveness, economic growth, poverty alleviation and a range of social and health benefits' through education (Rolleston 2016). This is the 'next phase' of quality education, which is particularly of concern in India and Vietnam. In both countries, competency-based curricula are in development. In India, the New Education Policy drafted in 2016 seeks to ensure that schools equip students with 'life skills' such as creativity, critical thinking and problem solving suited to the requirements of the developing 'knowledge-based' economy (MHRD 2016). Meanwhile, in Vietnam, a new curriculum under the General Renovation of Education seeks to develop skills such as communication, teamwork and problem solving through learner-centred approaches at primary and secondary level (World Bank 2015).

Measuring learning quality at secondary level

When measuring learning quality in Ethiopia, India and Vietnam, a key challenge was to design cognitive tests that assessed constructs appropriate for diverse contexts within each country (for example, across rural-urban divides). An additional challenge arose from our aim to create tests which were linked across all three countries, to allow for cross-country comparison of learning quality among 15-year-olds. The three broad domains for the cognitive tests at secondary level were Maths, functional English and Transferable Skills (with a focus on problem solving and critical thinking). As discussed in more detail below, testing Maths enables us to assess both knowledge and application within the same subject domain, while the decision to test functional English, problem solving and critical thinking was linked to our interest in transferable skills, particularly in relation to the labour market.

When developing cognitive tests, we considered the different priorities and areas of interest for each country alongside item-level analysis of data from large-scale pilot testing which took place across all three countries. It is important to note that these pilots were not conducted with samples chosen to be representative of the final survey samples, but rather to help us identify potential ‘floor’ and ‘ceiling’ effects among different groups (e.g. items which might be too easy or too difficult for students in urban vs. rural schools, and/or government vs. private schools). The pilots also enabled us to consider any differential functioning of items – for example, whether there was a systematic difference in difficulty levels of the same items when presented in different languages for different groups in Ethiopia.

Item-level analysis of pilot data was conducted using Classical Test Theory analysis and 2-parameter IRT analysis. Our analysis enabled us to select items for the final Maths and English tests based on the difficulty of each item, the ‘fit’ of the item as

measured by IRT analysis, and the functioning of the distractors (e.g., what the selection of a response tells us about how students of different ability levels understand the item). When assembling the final tests, we also took into consideration the balance of cognitive and content domains being assessed by the test as a whole, as well as including a range of items across suitable grade levels (see Azubuike, Moore & Iyer 2017 for a more detailed discussion of piloting and pilot data analysis).

Maths

Our Maths tests for the secondary school surveys were designed using Grønmo et al's (2015) assessment framework. This allowed us to test students' performance across three cognitive domains:

- **Knowledge**, which covers the facts, concepts, and procedures students need to know
- **Application**, which focuses on the ability of students to apply knowledge and the conceptual understanding to solve problems or answer questions
- **Reasoning**, which goes beyond the solution of routine problems to encompass unfamiliar situations, complex contexts, and multi-step problems (Grønmo et al 2015: 24).

The Maths tests were also based around the appropriate content domains for each country, which were identified using the maths curricula in the three countries:

- Basic number competency
- Integers, rational numbers, powers and bases
- Fractions, decimals, ratios and percentages
- Area, perimeter, volume and surface area

- Geometry and shapes
- Algebra
- Measurement, charts and graphs
- Reasoning, problem solving, and applications in daily life

In collaboration with test consultants in each country and together with Educational Initiatives, a research consultancy in India, specific Maths tests were designed for each country, with seven common items on each test which can be used for cross-country and cross-wave comparison. While our broad interest was in testing students' knowledge (linked to the curriculum), application and reasoning skills, the balance of items assessing these cognitive domains was determined based on different priorities within each country. For example, Young Lives' earlier findings (James & Rolleston 2015) and our own pilot data for the secondary school survey indicated that Ethiopian students are often performing below their expected grade level. Being able to answer grade-appropriate 'knowledge' items correctly would therefore arguably reflect 'quality learning' in Grades 7 and 8 in Ethiopia. As a result, 50% of the Ethiopia Wave 1 Maths test was made up of knowledge items, and the test also included lower grade level items. By contrast, existing Young Lives data (Rolleston et al 2013) and PISA 2012 and 2015 results point to generally high levels of mathematical knowledge among students in Vietnam. 65% of the items on the Wave 1 Vietnam Maths test therefore assessed Grade 10 students' ability to apply their mathematical knowledge or to use mathematical reasoning skills in less familiar contexts – skills which, it is often argued, the Vietnamese education system does not support.

Below are examples of items which assess mathematical knowledge, application and reasoning skills (Figures 3a, 3b and 3c respectively), along with results from pilots conducted in Ethiopia, India and Vietnam in 2016. These items are all examples of

those which functioned well as anchor items, as there is overlap in students' performance across the three countries, and therefore allow the tests to be put on the same scale. These items, along with nine other items which functioned in a similar way, were therefore selected as cross-country anchor items for Wave 1 of the secondary school surveys.

The pilot data for the items presented here broadly suggest that the reasoning item was harder than the knowledge and application items in Ethiopia and Vietnam, although the Grade 6 knowledge item in Figure 3a proved very difficult in Ethiopia and India. This is particularly notable since it is an item assessing Grade 6 level knowledge (and in a familiar, textbook format) which we would expect Grade 7 and 8 students in Ethiopia and Grade 9 students in India to have learned by this stage. The items also broadly reflect the differences that we would expect across the countries, with Vietnamese students at the higher end of the ability range, Indian students generally in the middle, and Ethiopian students towards the lower end.

Figure 3a: Cross-country item assessing mathematical knowledge

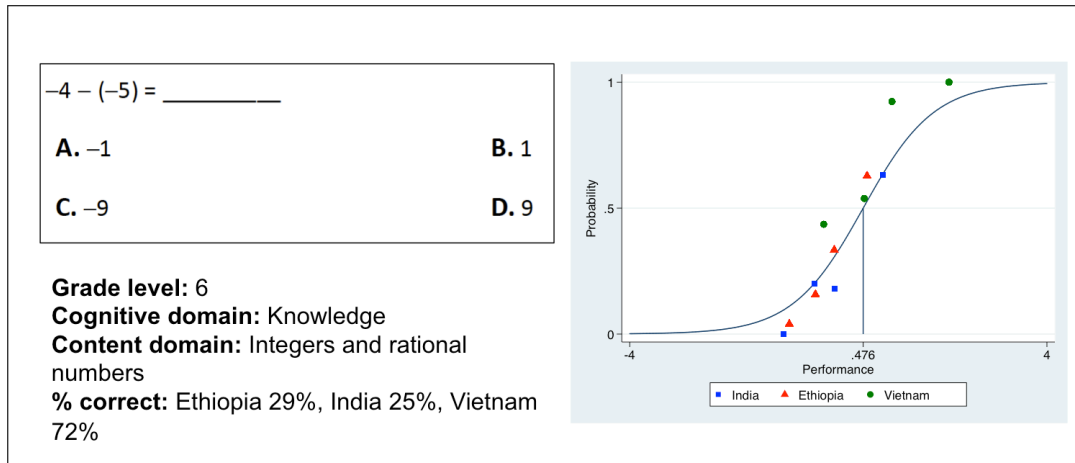


Figure 3b: Cross-country item assessing mathematical application

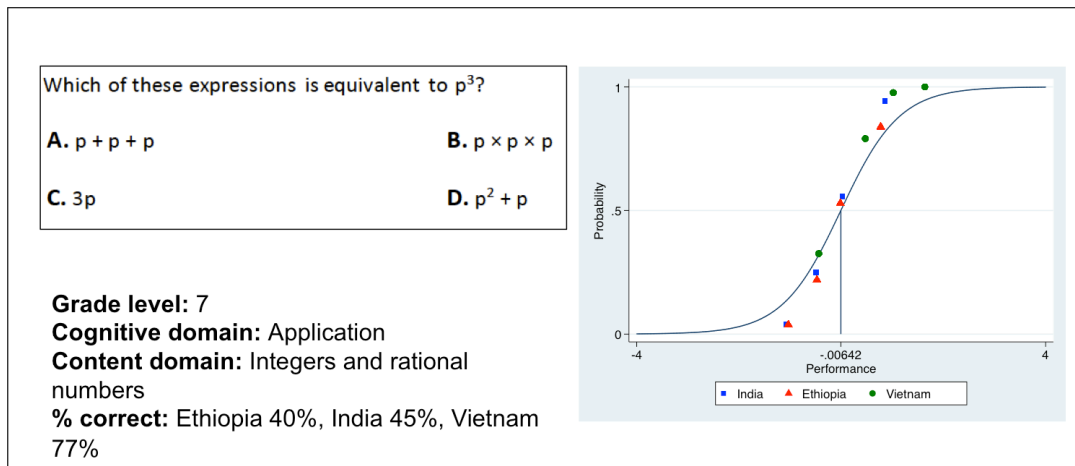
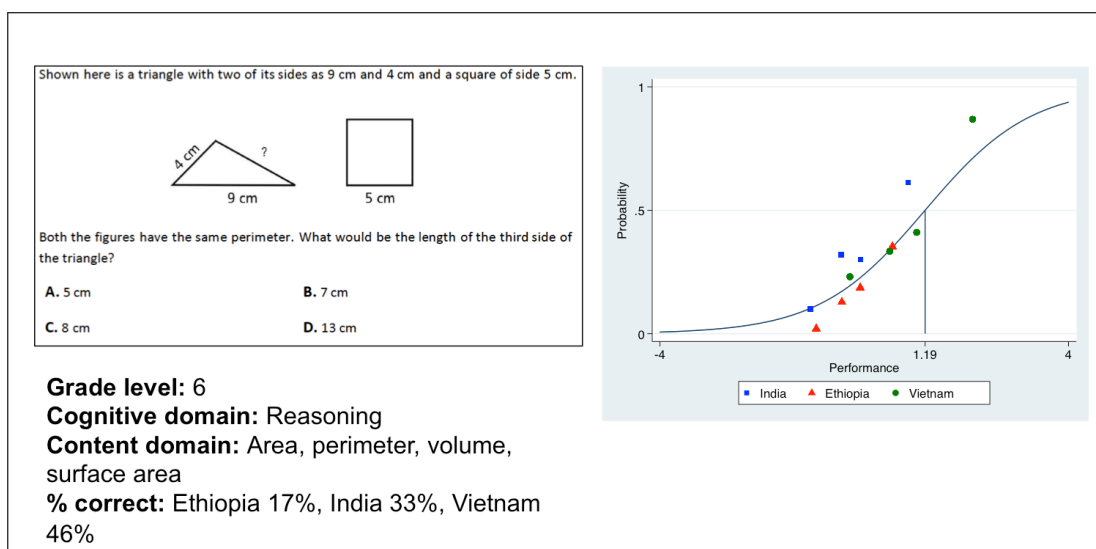


Figure 3c: Cross-country item assessing mathematical reasoning



Functional English

We included an English language test in the secondary school surveys due to the status of English as a transferable skill in all three study countries, with relevance for continuing education, labour market opportunities and social mobility in an increasingly globalised economy (Graddol 2010). Although the contexts in which children are exposed to English (both in and out of school) differ in Ethiopia, India and Vietnam, the language is perceived to have relevance to their lives by education stakeholders and policymakers in all three countries.

There are multiple ways of studying English, depending on the context in which a learner is exposed to the language and their reasons for learning it. For example, a student attending an English medium school system in an environment where they have little or no exposure to English outside school will learn differently from a student who is learning English and is frequently exposed to English media and people who speak some English. Both students need to learn the language, but they have different objectives and therefore different understandings of what ‘successful’ learning progress might look like. To ensure the relevance of our English test across diverse contexts, we aligned it to the Common European Framework of Reference for Languages (CEFR), which is based on the assumption that students learn a language by completing various ‘tasks’ relevant to their context, needs, resources and characteristics. CEFR defines six levels of language proficiency (known as the Common Reference Levels) based on what learners are able to do, ranging from A1 (most basic) to C2 (most advanced) (Council of Europe 2001).

In order to develop English tests for the secondary school surveys, we began by identifying the overarching construct to be tested, based on our understanding of how English is learned and used in the three study country contexts. The construct we aimed

to test was ‘functional English’, which has been defined as the “application of [...] skills in purposeful contexts and scenarios that reflect real-life situations” (OFQUAL 2011: 10). In this sense, our English test diverged somewhat from country-specific school curricula (unlike the Maths test), in light of our aim to capture skills which have relevance for students now or in their future. Due to practical and logistical considerations for conducting a large-scale survey, the test was limited to multiple-choice reading items, which meant that it only captured one dimension of the functional English construct.

Within this broader construct, we identified four skill areas to assess, working in collaboration with Educational Initiatives (as with the development of the Maths test). These are:

- **Word identification:** Identifying simple vocabulary to which students are likely to have been exposed, with particular focus on language relating to their everyday environment and to education.
- **Word meaning, knowledge and contextual vocabulary:** Identifying the meaning of unfamiliar words through their contextualised use in a sentence, or through identifying a synonym/antonym.
- **Sentence construction and comprehension:** Completing sentences correctly, using appropriate grammatical concepts, and combining sentences together.
- **Reading comprehension:** Reading a range of texts (stories, posters, factual passages) and comprehending both direct facts and implicit inferences from them.

As the functional English construct is slightly divorced from school curricula, we were able to identify a larger number of common items to be used across all three

countries. 10 common items were used in all three countries across Wave 1 and Wave 2, with additional items common between two countries in one or both waves. The assessment used in India was longer (50 items) than in Ethiopia and Vietnam (40 items), as we anticipated a greater range of ability levels within the India sample, which would be best captured by including items from a wider range of difficulty levels.

The composition of the tests varied across the three countries according to country-specific priorities and contexts, and in light of pilot item functioning. For example, in Ethiopia and Vietnam, we anticipated lower levels of exposure to English outside the classroom. Pilot data confirmed that items between CEFR levels A1-A2 functioned best, and so the Wave 1 test forms contained a larger proportion of ‘word identification’ items (25% and 20% respectively) with a lower level of difficulty. Meanwhile, pilot findings from India indicated slightly higher English levels (between CEFR levels A2-B2), which similarly matched our expectations from existing reports (e.g. Graddol 2010). As a result, the India Wave 1 test contained a smaller proportion of word identification items (8%) and a greater proportion of word meaning, knowledge and contextual vocabulary items (24%). Across all three countries, reading comprehension items functioned differently according to the difficulty level of the text and the type of question. The item shown in Figure 4c requires an answer which can be taken directly from the text (and as a result has a fairly low level of difficulty in India and Vietnam), while those items requiring more inference or understanding of the text had a higher level of difficulty in all of the countries (see Figure 4b). These three items largely reflect the levels of performance we anticipated, with India scores towards the higher end and Vietnam and Ethiopia towards the middle and lower end, and some overlap in scores between the three countries.

Figure 4a: Cross-country item assessing functional English – sentence construction

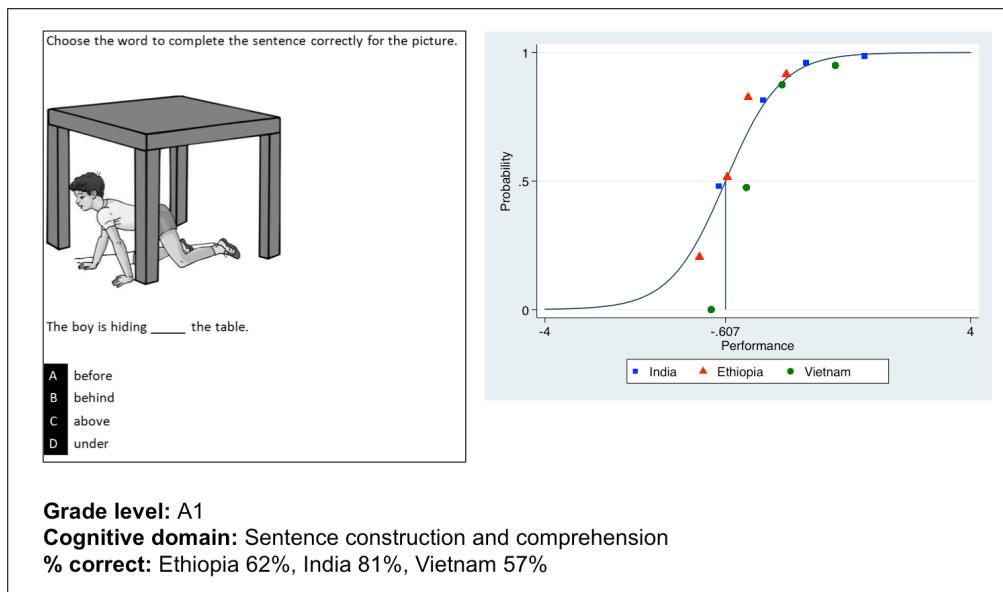


Figure 4b: Cross-country item assessing functional English – word meaning

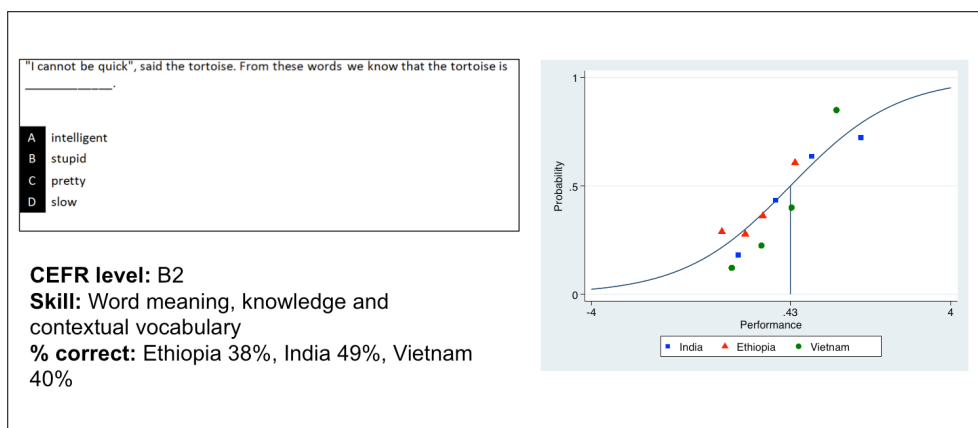
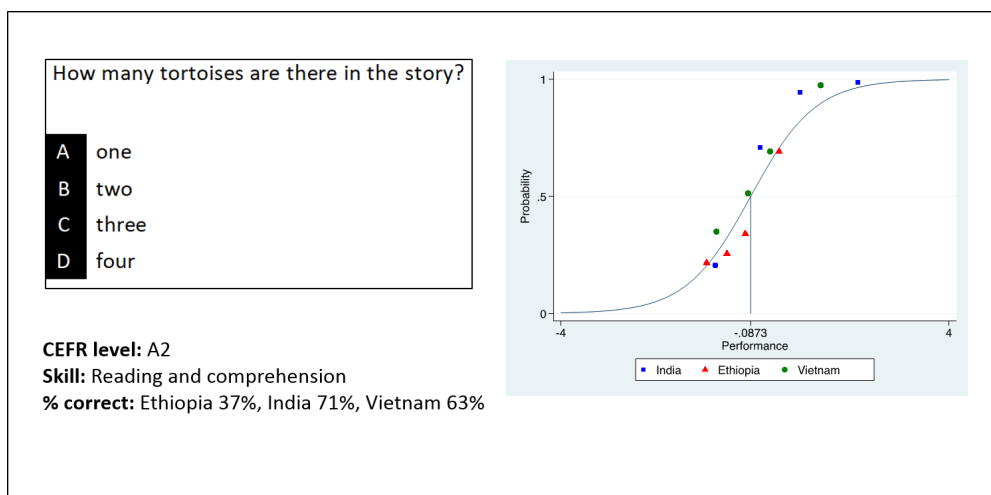


Figure 4c: Cross-country item assessing functional English – reading comprehension



Transferable Skills

In addition to repeated measures Maths and English tests, we designed a cross-sectional cognitive test to assess two specific ‘transferable skills’: problem solving and critical thinking. These are cross-curricular skills which can be developed across school subjects and in ‘real-life’ situations (Greiff, Holt & Funke 2013), rather than subject-specific skills. Problem solving can be defined as ‘an individual’s capacity to use cognitive processes to resolve real, cross-disciplinary situations where the solution path is not immediately obvious’ (OECD 2003, in Greiff, Holt & Funke 2013: 74). While problem solving may involve grappling with a complex problem, there is usually a definitive solution to be determined; by contrast, critical-thinking skills are those such as inference and evaluation which are applied to ill-structured problems, and for which ‘there are no definitive solutions’ (Kuhn 1991; Thomas & Lok 2015).

PISA remains the key source of problem-solving assessments for school students, particularly at age 15, and across international contexts. The PISA 2012 and PISA 2015 problem-solving assessments have tested more sophisticated types of problem solving, such as interactive and collaborative problem-solving skills, but these tests require computer-administered, adaptive testing. By contrast, the PISA 2003 assessments are pen-and-paper tests, which are more suited to our contexts and survey methods.

Problem-solving assessments from PISA 2003 were adapted for small-scale, qualitative pre-pilots in all three countries (see OECD 2003 for original PISA items). In some cases, we adapted the format of the items – for example, changing open-ended responses to multiple-choice responses, as we did not have the capacity to mark and analyze data from open-ended responses. In other cases, we adapted the content so that it was more suited to low-income, rural contexts – for example, an item based around

different menus was adapted to list local, more familiar types of food. Basic comprehension questions were also added to the problem-solving exercises, to enable us to distinguish between students' ability to understand the text and their ability to use problem-solving skills.

To assess critical thinking, we adapted CWRA+ (College Work Readiness Assessment) selected-response critical thinking items, which have been developed for use in the United States with middle school and high school students by the Council for Aid to Education (CAE 2015). Our process of adaptation for these items broadly followed the methodology adopted by Schendel & Tolmie (2016) in their adaptation of College Learning Assessment (CLA) items to assess critical thinking skills among college students in Rwanda. In particular, passages and response options were adapted to ensure names, locations and so on were relevant for each context in which they would be used.

Qualitative pre-pilots were conducted in all three countries to determine how suitable the adapted problem-solving items were for each of the contexts, and in India and Vietnam to determine the suitability of adapted critical-thinking items. Initial findings from these pre-pilots were encouraging in India and Vietnam, but less so in Ethiopia. While students in Ethiopia were able to answer basic comprehension questions, only a small proportion of students were able to answer problem-solving items correctly. Based on these pre-pilot findings, and a lack of policy emphasis on transferable skills in Ethiopia, it was therefore decided not to include problem-solving and critical-thinking assessments in the 2016-17 school survey in Ethiopia. Large-scale pilots of the Transferable Skills tests were therefore only conducted in India and Vietnam.

One problem situation adapted from PISA 2003 was ‘Journey’; in the original item, students were presented with a metro map, which we adapted to a bus map in order to be more accessible for students in our contexts. In response to a newly-added comprehension question for this item (which required students to retrieve basic information from the bus map), 64% of students in India ($n = 112$) and 91% of students in Vietnam ($n = 176$)⁴ answered this question correctly. The problem-solving questions for this item required students to determine the single route between two given points which is both quickest and cheapest. Students were given partial credit if they answered one of these items correctly, and full credit if they answered both items correctly. 41% of students in India and 52% of students in Vietnam gained partial credit on these items, with 10% and 28% respectively gaining full credit. These pilot results indicated that this problem situation functioned well in both countries, with a high proportion of students able to understand the basic information presented, a lower proportion demonstrating mid-level problem-solving skills, and an even smaller proportion demonstrating high-level problem-solving skills. This problem situation, along with two others in India and three others in Vietnam, was therefore included in the final Transferable Skills test administered at Wave 2.

In terms of critical thinking, a comprehension question added to one of the passages adapted from CWRA+ was answered correctly by 37% of students in India ($n = 107$) and 70% of students in Vietnam ($n = 179$). Distractor analysis conducted on the

⁴ IRT analysis was not used to analyze Transferable Skills pilot data, as one of the assumptions of IRT is that items are locally independent (i.e., responding correctly to one item is not statistically related to other items in the test). Classical Test Theory was therefore used to analyze this pilot data instead (see Iyer and Azubuike, forthcoming for a more detailed discussion of Transferable Skills pilot data analysis).

India pilot data revealed that 41% of students chose option A rather than C, the correct option. This indicated that answer option A was too ‘distracting’, and it was therefore revised for the final version of the Transferable Skills test in India to make it less distracting for students. Meanwhile, when responding to a critical thinking item for the same passage (which assessed students’ ability to evaluate the reliability of information provided), 31% of students in India and 59% of students in Vietnam answered the item correctly. Overall, critical thinking pilot results allowed us to further adapt items to ensure that they were appropriate for each country context, and in spite of relatively low pilot scores (particularly in India), two passages assessing students’ ability to ‘critique an argument’ and ‘critical reading and evaluation’ skills respectively functioned well enough to be included in the final Transferable Skills test in India and Vietnam.

Discussion

This paper has discussed the conceptualisation and measurement of learning quality at primary and secondary level across three diverse country contexts. At primary level in Vietnam and Ethiopia, learning quality in the Young Lives school surveys was understood in terms of basic numeracy and literacy skills; as a result, cognitive tests were linked to the Vietnamese and Ethiopian curricula respectively, and designed to measure progress on the core domains of Maths and reading comprehension. However, at secondary level, learning quality is ideally associated with more than the accumulation of basic skills and knowledge. Cognitive tests for the secondary school surveys in 2016-17 were therefore designed to assess ‘transferable’ or ‘21st century’ skills both within curriculum domains (Maths), and in cross-curricular domains (functional English, problem solving and critical thinking). Developing cognitive tests for 15-year-olds across different educational levels (upper primary, lower secondary and

upper secondary) and across diverse contexts in Ethiopia, India and Vietnam has also led to a more nuanced conceptualisation of learning quality; tests in all countries focusing on knowledge and the transfer of knowledge, but to varying extents. For example, there was a greater focus on application of knowledge in the Vietnam Maths tests; there was a wider range of difficulty levels on English tests in India; and instead of problem solving and critical thinking, Functional Amharic was assessed as a more relevant transferable skill in Ethiopia.

As discussed above, by following a school effectiveness design, the Young Lives primary school surveys in Vietnam and Ethiopia provided important insights into learning levels, learning progress, and the ‘value-added’ by one year of primary school (Rolleston et al 2013; James & Rolleston 2015). As in the primary school surveys, the school effectiveness design in the 2016-17 surveys means that, at upper primary and secondary levels, we will be able to consider:

- The progress students make over the course of one academic year;
- How these levels of progress are associated with home, school, teacher, class and student factors;
- The characteristics of schools with high and low ‘value-added’.

Moreover, the design of the cognitive tests being used in the secondary school surveys will support new insights into learning quality which are particularly relevant to secondary level, by allowing us to explore:

- ‘Meaningful’ learning: have students reached the levels of knowledge and skills expected for their grade, and to what extent are they able to apply these knowledge and skills in less familiar contexts?

- Transferable skills: to what extent are schools preparing students for the labour market, in terms of functional English language, problem-solving and critical-thinking skills?

Finally, the inclusion of cognitive test items which are common across all three countries enables us to explore learning levels and learning progress in Maths and functional English across Young Lives sites in Ethiopia, India and Vietnam on a common scale. Building upon the extensive Young Lives data which has already been collected in these sites through household and school surveys, this will provide insights into school effectiveness and school systems within and across three diverse country contexts.

Acknowledgements: The work discussed in this paper has been carried out by the authors together with the Young Lives education team – Caine Rolleston, Jack Rossiter, Bridget Azubuike and Zoe James. We would like to thank Caine Rolleston and Jack Rossiter for their comments on an early draft of this paper, and the reviewer for their useful suggestions for improvements to the paper during peer review.

References

- Aurino, E., Z. James & C. Rolleston. 2014. Young Lives Ethiopia school survey 2012-13: data overview report. Working Paper 134. Oxford: Young Lives.
- Azubuike, O. B., R. Moore & P. Iyer. 2017. The design and development of cross-country Maths and English tests in Ethiopia, India and Vietnam. Technical Note. Oxford: Young Lives.
- Council for Aid to Education. 2015. CWRA+: Technical FAQs. New York, NY: Council for Aid to Education.
- Council of Europe. 2001. Common European Framework for Reference Of Languages: Learning, Teaching, Assessment. Strasbourg: Language Policy Unit.

- Federal Ministry of Education. 2010. Education Sector Development Program IV (ESDP IV), 2010/11 – 2014/15, 2003 EC – 2007 EC. Addis Ababa: Federal Ministry of Education.
- Galab, S., Prudhika Reddy, P., & V. N. Reddy. 2014. Classroom process, teacher ability and student performance: evidence from school-based component of Young Lives in Undivided Andhra Pradesh. Working Paper 134. Hyderabad: Centre for Economic and Social Studies.
- Graddol, D. 2010 English Next: India. London: British Council.
- Greiff, S., D. V. Holt, & J. Funke. 2013. “Perspectives on problem solving in educational assessment: analytical, interactive, and collaborate problem solving”. *The Journal of Problem Solving* 5 (2): 71-91.
- Grønmo , L. S., M. Lindquist, A. Arora & I. V. S Mullis. 2015. TIMSS 2015 Mathematics Framework. Boston, MA: TIMSS & PIRLS International Study Centre.
- Guerrero, G., Leon, J., Rosales, E., Zapata, M., Freire, S., Saldarriaga, V. & S. Cueto. 2012. Young Lives school survey in Peru: Design and initial findings. Working paper 92. Oxford: Young Lives.
- Iyer, P. 2016. The design of the 2016-17 Young Lives School Survey in Vietnam. Technical Note 38. Oxford: Young Lives.
- James, Z. 2013. Young Lives school survey: the design and development of achievement tests in the Vietnam school survey, Round 1. Oxford: Young Lives.
- James, Z. 2014. Young Lives school survey: the design of achievement tests in the Ethiopia school survey, Round 2 (2012-13). Oxford: Young Lives.
- James, Z. & C. Rolleston. 2015 “School effectiveness in Ethiopia: challenges and opportunities”. Paper presented at the 13th UKFIET International Conference on Education and Development, University of Oxford, 16th September 2015.
- Kraftwohl, D. R. 2002 “A revision of Bloom’s Taxonomy: an overview”. *Theory into Practice* 41 (2): 212-218.
- Kuhn, T. 1991 *The skills of argument*. Cambridge: Cambridge University Press.
- Mayer, R. E. 2002 “Rote versus meaningful learning”. *Theory into Practice* 41 (2): 226-232.
- Ministry of Human Resource Development (MHRD). 2016. Some Inputs for Draft National Education Policy 2016. New Delhi: MHRD, Government of India.

- Moore, R. 2016. The design of the 2016-17 Young Lives School Survey in India. Technical Note 37. Oxford: Young Lives.
- OECD. 2003. PISA problem solving items and scoring guides. Paris: OECD Publishing.
- OECD. 2004. Problem solving for tomorrow's world: first measures of cross-curricular competencies from PISA 2003. Paris: OECD Publishing.
- OECD. 2015. Skills for social progress: the power of social and emotional skills. OECD Skills Studies. Paris: OECD Publishing.
- OFQUAL. 2011. Functional Skills Criteria for English. Entry 1, Entry 2, Entry 3, Level 1 and Level 2. Coventry: OFQUAL.
- Reynolds, D., C. Teddlie, B. Creemers, J. Sheerens & T. Townsend. 2000. "An introduction to school effectiveness research". In *The International Handbook of School Effectiveness Research*, edited by C. Teddlie & D. Reynolds, 3-25. London: Falmer Press.
- Reynolds, D., P. Sammons, B. De Fraine, T. Townsend, J. Van Damme. 2011. "Educational Effectiveness Research (EER): A State of the Art Review." Paper presented to the Annual Meeting of the International Congress for School Effectiveness and Improvement, Cyprus, 2011.
- Rolleston, C. 2013. "Who benefits from value-added? School effectiveness in Vietnam". Paper presented at 12th UKFIET International Conference on Education and Development, University of Oxford, 10th September 2013.
- Rolleston, C. 2016 Escaping a low-level equilibrium of educational quality. RISE Working Paper 16/008.
- Rolleston, C., Z. James, L. Pasquier-Doumer & T. N. Thi Minh Tam. 2013. Making progress: report of the Young Lives School Survey in Vietnam. Working Paper 100. Oxford: Young Lives.
- Rossiter, J. 2016. The design of the 2016-17 Young Lives School Survey in Ethiopia. Technical Note 36. Oxford: Young Lives.
- Scheerens, J., H. Lutyen & J. van Ravens. 2011. "Measuring educational quality by means of indicators". In *Perspectives on Educational Quality: Illustrative outcomes on primary and secondary schooling in the Netherlands*, edited by J. Scheerens, H. Luyten & J. van Ravens, 35 - 50. Dordrecht: Springer Netherlands.

- Schendel, R. & A. Tolmie. 2016. "Beyond translation: adapting a performance-task-based assessment of critical thinking ability for use in Rwanda". *Assessment & Evaluation in Higher Education*, DOI: 10.1080/02602938.2016.1177484.
- Thomas, K. & B. Lok. 2015. "Teaching critical thinking: an operational framework". In *The Palgrave Handbook of Critical Thinking in Higher Education*, edited by M. Davies & R. Barnett. New York, NY: Palgrave MacMillan.
- Thomas, S.M., L. Kyriakides & T. Townsend 2016. "Educational Effectiveness Research in New and Emerging Contexts". In *The International Handbook of Educational Effectiveness: Research, Policy and Practice*, edited by C. Chapman, D. Muijs, D. Reynolds, P. Sammons & C. Teddlie, 220-245. London: Routledge.
- World Bank. 2009. *Secondary Education in India: Universalising opportunity*. Washington DC: Human Development Unit, South Asia Region.
- World Bank. 2014. *Skilling up Vietnam: Preparing the workforce for a modern market economy*. Main report. Hanoi: Vietnam Development Information Centre.
- World Bank. 2015. *Project Appraisal Document for a Renovation of General Education Project, Vietnam*. Washington, D.C.: World Bank