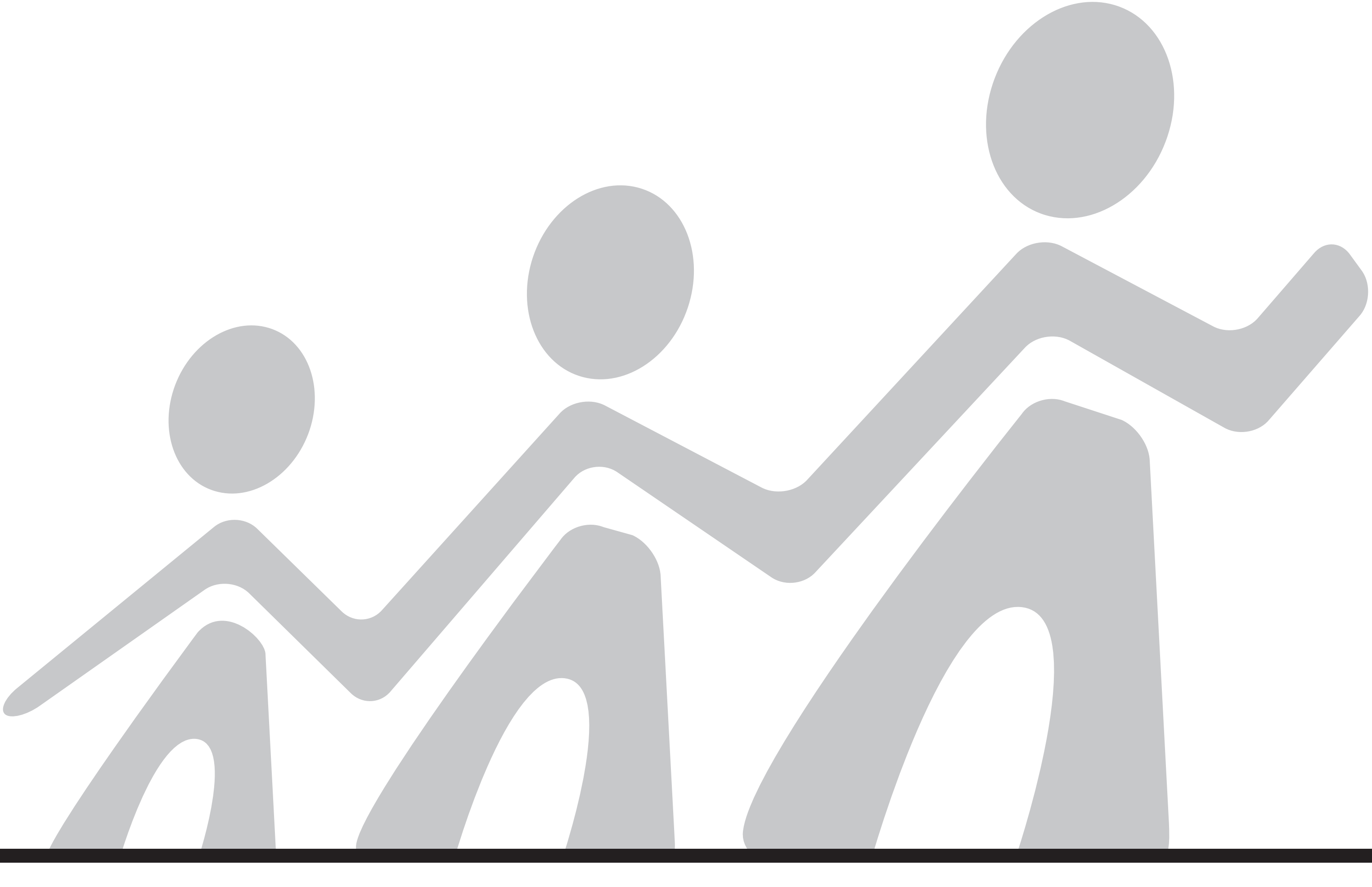
Technical Note 50

Young Lives

Using Scale-Anchoring to Interpret the Young Lives 2016–17 Achievement Scale

Zoe James and Jack Rossiter



www.younglives.org.uk

DECEMBER 2018

Using Scale-Anchoring to Interpret the Young Lives 2016–17 Achievement Scale

Zoe James and Jack Rossiter

First published by Young Lives in December 2018

Printed on FSC-certified paper from traceable and sustainable sources.

About Young Lives

Young Lives is an international study of childhood poverty, following the lives of 12,000 children in four countries (Ethiopia, India, Peru and Vietnam) over 15 years. **www.younglives.org.uk**

Young Lives is core-funded by UK aid from the Department for International Development (DFID).

The views expressed are those of the author(s). They are not necessarily those of, or endorsed by, Young Lives, the University of Oxford, DFID or other funders.

Core-funded by

UKaid
from the British people

Young Lives, Oxford Department of International Development (ODID), University of Oxford,
Queen Elizabeth House, 3 Mansfield Road, Oxford OX1 3TB, UK

Tel: +44 (0)1865 281751 • Email: younglives@younglives.org.uk

Contents

The authors

Zoe James was a Researcher in the Young Lives education team from 2010 to March 2016. She continues to be closely involved in the design and development of the Young Lives school surveys, and has broad interests in education-related issues in low and middle-income countries. Zoe recently completed her PhD at UCL Institute of Education, where her doctoral research examined the relationship between language of instruction and pupil learning outcomes in Ethiopia, including a focus on issues of cross-language and cross-cultural educational measurement. Her previous research has included work on issues of school effectiveness in Ethiopia and Vietnam, and privatisation of education in India.

Jack Rossiter is an Education Research Officer at Young Lives, working on Ethiopia, and a Research Associate at the Oxford Department of International Development. His research interests include early learning, the economics of education, cognitive skills development and educational assessment. Jack is involved in the delivery of the Young Lives school effectiveness survey in Ethiopia and policy-led research support to the Government of Ethiopia, to improve the design and implementation of pre-primary education.

Acknowledgements

This work has benefited from guidance from Bridget Azubuike, Padmini Iyer, Rhiannon Moore and Caine Rolleston.

1. Introduction

An important dimension of the Young Lives school surveys in Ethiopia, India, Peru and Vietnam has been the inclusion of assessments in selected cognitive domains. In the 2016-17 secondary school survey (Iyer and Moore 2017), assessments of mathematics and English were administered at the beginning and end of the school year in Ethiopia, India and Vietnam (Azubuike et al. 2017).[1]

Assessments included a set of 'common items' which allowed linkages to be made across countries and across survey 'waves' (Wave 1, 'W1', at the beginning of each country's academic year, and Wave 2, 'W2', at the end of the academic year). Results were summarised using techniques from item response theory (IRT), with two waves of data from each country and for each subject being analysed concurrently (Edelen and Reeve 2007). This approach permits the creation of a single interval scale, against which the performance of each student can be reported. In the case of the 2016-17 school surveys, each scale takes a mean of 500 and standard deviation of 100 at W1 (Moore et al. 2017).

A single scale offers a measure against which to compare performance of students – both over the course of the school year, and in relation to each other. However, a 'norm-referenced' achievement scale of this type, on which the performances of individual students are described in terms of how they score in relation to other students, offers little information on the competencies associated with different locations on the scale (Beaton and Allen 1992). Such competency information can help to bring to life an otherwise abstract numerical scale and is potentially valuable for policy and curricular reform.

This technical note presents the results of an exploratory 'scale-anchoring' exercise, which links items to achievement levels to produce performance-level descriptors of what students have demonstrated they 'know and can do' (Sinharay et al. 2011). According to this exercise, an achievement score distribution is reviewed, and performance levels identified along the scale. The types of items that students at each 'level' are typically able to answer correctly are then examined, with a view to generating a 'competency statement' for each level of performance. The note uses mathematics assessment data from the Young Lives 2016-17 school survey in India before extending the analysis to include Ethiopia, India and Vietnam in an exploratory cross-country scale.[2]

2. Methodological approach

Constructing scales is a well-developed function of educational measurement but communicating the meanings of scales to policymakers or to the public is not always effective (Beaton and Allen 1992). Scale-anchoring for a test involves a substantial amount of work – and subjective judgment – to summarise the relationships between tasks that

[1] Young Lives countries include Ethiopia, India (the states of Andhra Pradesh and Telangana), Peru and Vietnam. The 2016-17 round of cross-country school surveys did not include Peru. Instead, in 2017 a standalone school survey was conducted in Peru, with a single measure of student achievement, which cannot be linked to assessments in the other three countries. As a result, Peru is not included on scales presented in this technical note.

[2] As some school types were census sampled, and others were randomly sampled, sampling weights are required for the data to be representative of pupils in different school types at the Young Lives site level in India. For more information on sampling weights applied in India see Moore et al. 2017.

students can perform and their observed test scores (Sinharay et al. 2011). Large-scale assessments such as the National Assessment of Educational Progress (NAEP), the Trends in International Mathematics and Science Study (TIMSS), and the Programme for International Student Assessment (PISA) use scale-anchoring methods to aid interpretation of what students at selected points on each scale 'know and can do'. Although different in their approaches, we have drawn on each to identify the four steps that have been followed in this technical note:

1. select the score that represents student performance in mathematics;
2. identify performance thresholds and reference groups of students;
3. identify item-allocation criteria; and
4. generate competency statements.

2.1. Select the score that represents student performance in mathematics

The first step in the scale-anchoring exercise was to select the score that represents student performance in mathematics. In cross-sectional assessments common to NEAP, TIMSS, PISA and others, students sit one assessment with a common set of items. As a result, there is only one score per student, which serves as the indicator of student performance. Each student in the Young Lives 2016-17 school survey, however, had two distinct mathematics scores available for consideration: their W1 score and their W2 score.[3]

The use of concurrent calibration to scale the data suggests that information from every item, whether administered at W1 or W2, or both, informs the final student scores for each survey wave. With this in mind, the end-of-year W2 score was selected since, conceptually, this offers a better indicator of what students 'can do' in Grade 9 in survey schools. By contrast, W1 was conducted at the start of the school year so is perhaps more indicative of what students could do at the end of Grade 8.

This selection does not, however, restrict the anchoring exercise to the pool of W2 items. All items from W1 and W2 will be associated with student performance at W2. Put another way, the performance of students at each level of W2 score will be examined on all W1 and W2 items. The interpretation of the different competency statements generated by this exercise is therefore '*what students at different levels of achievement at the end of Grade 9 can typically do over the course of the Grade 9 school year*'.

2.2. Identify performance thresholds and reference groups of students

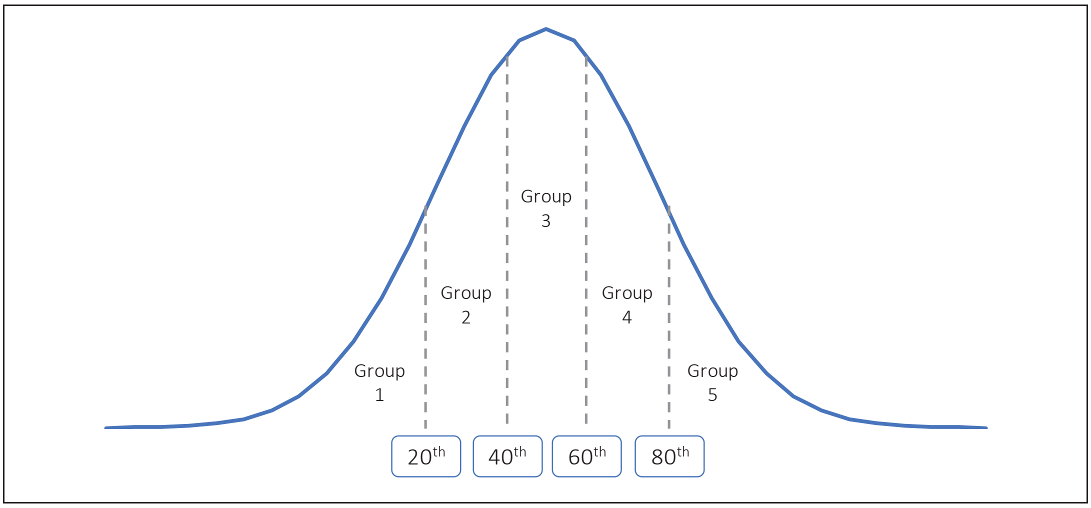
The second step in the process was the identification of thresholds in the performance data (i.e. along the scale), which split students into groups. Performance on items can then be examined in relation to these thresholds. Two approaches were identified.

The first approach splits the entire student sample into groups, according to some pre-specified cut-points in the performance distribution. We refer to this as the 'groups' approach. For example, a 'data-driven' approach might split the entire student sample into five groups,

[3] Some students present at W1 were not present at W2. These students remain in the sample for other analyses, but for all scale-anchoring exercises, so that full item response information is available, only students who completed assessments at W1 and W2 are retained.

representing those below the 20th percentile of W2 achievement, those between the 20th and 40th percentiles of W2 achievement, and so on (Figure 1). Alternatively, a review of the performance distribution might suggest that splitting the student sample into five groups based on four easily communicated but somewhat arbitrary points in the performance distribution – for example, 300, 400, 500 and 600 – may yield a similar result. Average item-level performance of students in each of these five groups could then be examined.

Figure 1. *Example reference groups if student sample is split according to percentile thresholds*

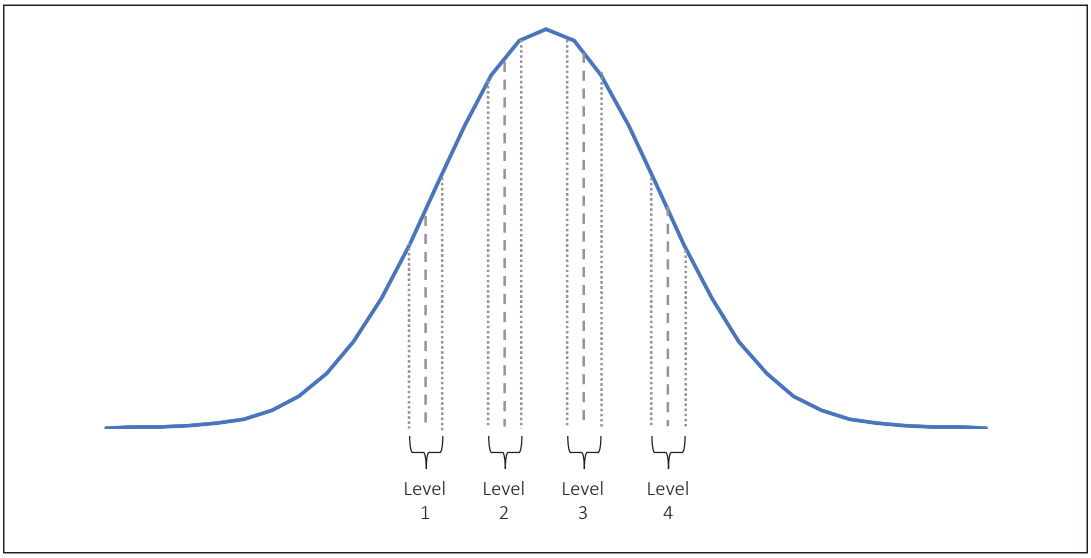


While having the advantage of using the whole sample, this approach inevitably means that a wide distribution of students, in terms of their mathematics performance, is grouped together. For example, students at the 19th percentile are grouped with students at the 5th percentile, but they may be expected to have more in common with those at the 21st percentile, who fall into a different performance level. This heterogeneity of skills within levels can make subsequent stages of the process more challenging, insofar as the exercise is trying to identify distinct competencies between groups of students at different levels. As such, where previous studies have adopted this approach, variation in response patterns 'within group' is often examined as part of the exercise (for example, PISA uses a group-oriented approach, see OECD 2014).

A second approach uses researcher-selected 'benchmarks' around which a 'bandwidth' is established (Figure 2). We refer to this as a 'levels' approach. As with the previous approach, benchmarks can be selected from the data (e.g. the 20th, 40th, 60th and 80th percentiles), or can be selected to reflect easily communicable (but arbitrary) scores, such as 300, 400, 500 and 600. A bandwidth such as +/-10 points around the benchmark is defined and the sub-sample of students falling within that range of performance represents that level. The bandwidth can then be adjusted, for example between +/-10 points or +/-20 points, to find an appropriate balance between the size of the sub-sample falling within each bandwidth and the distinctiveness of competencies that are represented at each level.[4]

[4] Note that TIMMS and PIRLS use +/-5 points and are still able to have large sample sizes at each level owing to their large overall sample (Mullis 2012).

Figure 2. *Example reference levels*

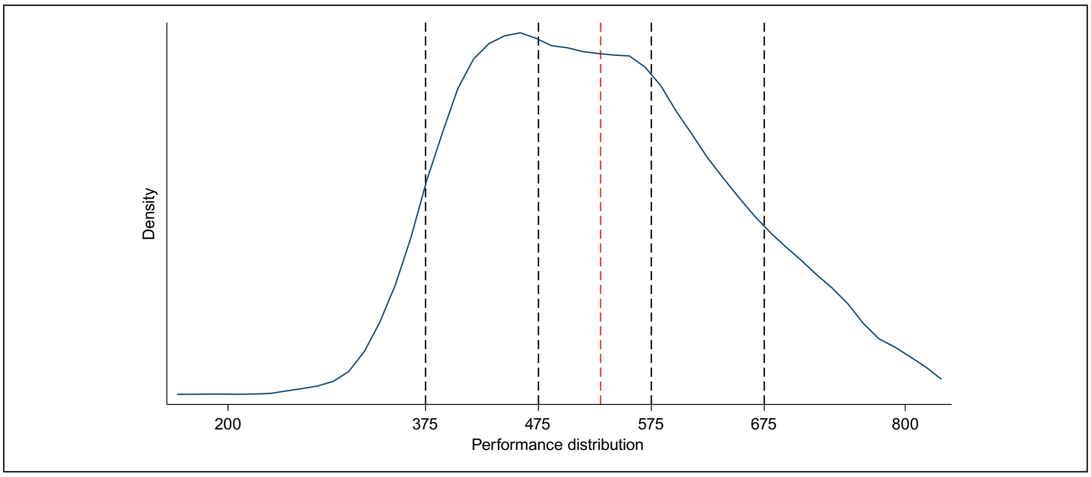


The 'levels' approach offers certain advantages, particularly thanks to its focus on a sub-sample of students with a narrow range of performance, making each level more easily distinguishable from the next. However, narrow bandwidths can lead to an over-reliance on data from a small sub-sample of students. An approach which approximates this is adopted in TIMSS and PIRLS (Mullis 2012).

A 'levels' approach was favoured over a 'groups' approach, since it offers opportunities for greater precision and the identification of distinct competency levels. The selection of benchmarks and bandwidths was made by balancing: (i) the need for sufficient students at each level so that they can be considered to represent the sample (i.e. avoid bandwidths that are too narrow); (ii) the need for levels that represent appropriately homogenous student performance (i.e. avoid bandwidths that are too wide); (iii) the need for levels that are distinct from one another in terms of the competencies they represent (i.e. not too many levels nor levels that are too close to one another); and (iv) the need for sufficient items to anchor at each level (based on work with item-level data and linked to the process outlined in Section 2.3).

Based on these considerations, four levels were defined at 375, 475, 575 and 675, along with a common bandwidth of +/-20 points. Figure 3 presents the distribution of W2 mathematics performance for students in the India sample, with the mean of the W2 score (approximately 530 points) depicted by the red dotted line and the four levels shown with black dotted lines.

Figure 3. *Wave 2 mathematics score and proficiency levels at 375, 475, 575 and 675*



Source: Young Lives 2016-17 school survey, data for India.

The number of students that fall into each of these levels is illustrated in Table 1, together with the percentiles of the achievement distribution to which they relate. The use of a relatively wide bandwidth (+/-20 points) ensures that the sample size at each level does not fall below 300 students (unweighted data).[5] Owing to the distribution of test scores, there are fewer students in the 375 and 675 levels than at the 475 and 575 equivalents.

Just under 40 per cent of the student sample falls into one of the four levels and all subsequent statements about proficiency are based on data for this sub-sample. Accordingly, the results provide a snapshot of what students at different levels of performance have demonstrated that they can do during Grade 9.

Table 1. *Description of sub-sample at each performance level*

Level	'375'	'475'	'575'	'675'
Approximate percentile	6th	36th	67th	88th
Performance range	355 – 395	455 – 495	555 – 595	655 – 695
Number of students (weighted)	568	994	935	486
Number of students (unweighted)	851	1149	678	305

2.3. Identify item-allocation criteria

Having selected levels, it is then necessary to compute the percentage of students at each level who answered each of the items correctly. All items were multiple choice and each assessment, at W1 and at W2, contained 40 items. In sum, there were 51 unique items across both tests. Due to poor item function, four items were discarded from the scale creation process and so were not considered in the scale-anchoring procedure. Of the remaining 47 items, 29 were common to both W1 and W2 and 18 items appeared in only one wave. For each of these items, the percentage of students at each level who answered each item correctly was computed.

[5] Where unweighted data are used, for effective sample size comparison, this is identified. Otherwise, all data for scale-anchoring analysis are weighted.

A simple rule (see Table 2) was applied to allocate items to levels, drawing on the approach of TIMMS and PIRLS (Mullis 2012). An item 'anchors' to a level if 65 per cent of students at that level answered the item correctly *and*, to account for discrimination between levels, fewer than 50 per cent of students at the next lowest level answered the item correctly (the latter criterion being redundant where an item had 65 per cent of students in the lowest level answering it correctly).

However, as in TIMMS and PIRLS, to allow as many items as possible to be included in the anchoring exercise, this rule was relaxed in two stages. In the first instance, items that 'almost anchored' were included, where at least 55 per cent of students answered the item correctly, *and* fewer than 50 per cent of students at the next lowest level answered the item correctly. Finally, as in TIMMS and PIRLS, items that 'weakly anchored', that is, that met only the 55 per cent correct criterion, were also identified. Any item for which fewer than 55 per cent of students at the '675' level answered correctly did not anchor.

Table 2. *Anchoring rules applied in three steps*

Step	Anchor strength	Rule at each level
1	Anchored	≥ 65 per cent of students at the level answer the item correctly; *and* < 50 per cent of students at the next lowest level answer the item correctly.
2	Almost anchored	≥ 55 per cent of students at the level answer the item correctly; *and* < 50 per cent of students at the next lowest level answer the item correctly.
3	Weakly anchored	≥ 55 per cent of students at the level answer the item correctly.

For common items, the percentage of students at each level who answered each item correctly was computed for each wave and results compared. If a common item anchored at different levels, or with different strength across waves, a final set of rules was applied (Table 3).

Recalling that the purpose of the scale-anchoring exercise is to ascertain what students at different parts of the W2 achievement distribution can do during the academic year, if an item anchored in one wave but not another, it was allocated to its successful anchor level. Where items anchored to different levels in different waves, the wave with the strongest anchor was selected. When an item anchored to different levels in different waves, but at the same strength (i.e. 'anchored,' 'almost anchored' or 'weakly anchored'), the W2 stage was used.

Table 3. *Anchoring rules for common items*

Step	Situation	Rule
1	Item anchors at same level in each wave	Item anchors at that level
2	Item anchors at one wave but not at the other wave	Item anchors to successful anchor level
3	Item anchors to different levels in different waves, and with different strength	Item anchors to level with the strongest anchor
4	Item anchors to different levels in different waves, but with same strength	Item anchors to W2 level

Using levels at 375, 475, 575, 675 and a bandwidth of +/-20 points, the 47 items were allocated as shown in Table 4. Of the 47 items, 36 were successfully allocated, with the majority 'anchoring' or 'almost anchoring'. The 11 items that failed to anchor are all items that were administered in only one wave, with seven being administered only in W1, and three only in W2. In almost all cases, they are very difficult items, with very few students getting them correct.

Table 4. *Item allocation to levels, according to anchor strength*

Level	'375'	'475'	'575'	'675'	Total
Anchored	3	6	4	5	18
Almost anchored	0	3	10	2	15
Weakly anchored	0	0	2	1	3
Total	3	9	16	8	36
Failed to anchor	11				

2.4. Generate competency statements

Since the assessments are not comprehensive curriculum-related assessments of mathematics, and the sample is not regionally or national representative, these levels must be interpreted as test- and sample-specific. The four levels therefore indicate what students who are 'Low' (375), 'Intermediate Low' (475), 'Intermediate High' (575), or 'High' (675) achievers, on this assessment and in this sample, are typically able to do.

Students at the 'Low' level sit well below mean performance in this sample, while those at the 'Intermediate Low' level sit just below mean performance. Students at the 'Intermediate High' level are above-average performers, while those at the 'High' level are approaching the top of the achievement distribution. Terms such as 'basic' or 'proficient' were avoided for the purposes of the competency level descriptors, since these terms involve some sort of judgment about what students at this level should be *expected* to be able to do. Instead, terms such as 'Low' and 'Intermediate' can be interpreted in relation to the broader test-score scale, without making statements about whether performance at each level is above or below some externally determined (and/or validated) expectation.

Each anchored item was reviewed according to content domain, cognitive domain and estimated target grade. From these, generalised competency statements were produced to provide indicators of what students can do at each level. It should be noted that these competency levels draw on a relatively small sample of items and therefore provide only indicative evidence of competence in each stated domain or skill. In addition, where the number of items is large (for example at the 'Intermediate High' level) and the content or cognitive coverage broad, available information can end up being lost, distilled or diluted in constructing a single competency statement. Nonetheless, the results are instructive in unpicking what students in this sample can do.

3. Competency statements

Figure 4 presents summary statements of what students can do at each level. Detailed descriptions and example items for each level follow.

Figure 4. *Summary competency statements for each level, Low to High*

Low

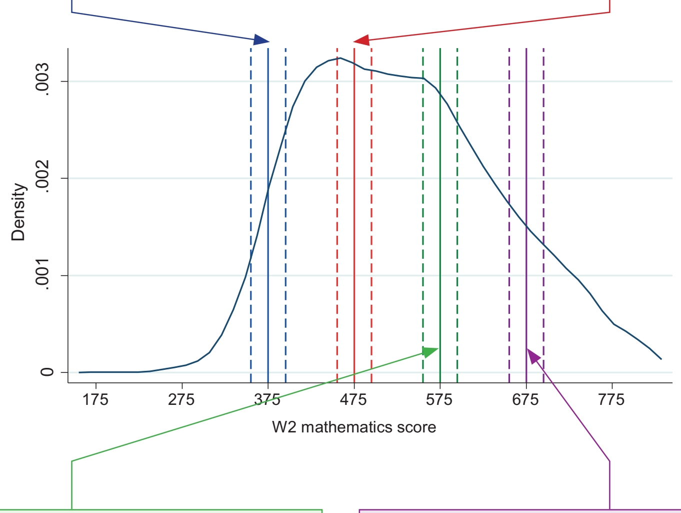
Students can typically answer simple, single-staged mechanical operations presented in a familiar way and involving direct application of simple mathematical functions.

Students can: identify prime numbers; solve double-digit addition; identify congruent triangles.

Intermediate Low

Students can typically answer single and sometimes two-stage mechanical operations presented in a familiar way; some applied problems where relevant information is readily available; can begin to translate word-based problems into mathematical expressions.

Students can: work with 3-digit division; understand place value; use the cube root and exponents; work with shapes and fractions and basic geometry and identify shape-based patterns; extract information from simple graphs.



Intermediate High

Students can typically answer more complex mathematical procedures often involving two stages and presented in both a familiar and unfamiliar way; are able to answer applied questions drawing on a broader array of mathematical competencies; can identify information from multiple sources and convert this to relevant mathematical operations.

Students can: add positive and negative integers; link fractions and decimals; apply the law of exponents; solve complex algebra; demonstrate stronger spatial reasoning; understand volume; extract data from histograms; identify number patterns.

High

Students at this level typically have more developed problem-solving and reasoning capabilities; can answer multi-stage applied problems presented in familiar and unfamiliar contexts; can employ and combine understanding of more complex mechanical operations and functions, and identify and use information from multiple sources to answer mathematical problems.

Students can: subtract two negative numbers; conduct multi-stage arithmetic with decimals; complete more complex algebra, using data from tables; apply the concept of averages to word problems; distinguish fractions in terms of size and equivalence; identify letter patterns.

3.1. Competency level: Low

Number of items: 3

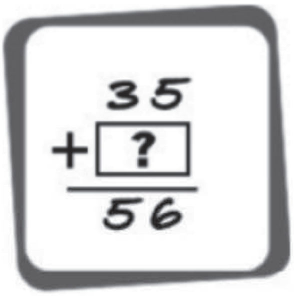
Students at this level can typically answer relatively routine questions presented in a clear and familiar way, usually requiring a single operation, or knowledge of a single mathematical concept, and simple computation. Students are typically able to directly apply relatively basic knowledge and procedures, and thereby demonstrate understanding of the assessed underlying skill.

At this level, students can successfully identify prime numbers, solve a double-digit addition problem, and identify congruent triangles.

On average, items at this level correspond to Grade 6 competencies.

Example item (a)

A number added to 35 gives 56.



What is the number?

A. 12

B. 20

C. 21

D. 91

Example item (b)

Which of the following is a prime number?

A. 5

B. 15

C. 25

D. 35

3.2. Competency level: Intermediate Low

Number of items: 9

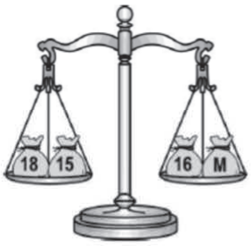
Students at this level begin to be able to answer two-stage problems, or problems requiring understanding of more than one mathematical function or concept. Students start to be able to correctly answer questions of a more applied nature where all relevant information is readily available and presented in familiar contexts, and are more able to identify and combine information from multiple sources and translate it into mathematical expressions to reach the correct answer.

At this level students can typically work with 3-digit division, demonstrate understanding of place value, and can answer more complex mathematical functions, such as the cube root, and exponents. Students can typically work with shapes and fractions and can apply mathematical reasoning to simple concepts of geometry as well as being able to discern simple shape-based patterns. Students can also identify and extract information from simple graphs. Students also begin to be able to translate word-based problems into mathematical expressions, and to correctly answer real-world, applied questions.

On average, items at this level correspond to competencies which lie between what is taught in Grades 6 and 7.

Example item (c)

Jim has balanced some bags of marbles. All the marbles are of the same weight. The number of marbles in each bag is written on it.



How many marbles are there in the bag marked M?

A. 13

B. 17

C. 33

D. 49

Example item (d)

7341 is completely divisible by ______

A. 3

B. 5

C. 7

D. 9

3.3. Competency level: Intermediate High

Number of items: 16

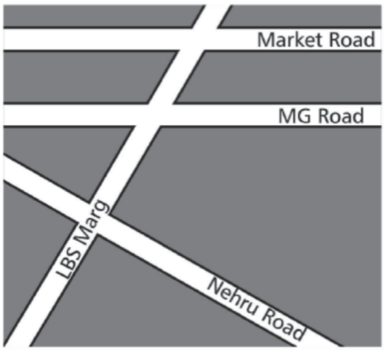
Students at this level can typically conduct more complex mathematical procedures, often arranged as two-stage problems involving understanding of multiple mathematical functions. Students at this level are typically able to answer a larger proportion of applied questions, drawing on their knowledge and understanding of key mathematical functions and procedures. Students can identify information from multiple sources, for example text, pictures, and tables, and convert this to the relevant mathematical operation.

Students are typically able to arrange negative and positive numbers in ascending order, conduct double-digit addition with both positive and negative integers, understand the link between fractions and decimals, and can apply the law of exponents. Students are more likely to be able to answer complex algebraic problems and can typically rearrange information to solve equations. Students at this level demonstrate stronger spatial reasoning, for example identifying shapes presented at different angles and can apply their understanding of elements of geometry (for example, perpendicular lines), to problems expressed in words and pictures. Students can also work with data presented in different forms, can extract information from a histogram, and can identify number patterns.

On average, items at this level correspond to competencies at Grade 7 of the curriculum.

Example item (e)

Look at the map below.



Which of these streets are perpendicular to each other?

A. MR Road and LBS Marg

B. Market Road and MG Road

C. LBS Marg and Nehru Road

D. (None of the streets shown in the map are perpendicular to each other).

3.4. Competency level: High

Number of items: 8

At this level, students can typically answer more complex mathematical problems and demonstrate stronger problem solving and reasoning capabilities. Students typically have good knowledge and understanding of key mathematical functions which they can implement to answer questions framed in a variety of familiar and unfamiliar contexts. Students can answer more complicated application-oriented questions and are typically able to tackle multi-stage problems, including word-based problems.

Students at this level are typically able to complete subtraction problems with two negative numbers and conduct multi-stage arithmetic with decimals. They are also able to complete increasingly complex algebraic problems and can use information in tables to construct algebra problems. Students are typically able to distinguish fractions in terms of size and equivalence, can answer word problems which require understanding of complex underlying mathematical problems, such as averages, and which involve multiple stages of calculation, and are more likely to successfully identify letter patterns.

On average, items at this level correspond to competencies at between Grades 7 and 8 of the curriculum.

Example item (f)

Rahul's father is 6 times as old as Rahul. Rahul's mother is 25 years old. The average age of this family of three is 20 years. How old is Rahul?

A. 15 years

B. 10 years

C. 7 years

D. 5 years

Example item (g)

The table below shows the relationship between the x and y

x	2	3	4	5
y	4	7	10	13

Which of these equations expresses this relation?

A. y = 2x + 2

B. y = x + 2

C. y = 4x – 4

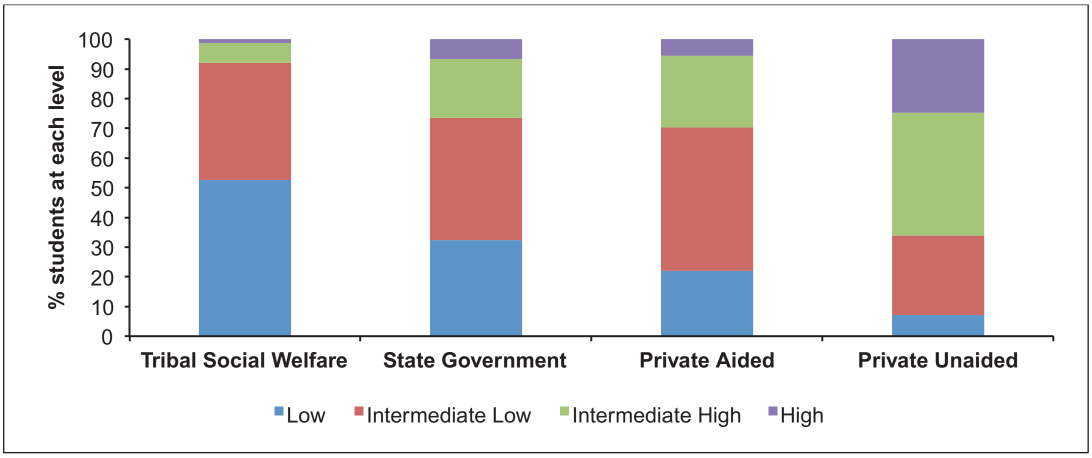
D. y = 3x - 2

4. Competency levels, school and student characteristics

After a scale-anchoring process has added some descriptive meaning to numerical scores, it is possible to examine the distribution of students across competency levels. While it should be borne in mind that only 40 per cent of the student sample informs these breakdowns, they offer an example of the potentially instructive results that can be generated after performance levels are constructed. Results are purely descriptive, and no tests for statistical significance were conducted.

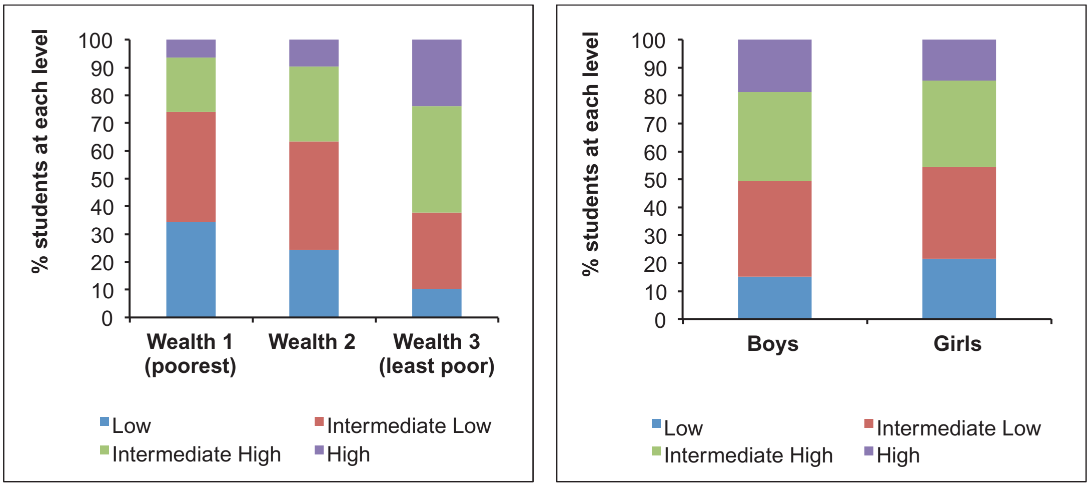
Looking first at performance across school management, just over 50 per cent of students in Tribal Social Welfare schools perform at the 'Low' level and a further 40 per cent perform at the 'Intermediate Low' level (Figure 5). Less than 10 per cent are at the 'Intermediate High' or 'High' levels. By contrast, less than 10 per cent of students in Private Unaided schools perform at the 'Low' level, while over 60 per cent of students in these schools are at 'Intermediate High' or 'High' levels.

Figure 5. *School management type and performance levels*



These variations likely reflect systematic variation in pupil backgrounds between different types of schools. This is drawn out in Figure 6, where it is evident that the poorest third of students (as indicated by a composite index of durable assets), is much more likely to fall into the 'Low' or 'Intermediate Low' levels, than their least-poor counterparts. Less distinct patterns are evident by sex, but slightly more girls fall into the 'Low' and 'Intermediate Low' levels than boys.

Figure 6. *(Left) poverty and performance levels and (Right) sex and performance levels*

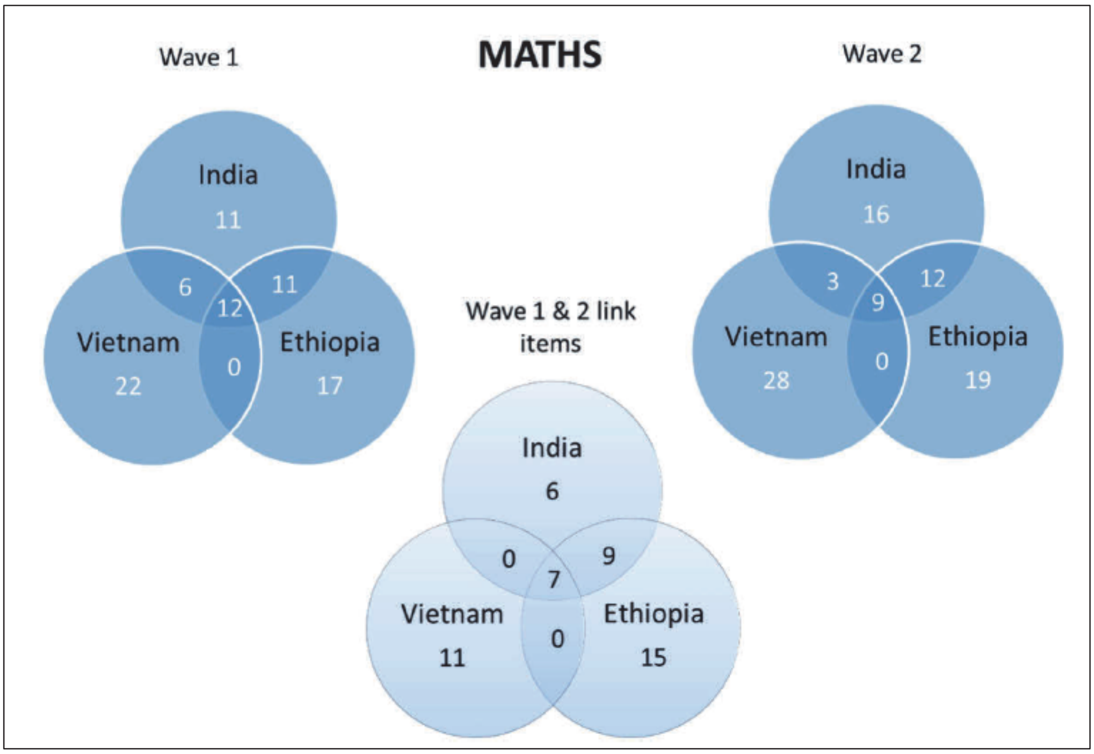


5. Extension: an exploratory cross-country scale-anchoring exercise

An important objective of the Young Lives 2016-17 school surveys was the creation of a single interval scale of mathematics achievement (Azubuike et al. 2017). This would be made possible by the inclusion of 'link items' and would allow the comparison of student achievement across countries and over time (Das and Zajonc 2010). With such a scale, for example, average achievement of 450 scale points in one country has a direct interpretation alongside average achievement of 520 scale points in a second country. It allows country distributions to be modelled together and there is no need to resort to standardisation or subjective judgment of which level of achievement, or achievement progress, is the greater.

In aiming for a single scale, cognitive assessments were developed with 'unique' items and 'link' items. Unique items were specific to the country-wave combination (i.e. items that were administered only in Vietnam at W1), while link items overlapped between countries or/and across rounds (Azubuike et al. 2017). Sometimes, link items spanned two countries at one round (e.g. administered in Ethiopia and India at W2), at other times they linked all countries and rounds (e.g. administered in Ethiopia, India and Vietnam at both W1 and W2). The distribution of unique and link items across mathematics assessments is shown in Figure 7.

Figure 7. *Cross-country and cross-wave link items for mathematics*



Source: Azubuike et al. 2017.

As for the India-only case, a cross-country scale offers a measure against which to compare performance of students and student groups – across countries, over the course of the school year, and in relation to each other. However, it offers little information on the competencies associated with different locations on the scale and how country distributions relate to competency levels.

This section summarises the results of an exploratory scale-anchoring exercise, using the Young Lives secondary school survey cross-country scale of mathematics achievement. The scale involves 111 mathematics items, each administered in some combination of countries (Ethiopia, India, Vietnam) and survey waves (W1 and W2).[6]

The cross-country scale-anchoring exercise followed the steps already set out for the India scale-anchoring process. First, the W2 mathematics score was selected as the performance reference. As a result, the conceptual basis for the competency statements that result is '*what students at different levels of achievement at the end of Grade 7/8/9/10 can typically do over the course of the Grade 7/8/9/10 school year*'. The use of multiple grades may seem a little complicated, but it reflects only the differences in education system structure across countries. In each country, grades were selected to target children aged 14-15 years old. These happen to be in different grades (Grade 7/8 in Ethiopia, 9 in India and 10 in Vietnam).

[6] An 'item' is an individual item administered in the same combination of countries in all survey waves in which it is administered. For example, where the same individual item (e.g. 2+4), was administered in Ethiopia in W1, and in Ethiopia and India in W2, this would be counted as two items. If this same individual item (e.g. 2+4) had been administered in both Ethiopia and India in both W1 and W2, this would be counted as one item.

Note that there should be no expectation of equivalent performance across these countries in these grades.

Second, benchmarks were identified to reflect the range of scores evident both across and within countries, and therefore focused on purposively determined score points, rather than percentiles of performance on the cross-country scale. Table 5 summarises the W2 mathematics score at specific score percentiles for the whole sample, and by country. It was important that the selected benchmarks captured the range of performance in both Ethiopia (from a low of 340 points at the 5th percentile) and Vietnam (to a high of 783 points at the 95th percentile), while also offering insight into variation between students towards the middle of the performance distribution. Five benchmarks were selected, at 350, 450, 550, 650 and 750 points on the test-score scale and labelled L1, L2, L3, L4 and L5.

Table 5. *Wave 2 mathematics scores at country-specific score percentiles (weighted)*[7]

Percentile	Whole sample	Ethiopia	India	Vietnam
5th	361	340	387	462
10th	384	357	407	488
25th	433	389	447	543
50th	507	437	512	612
75th	598	495	582	684
90th	682	557	655	753
95th	730	593	695	783

For each benchmark, a bandwidth of +/-20 score points was used to identify the subsample of students who would be examined during the scale-anchoring process.[8] This offered large sample sizes at each level, while also ensuring each level reflected quite distinct scores. Table 6 summarises the sample size at each level, by country. The country-specific make-up of each level should be noted; students in Ethiopia dominate the lower level(s), while students in Vietnam dominate the higher level(s).

Table 6. *Sample size at each level, by country (weighted and unweighted)*

Level	Sample size							
	Unweighted				Weighted			
	ET	IN	VN	TOT	ET	IN	VN	TOT
L1 (350)	1148	256	6	1410	1141	158	3	1302
L2 (450)	1776	1605	471	3852	1765	1243	349	3357
L3 (550)	735	838	1079	2652	730	1109	975	2814
L4 (650)	118	286	969	1373	117	488	1048	1653
L5 (750)	16	36	409	461	16	87	519	621
Total	3793	3021	2934	9748	3769	3084	1895	9748

[7] As some school types were census sampled, and others were randomly sampled, sampling weights are required for the data to be representative of pupils at the Young Lives site level in India and Vietnam. For more information on sampling weights applied in India see Moore et al. 2017, and in Vietnam see Iyer et al. 2017.

[8] Alternative bandwidths of +/-15 and +/-10 were also explored, with little impact on item allocation. Accordingly, +/-20 was retained to increase the sample size at each level.

Items were then allocated to levels, again drawing on the approach of TIMMS and PIRLS (Mullis 2012), as applied for the India-only case (see Section 2.3). For the cross-country scale, items were anchored regardless of the country in which they were administered. This is conceptually somewhat complex, since this implies that an item may anchor to a level even if very few of the students who comprise that level sat that item. However, if the scale-creation process was reliable, it should support this sort of exercise.

In the cross-country case, for items administered in both W1 and W2 that anchored to different levels in each wave, the strongest anchor was selected. Where such items anchored at the same level of strength at both W1 and W2, the W2 level was selected. Where there were two unique items in terms of the countries in which they were administered in W1 and W2, but where the content was the same in both items, the level to which a multi-country version of that item anchored was usually selected. Table 7 summarises the way in which items anchored to levels, according to the country-combination in which they were administered.

Table 7. *Item allocation (five levels, items by countries of administration)*

Country combination	L1	L2	L3	L4	L5	No anchor	Total
All countries	0	1	5	6	1	0	13
Ethiopia and India	0	4	6	4	0	0	14
India and Vietnam	1	0	1	4	3	0	9
Ethiopia	2	4	10	5	0	0	21
India	0	0	9	6	3	0	18
Vietnam	0	0	6	16	14	0	36
Total	3	9	37	41	21	0	111

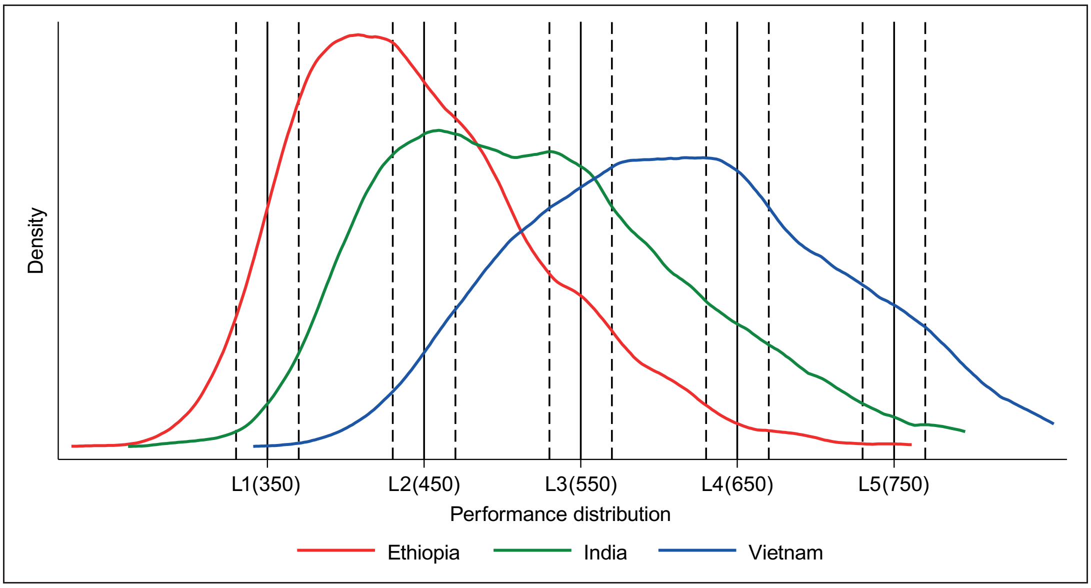
It should be noted that there are only 13 items which are common to all three countries, a further 14 common to both Ethiopia and India, and a further nine common to India and Vietnam. It is these common items on which the creation of the cross-country scale depends. The largest proportion of items anchor at L3 and L4, while far fewer items anchor at the lower end of the performance distribution, likely related in part to the small number of students at this level. The dominance of cross-country items at L3 and L4 reflects the fact that this is the part of the performance distribution in which there is strong 'common support,' and therefore in which it is possible to identify items suitable for students in all three countries.

Having anchored the 111 items to the five identified levels, items were reviewed based on content domain, cognitive domain, similarities and differences in competencies required, to generate synthesised statements about what students at each level were typically able to do. This is a subjective exercise, which can only ever be approximate and is made more challenging by the dominance of items in L3 and L4, and the large number of single-country items (Sinharay et al. 2011). Resulting competency statements should therefore be seen as illustrative of what might be done with this sort of cross-country data, rather than providing an immutable description of what students are able to do in these three countries.

6. Cross-country distribution and allocation of students to levels

Figure 8 presents the distribution of scores in each country on the cross-country scale, together with the specified levels (L1-L5). Table 8 then summarises the country-wise composition of the student sample that falls *between* each level, that is, it answers the question: for every 100 students that sit between L1 and L2, what share are from Ethiopia, India and Vietnam, respectively?

Figure 8. *Wave 2 mathematics achievement distribution by country with levels overlaid (weighted)*



Although there are substantial overlaps between country distributions, there are also important differences in the share of students, from each country, that falls between levels. Below L1, the student sample is overwhelmingly Ethiopian, while between L1–L2, the student sample is just under 70 per cent Ethiopian, and just under 30 per cent Indian, with very few Vietnamese students performing at this level. Between L2–L3, and L3–L4, the student sample is a mix of students from all three countries, while between L4–L5, Vietnamese students begin to dominate, at 70 per cent of the sample and there are very few Ethiopian students (under 5 per cent). Above L5, the student sample is 86 per cent Vietnamese, while only 1 per cent of students are Ethiopian and the remainder Indian.

Table 8. *Country proportions of students between each level*

Country	Country proportions of students between each level (weighted, expressed as each country's share of the students between each level)					
	< L1 (< 349) (%)	L1 – L2 (350–449) (%)	L2 – L3 (450–549) (%)	L3 – L4 (550–649) (%)	L4 – L5 (650–749) (%)	> L5 (>750) (%)
Ethiopia	92	68	39	17	4	1
India	8	28	38	34	26	13
Vietnam	<1?	4	23	49	70	86
Total	100	100	100	100	100	100

Notes: This table represents the country shares of students between each level. For example, for students between L1 and L2, what proportion is Ethiopian, Indian, Vietnamese, respectively. It does not show the distribution of students within-country across levels; that is, it is not that 86 per cent of students in Vietnam scored above L5, rather, of the students that scored above L5, 86 per cent were from Vietnam.

While this breakdown of the distribution helps to unpick the country-specific composition of the student subsamples between levels, it is perhaps misleading insofar as it fails to reflect the within-country distribution of students *across the score range*. Within-country distributions are shown graphically in Figure 8 and Table 9 presents the distributions of students across levels, that is, it answers the question: for the student sample in Vietnam (and for each other country), what percentage falls between L1–L2, L2–L3, L3–L4 and L4–L5, respectively? It should be noted that very few students fall below L1 in any country, while students in each country are often concentrated in particular parts of the general cross-country distribution.

Table 9. *Within-country distribution of students between levels (weighted)*

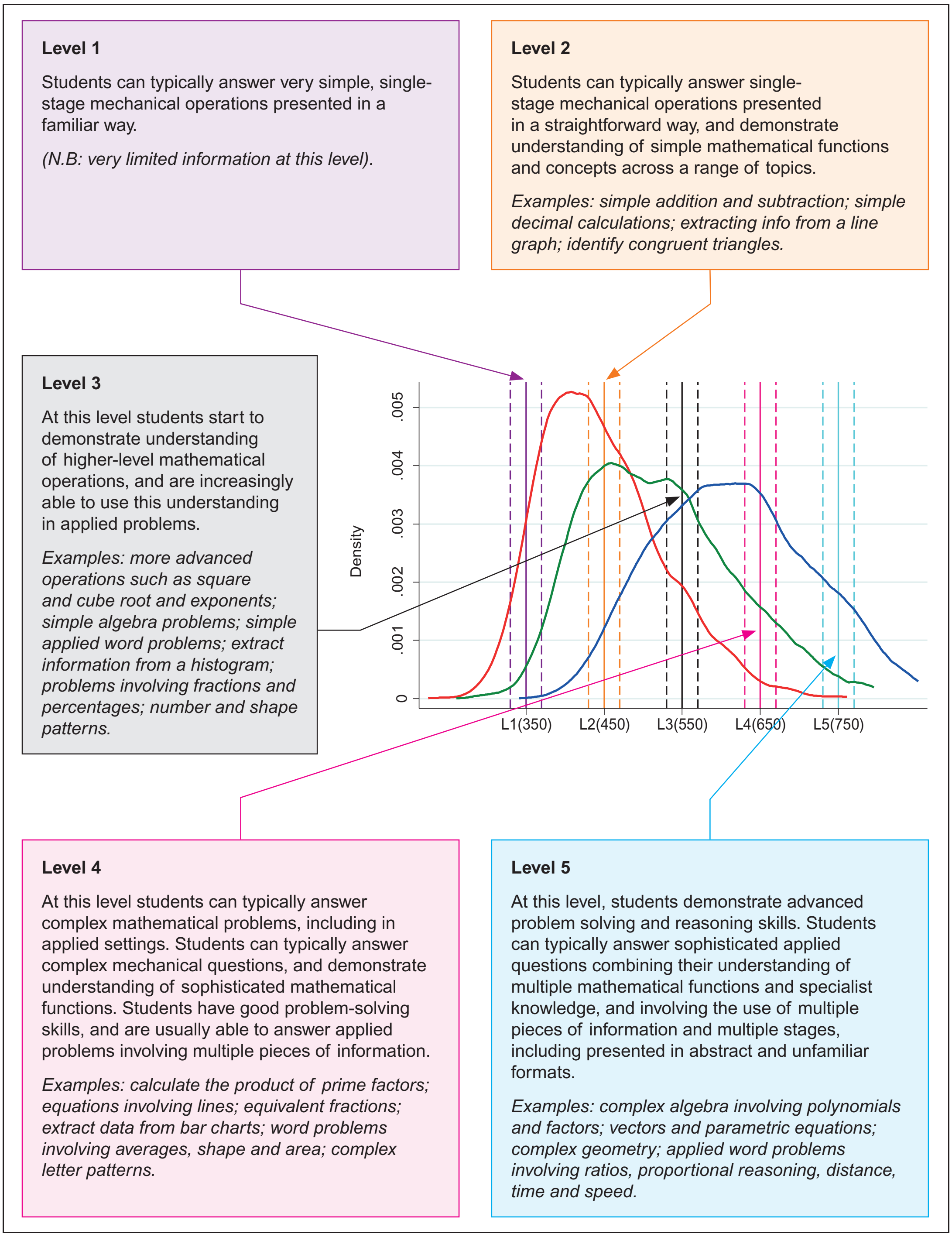
Level	Country		
	Ethiopia (%)	India (%)	Vietnam (%)
< L1 (< 349)	8	1	< 0.5
L1 – L2 (350 – 449)	**49**	**25**	3
L2 – L3 (450 – 549)	**32**	**38**	**24**
L3 – L4 (550 – 649)	10	**25**	**37**
L4 – L5 (650 – 749)	1	9	**25**
> L5 (> 750)	< 1?	2	11
Total	100	100	100

In Ethiopia, nearly half of the student sample falls between L1–L2, while over 30 per cent falls between L2–L3, and the remainder falls either below L1, or between L3–L4. Very few students exceed L4 (below 1 per cent). In India, meanwhile, the largest proportion of students falls between L2–L3 (38 per cent), while a further 25 per cent falls into both L1–L2 and L3–L4. Just under 10 per cent of the student sample falls between L4–L5 and no more than 2 per cent falls below L1, or above L5. In Vietnam, the single largest proportion of students falls between L3–L4 (just under 40 per cent). A further 24 per cent of the student sample falls between L2–L3 and 25 per cent between L4–L5. A final 11 per cent exceeds L5. Less than 4 per cent of the Vietnam sample falls below L2.

7. Cross-country competency statements

Figure 9 presents summary competency statements for each level. Detailed descriptions follow.

Figure 9. *Summary competency statements*



Level 1

Students can typically answer very simple, single-stage mechanical operations presented in a familiar way.

(N.B: very limited information at this level).

Level 2

Students can typically answer single-stage mechanical operations presented in a straightforward way, and demonstrate understanding of simple mathematical functions and concepts across a range of topics.

Examples: simple addition and subtraction; simple decimal calculations; extracting info from a line graph; identify congruent triangles.

Level 3

At this level students start to demonstrate understanding of higher-level mathematical operations, and are increasingly able to use this understanding in applied problems.

Examples: more advanced operations such as square and cube root and exponents; simple algebra problems; simple applied word problems; extract information from a histogram; problems involving fractions and percentages; number and shape patterns.

Level 4

At this level students can typically answer complex mathematical problems, including in applied settings. Students can typically answer complex mechanical questions, and demonstrate understanding of sophisticated mathematical functions. Students have good problem-solving skills, and are usually able to answer applied problems involving multiple pieces of information.

Examples: calculate the product of prime factors; equations involving lines; equivalent fractions; extract data from bar charts; word problems involving averages, shape and area; complex letter patterns.

Level 5

At this level, students demonstrate advanced problem solving and reasoning skills. Students can typically answer sophisticated applied questions combining their understanding of multiple mathematical functions and specialist knowledge, and involving the use of multiple pieces of information and multiple stages, including presented in abstract and unfamiliar formats.

Examples: complex algebra involving polynomials and factors; vectors and parametric equations; complex geometry; applied word problems involving ratios, proportional reasoning, distance, time and speed.

Level 1 (L1, 350)

Number of items: 3

We have very limited information about what students at this level of achievement can do, so this level should be treated with caution.

Students can typically answer very simple, single-stage mechanical operations presented in a familiar way.

At this level students can usually solve a simple subtraction; identify prime numbers; and evaluate relative volumes.

Level 2 (L2, 450)

Number of items: 9

Students can typically answer single-stage mechanical operations presented in a straightforward way and demonstrate understanding of simple mathematical functions and concepts across a range of topics.

At this level students can usually solve simple addition and subtraction problems, including involving decimals and rounding; can identify place value in numbers up to 5 digits; can extract data from a line graph; can solve simple algebra problems; and can identify congruent triangles.

Level 3 (L3, 550)

Number of items: 37

At this level students start to demonstrate understanding of higher-level mathematical operations and are increasingly able to use this understanding in applied problems.

Students can typically answer more demanding single-stage mechanical operations; demonstrate understanding of more advanced mathematical functions such as square and cube root and exponents; can answer questions involving shapes, fractions and simple geometry; can answer more applied questions including simple word problems where all relevant information is available, and can answer questions involving spatial reasoning.

Students are typically able to successfully answer simple division questions; conduct simple calculations involving decimals and positive and negative integers; calculate the cube and square root of a number; appraise the relative order of 7-digit numbers; simplify fractions; solve simple algebra problems; and convert fractions to decimals. Students can demonstrate understanding of and use exponents; work with number lines; extract information from a histogram; appraise alternative presentations of data; compare volume of 3d shapes; and work with fractions and percentages, including in relation to shapes.

Students at this level are also able to correctly answer questions framed in an applied way, including word problems where all relevant information is readily available; work with number and shape patterns; and can solve simple problems involving fractions and ratios.

Level 4 (L4, 650)

Number of items: 41

At this level students can typically answer complex mathematical problems, including in applied settings.

Students can typically answer complex mechanical questions and demonstrate understanding of sophisticated mathematical functions. Students have good problem-solving skills and are usually able to answer applied problems involving multiple pieces of information.

Students at this level are typically able to answer mechanical questions involving subtraction of two single-digit negative integers; arrange positive and negative integers in ascending order; convert decimals and percentages; and can calculate the product of prime factors. Students also demonstrate a good understanding of complex algebra and can typically answer questions including: equations of intersecting lines; identifying the value of a variable based on a given inequality; questions involving roots; and questions involving parabolas and vectors.

Students also demonstrate understanding of equivalent fractions; are proficient in more sophisticated geometric operations including the calculation of the area of both a rectangle and a trapezium; and can identify shapes from information about angles and sides; students can also work with data and are able to extract multiple pieces of information from a bar chart to solve a problem.

Students at this level also demonstrate good problem-solving skills, and they can answer applied problems which involve multiple stages and different topics, including problems involving averages, shapes and area. Students can also solve problems involving complex letter patterns.

Level 5 (L5, 750)

Number of items: 21

At this level, students demonstrate advanced problem solving and reasoning skills.

Students can typically answer sophisticated applied questions combining their understanding of multiple mathematical functions and involving the use of multiple pieces of information and multiple stages, including presented in abstract and unfamiliar formats.

Students can typically successfully answer complicated mechanical operations involving specialist knowledge, including complex algebra involving polynomials and factors; can identify equivalent numbers with exponents; vectors and parametric equations; and demonstrate understanding of the conditions under which equations have roots. Students have a strong understanding of complex geometry and can calculate angles inscribed within a circle. Students can also interpret data from a pie chart.

Students can also successfully answer applied word problems involving ratios; the perimeter of a quadrilateral; weight; percentages; proportional reasoning; distance, time and speed, and can use their understanding of multiple functions to solve complex problems.

8. Summary

This technical note has summarised two exploratory scale-anchoring exercises conducted with data from mathematics assessments administered in the Young Lives 2016-17 school surveys of Ethiopia, India and Vietnam.

It first introduced a limitation of ‘norm-referenced’ achievement scales, namely that students may be compared to one another, but that their achievement scores bear no relation to expected or desirable skills or competencies.

There are several ways to tackle this research problem and this note proposed a ‘scale-anchoring’ exercise which establishes a set of distinct performance levels, to which items are ‘anchored’. These levels may then be described in terms of the mathematical skills and competencies that they represent, bringing to life an otherwise abstract numerical scale, with potential value for policy dialogue and curricular reform.

The exercise draws on the approach adopted in TIMMS and PIRLS (Mullis 2012), focusing on specific benchmarks in the data and the performance of the sub-sample of students that fall within a specified bandwidth of each benchmark. The note outlined four steps taken to: (i) select scores that represent distinct levels of student performance in mathematics; (ii) identify performance thresholds and reference groups of students; (iii) identify item-allocation criteria; and (iv) generate competency statements. It then presented achievement distributions and competency statements for the India-only case and for an exploratory cross-country case.

This approach returned four levels of achievement for the India-only sample and five levels for the cross-country sample. As levels increase, in each case, so does the mathematical skill that they reflect. Clear – and logical – distinctions are evident between the types of items that students at each level are typically able to answer correctly, while variation in the characteristics of students who fall between levels is also evident.

Throughout the note, challenges and limitations have been discussed. Two methodological challenges are relevant to both the India-only and cross-country exercises: (i) the ‘levels’ approach uses only a sub-sample of the data, thereby failing to capitalise on all information available from student responses; and (ii) the process of developing competency statements concentrates a huge amount of item-information into a single, hopefully concise, statement of proficiency. Both limitations hinge on qualitative judgments: (i) how wide should the bandwidths be and so how large is the retained student sample; and (ii) how should item information be combined into a valid statement of proficiency. The former can be tested and decided upon within the process and its impact on how/where items anchor can contribute to robustness claims, or may highlight concerns with the reliability of the approach. The latter, to be done well, requires extensive item information, curriculum information and awareness of specific and sometimes complex mathematics skills and procedures.

The cross-country exercise faces further challenges, since it relies heavily on the validity of a single cross-country scale, which is generated from non-equivalent forms with anchor items across countries. This is both conceptually and methodologically complex. In curriculum-linked assessments, the number of cross-country anchor items may, by necessity, be small. There may also be statistical limitations, based on linking errors for example, to the creation of a single-scale across three populations with quite different curricula and non-common sample sizes or sampling methodologies. The conceptual validity of such an exercise is also challenging. Testing the statistical or conceptual limitations of a cross-country scale was not the purpose of this note, but the development of hybrid assessments of this type is an area of research interest (for example, see Wagner 2011).

Thereafter, if competency statements are going to have policy relevance (and at the very least ‘face validity’), then it is not obvious whether to allow items to anchor regardless of the country in which they were administered. This is conceptually somewhat complex, since this implies that an item may anchor to a level even if very few of the students who comprise that level sat that item. Relatedly, competency statements for lowest and highest levels are based, in large part, on responses from students in one of the three countries (Ethiopia at the lower end and Vietnam at the higher end). The lack of common support at the extremes means that students in Ethiopia (Vietnam) with highest (lowest) performance are then described in terms of items that they would never have sat, neither items that would have much relevance to Ethiopia’s (Vietnam’s) curriculum. The validity of such an approach relies in turn on the validity of the scale itself, and it is important that such complexities and challenges be borne in mind in any attempts at equivalent exercises, or in using such data.

However, while acknowledging and investigating some of these limitations in this note, we feel that this exploratory approach to scale-anchoring offers an alternative way of evaluating and presenting student assessment data. If done well it can contribute to policy dialogue and awareness around the shortage or abundance of relevant skills which, ultimately, might contribute to progress on the acquisition of such skills.

References

Azubuike, O.B., R. Moore, and P. Iyer (2017) *Young Lives School Surveys, 2016–17: The Design and Development of Cross-Country Maths and English Tests in Ethiopia, India and Vietnam*, Technical Note 39, Oxford: Young Lives.

Das, J., and T. Zajonc (2010) ‘India Shining and Bharat Drowning: Comparing Two Indian States to the Worldwide Distribution in Mathematics Achievement’, *Journal of Development Economics* 92: 175–187.

Edelen, M., and B. Reeve (2007) ‘Applying Item Response Theory (IRT) Modelling to Questionnaire Development, Evaluation, and Refinement’, Quality of Life Research 16.1: 5-18.

Iyer, P., O.B. Azubuike, and C. Rolleston (2017) ‘Young Lives School Survey 2016-17: Evidence from Vietnam’, Country Report, Oxford: Young Lives.

Iyer, P., and R. Moore (2017) ‘Measuring Learning Quality in Ethiopia, India and Vietnam: From Primary to Secondary School Effectiveness, *Compare: A Journal of Comparative and International Education* 47.6: 908-924.

Moore, R., O.B. Azubuike, P. Reddy, C. Rolleston, and R. Singh (2017) ‘Young Lives School Survey 2016-17: Evidence from India’, Country Report, Oxford: Young Lives.

Mullis, I. (2012) ‘Using Scale Anchoring to Interpret the TIMSS and PIRLS 2011 Achievement Scales’, in M.O. Martin and I.V.S. Mullis (eds.) *Methods and Procedures in TIMSS and PIRLS 2011*, Chestnut Hill, MA: IEA.

OECD (2014) ‘PISA 2012 Technical Report’, Paris: OECD.

Wagner, D.A. (2011) ‘Smaller, Quicker, Cheaper: Improving Learning Assessments for Developing Countries’, Paris: UNESCO-IIEP.

Using Scale-Anchoring to Interpret the Young Lives 2016–17 Achievement Scale

An important dimension of the Young Lives school surveys in Ethiopia, India, Peru and Vietnam has been the inclusion of assessments in selected cognitive domains. In the 2016-17 secondary school survey, assessments of mathematics and English were administered at the beginning and end of the school year in Ethiopia, India and Vietnam.

This technical note presents the results of two exploratory `scale-anchoring' exercises, which link items to achievement levels to produce performance-level descriptors of what students have demonstrated they know and can do. The note uses mathematics assessment data from the 2016-17 school survey in India before extending the analysis to include Ethiopia, India and Vietnam in a cross-country scale.

Young Lives
An International Study of Childhood Poverty

About Young Lives

Young Lives is an international study of childhood poverty, involving 12,000 children in four countries over 15 years. It is led by a team in the Department of International Development at the University of Oxford in association with research and policy partners in the four study countries: Ethiopia, India, Peru and Vietnam.

Through researching different aspects of children's lives, we seek to improve policies and programmes for children.

Young Lives Partners

Young Lives is coordinated by a small team based at the University of Oxford, led by Professor Jo Boyden.

- *Ethiopian Development Research Institute, Ethiopia*
- *Pankhurst Development Research and Consulting plc, Ethiopia*
- *Centre for Economic and Social Studies, Hyderabad, India*
- *Sri Padmavathi Mahila Visvavidyalayam (Women's University), Andhra Pradesh, India*
- *Grupo de Análisis para el Desarollo (GRADE), Peru*
- *Instituto de Investigación Nutricional (IIN), Peru*
- *Centre for Analysis and Forecasting, Vietnamese Academy of Social Sciences, Vietnam*
- *General Statistics Office, Vietnam*
- *Oxford Department of International Development, University of Oxford, UK*

Contact:
Young Lives
Oxford Department of
International Development,
University of Oxford,
Mansfield Road,
Oxford OX1 3TB, UK
Tel: +44 (0)1865 281751
Email: younglives@younglives.org.uk
Website: www.younglives.org.uk

Young Lives