Technical Note 39

Young Lives

Young Lives School Surveys, 2016–17

The Design and Development of Cross-Country Maths and English Tests in Ethiopia, India and Vietnam

Obiageri Bridget Azubuike, Rhiannon Moore and Padmini Iyer

www.younglives.org.uk

MAY 2017

Young Lives School Surveys, 2016–17: The Design and Development of Cross-Country Maths and English Tests in Ethiopia, India and Vietnam

Obiageri Bridget Azubuike, Rhiannon Moore and Padmini Iyer

First published by Young Lives in May 2017

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 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 • E-mail: younglives@younglives.org.uk

Contents

Acknowledgements 4

1. Introduction to the survey 4

2. Maths assessment framework 5

3. English assessment framework 5

4. Item development and pilot sample 7

4.1. Pilot Sample 8

5. Test Data Analysis 9

5.1. Classical Test Theory (CTT) 9

5.2. Item Response Theory (IRT) 9

6. Pilot Data Preparation and Analysis 11

6.1. Analysis of pilot Maths test results 11

6.2. Analysis of pilot English test results 14

6.3. Interpretation of Item Characteristic Curves (ICC) 17

6.4. Distractor Analysis 18

7. Final Test Form Development 19

References 21

Appendix 1 22

Acknowledgements

The work discussed in this technical note was conducted by the Young Lives education team at the University of Oxford. The authors would like to thank Caine Rolleston, Jack Rossiter, and Zoe James who were all involved in the selection of cross-country Maths and English test items for the 2016-17 school survey, and the development of instruments used for the school survey in Ethiopia, India and Vietnam. The authors would also like to thank Educational Initiatives (in particular Nishchal Shukla and Gayatri Vaidya) and Phan Thi Luyen for their work on developing test items used in the surveys.

1. Introduction to the survey

Young Lives is an international study of childhood poverty in Ethiopia, India (Andhra Pradesh and Telangana), Peru and Vietnam. Since 2002, Young Lives household surveys have followed the lives of 12,000 children in these four countries in two age cohorts: an ‘Older Cohort’ born in 1994-95, and a ‘Younger Cohort’ born in 2001-02.

In 2010, the study introduced a series of school surveys in all four countries, which included a sub-sample of children in the Younger Cohort. Between 2010 and 2013 the school surveys examined issues of school quality and effectiveness in primary schools in Young Lives sites in Ethiopia, India (Andhra Pradesh and Telangana), Peru and Vietnam.

Building upon the design of the primary school surveys, the 2016-17 Young Lives school surveys examine school effectiveness at upper primary level in Ethiopia, and at secondary level in India and Vietnam (see Rossiter 2016, Moore 2016 and Iyer 2016 for a more detailed discussion of the school surveys in each country). The surveys examine school effectiveness through multiple outcome measures, including students’ learning progress in Maths and functional English. This involved the administration of Maths and functional English tests at the beginning and end of the school year (Wave 1 and Wave 2 of data collection respectively) in order to assess students’ learning progress in these domains.

This technical note focuses on the design and development of Maths and functional English cognitive tests for the 2016-17 school surveys. The note includes a discussion of the assessment frameworks used to design the tests, and the process of developing, piloting and selecting items for tests that were contextually relevant while also allowing for cross-country comparability across Ethiopia, India and Vietnam.

2. Maths assessment framework

In the 2016-17 school surveys, we conceptualise learning quality both in terms of progress on curriculum knowledge *and* students' ability to apply their knowledge and skills in unfamiliar contexts[1].

We identified the TIMSS Maths Assessment Framework as a useful way of assessing students' mathematical ability in these terms, as it distinguishes between the following three mathematical cognitive domains:

- Knowing: the facts, concepts and procedures students need to know;
- Applying: the ability of students to apply knowledge and the conceptual understanding to solve problems or answer questions;
- Reasoning: going beyond the solution of routine problems to encompass unfamiliar situations, complex contexts, and multi-step problems (Grønmo et al 2015: 24)

In addition to these cognitive domains, the Young Lives Maths tests are based around the appropriate mathematical content domains for each country. Using maths curricula for the survey grades in Ethiopia (Grades 7-8), India (Grade 9) and Vietnam (Grade 10), eight common content domains were identified:

- 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

3. English assessment framework

English language tests were included as part of the Young Lives school surveys in 2016-17 due to the status of English in Ethiopia, India and Vietnam as a 'transferable skill', with relevance for continuing education, labour market opportunities and social mobility (Graddol 2010). While students are exposed to English in varying contexts across the three countries (both within and beyond school), the language is seen as increasingly relevant by policymakers and individuals alike.

[1] See Iyer & Moore (2017) for a more detailed discussion of the way in which quality learning has been conceptualised in the 2016-17 school surveys.

The construct assessed in the Young Lives English tests is 'functional English', which is defined as the "application of […] skills in purposeful contexts and scenarios that reflect real-life situations" (OFQUAL 2011: 10). In this sense, the English tests diverge somewhat from the school curriculum in the three study countries, as they measure students' ability to use English in ways which have relevance for them. Due to practical and logistical considerations of conducting a large-scale survey, the test is comprised of multiple-choice questions – a limitation which means that it only captures reading skills, which are just one dimension of the functional English construct.

The Young Lives English tests are aligned with the Common European Framework of Reference for Languages (CEFR) developed by the Council of Europe (2001). This framework 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). The Common Reference Levels are detailed in Table 1.

Table 1: *CEFR Common Reference Levels (Council of Europe 2001)*

CEFR Level	Common reference level indicators
A1	Can understand and use familiar and everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.
A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.
B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topic issue giving the advantages and disadvantages of various options.
C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.

Within the functional English construct, the Young Lives English tests focus on the types of skills which reflect the ways in which the 15-year-olds in our study countries currently use English, or may need to do so in the future. The following four skill domains were identified:

- **Word identification**: Identifying simple vocabulary which students are likely to have been exposed to. With particular focus on language relating to their everyday environment and to education, questions in this skill bracket are particularly suited to learners at a lower level.

- **Word meaning and contextual vocabulary**: Identifying the meaning of unfamiliar words through their contextualised use in a sentence, or through identifying a synonym/antonym. Questions relating to this skill are of particular relevance for those who are likely to have greater exposure to English out school, or those who have a higher level of English.
- **Sentence construction and comprehension:** Completing sentences correctly, using appropriate grammatical concepts, and combining sentences together. Questions relating to this skill can be at a range of levels, but require learners to have some degree of understanding of the meaning of complete sentences.
- **Reading and comprehension:** Reading a range of texts (stories, posters, factual passages) and comprehending both direct facts and implicit inferences from them. Questions relating to this skill can be at a range of levels, but require learners to be able to read and have some understanding of English texts. Questions relating to implicit inferences rather than direct facts require a higher level of English language ability.

4. Item development and pilot sample

A total of 220 Maths items and 124 English items were selected for piloting across the three countries from January-March 2016. These items were developed in collaboration with Educational Initiatives, an educational research and assessment consultancy based in India, with the Ethiopia Ministry of Education’s Maths and Science Improvement Centre (MSIC), and with a Maths test consultant from the Vietnam Institute of Educational Sciences (VNIES). Maths and English items were mapped according to the assessment frameworks described above, with the complexity of items assessing each content and cognitive domain determined according to curriculum content, grade level (for the Maths tests), and CEFR level (for the English tests) within each country.

Maths and English items were pre-piloted and piloted in the three countries from March – May 2016. Qualitative pre-piloting with students aimed to check the suitability of item difficulty and content, and to identify any issues with translation. Maths and English teachers were also consulted during pre-pilots regarding the suitability of the tests for the target grades. Following revisions to the items based on student and teacher feedback, larger-scale pilots were conducted in each country.

For the Maths tests, three pilot forms (A, B, C) were prepared and administered in multiple-choice format using items from the combined item pool. In each country, a slightly different form length was required, depending on anticipated student literacy and the breadth of the curriculum, ranging from 39 items per form in Ethiopia to 43 items per form in Vietnam. A total of 45 pilot items were common across the three countries. The use of multiple forms allowed greater coverage of items in the pool, maximising the data available for later item selection.

For the English tests, two forms (A and B) were used in the pilot in each country, with a total of 86 cross-country common pilot items. In each country, a different form length was also required, depending on literacy and exposure to English language; 60 items were therefore included on each pilot English form in India, and 45 in Ethiopia and Vietnam. The English test was also administered in multiple-choice format.

4.1. Pilot Sample

Across the three countries, the pilot sample was drawn from different types of schools, different regions and localities. The aim was not for the pilot sample to be representative of the full sample, but rather for selected schools and students to reach the extremes of expected performance in each subject to identify potential ‘floor’ and ceiling’ effects, and to cover all language groups. Table 2 below provides an overview of the pilot sample in the three countries.

Table 2: *Sample size, location, school ownership and language of test in the pilot survey*

	Ethiopia	India	Vietnam
Number of students	1,232	593	486
Number of provinces / districts / regions	6	2	2
Locations	Urban and rural	Urban and rural	Urban and rural
Ownership	Private and government	Private and government	Government
Languages used	Bilingual Amharic/English Af Oromo Af Somali Tigrigna English	Bilingual: Telugu/English Urdu/English	Vietnamese

In Ethiopia, the pilot sample included 1,232 students across six regional states (Addis Ababa, Afar, Oromia, Somali, SNNP and Tigray). Both rural and urban schools were included and test languages were selected according to the official language of instruction in maths, at Grades 7 and 8, of the respective region. In regions that transition from Amharic to English between Grade 6 and 7, a bilingual form was used, with each item presented in both Amharic and English.

In India, the sample consisted of 593 students attending government schools, private schools and tribal welfare schools[2]. The sample was drawn from both rural and urban areas in two districts in Andhra Pradesh: Kurnool and West Godavari. A bilingual form was used in India, with each item presented in both English and Telugu. A smaller qualitative pilot was also undertaken with Urdu medium students at a later date, using a bilingual form with each item presented in both English and Urdu.

In Vietnam, the pilot sample included 486 students attending government schools, across both rural and urban areas in two provinces: Da Nang and Lao Cai. All forms were prepared and administered in Vietnamese.

Administering cross-country common items supported analysis which enabled us to evaluate what students in the three countries can do, and their learning levels in Maths and functional English. The following sections focus on the selection and validation of these cross-country items; analysis of students’ learning achievement, including benchmarking against curriculum expectations within each country, will be presented in forthcoming country and cross-country reports of the Young Lives 2016-17 school surveys.

[2] Tribal / Social Welfare schools are schools for children from minority groups in rural areas of India. They are mostly residential schools.

5. Test Data Analysis

Following pilot data collection in each country, test items were analysed to generate a range of statistics that would aid the item selection process for Wave 1 (beginning of the academic year) and Wave 2 (end of the academic year) tests in each country. To assess the reliability and validity of each item and the tests as a whole, analysis of pilot data was conducted using Classical Test Theory (CTT) and Item Response Theory (IRT). These analytical methods return a variety of test and item statistics, which were used to inform item selection. An explanation of the statistics provided by each analytical approach follows.

5.1. Classical Test Theory (CTT)

Classical test theory is a body of related psychometric theory which can be used to explain the difficulty of test items, insights into test score reliability, and the extent to which a test item conforms to items in the rest of the test (see Krishnan, 2013). In conducting CTT analysis, we aim to assess how reliably a student's test score measures the outcome of interest (e.g. a student's math skills).

With CTT, we can investigate the following indicators;

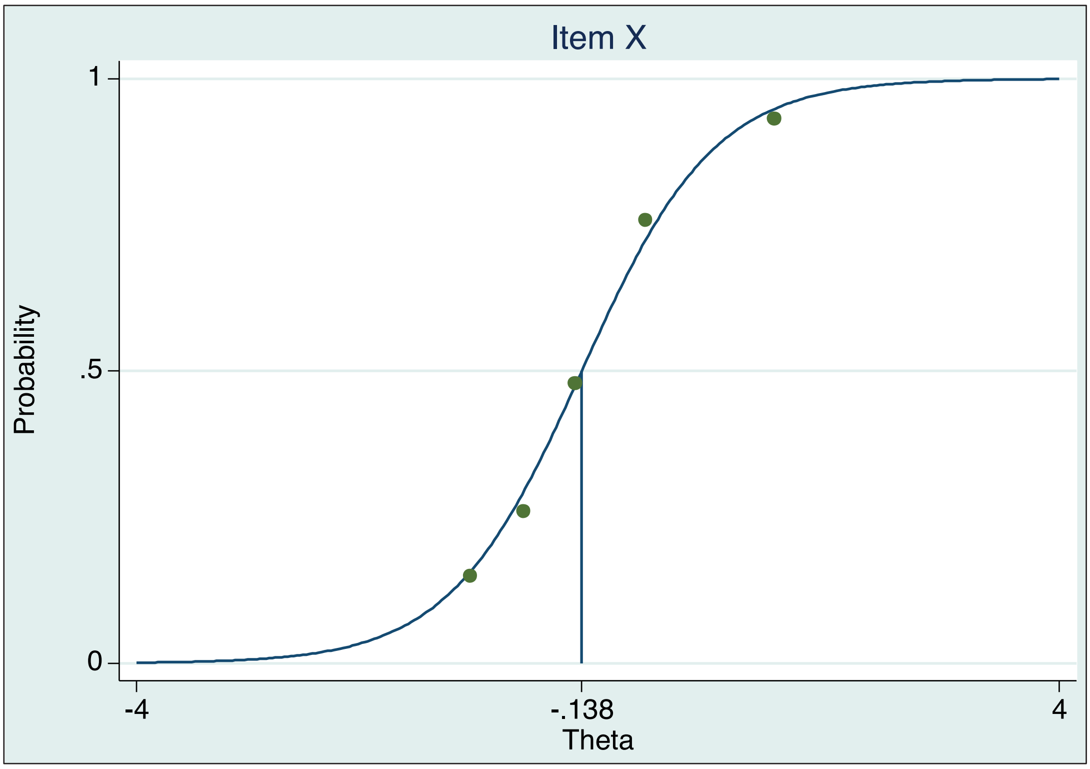
- **Item difficulty or *p*- values**: the difficulty index, indicated by *p*, is calculated as the ratio of the number of persons who answer an item correctly to the total number of test takers (percentage correct). Items with lower *p* values indicate higher difficulty and high *p* values indicate easier items. A general recommendation is to use items with *p* values within a range of 0.40 to 0.60 (with an average of 50% getting the item correct).
- **Item discrimination:** item discrimination index, which stands for the difference between the percentage of high performers and the percentage of low performers. It indicates the relationship between how well children did on an item and their total test score. This index ranges from –1 to 1, where positive values indicate that the item is discriminating in favour of high achievers, and negative values indicate that the item is discriminating against high achievers.
- **Reliability:** A reliable test is one we can trust or use to measure a person's performance approximately the same way each time. There are different internal consistency measures that can be used. The most commonly used is the Cronbach's alpha, which provides an indication of the average correlation among all items that make up the test. Cronbach's coefficient alpha evaluates the degree to which different items "pull together" the same content area – and therefore provide a reliable estimate of the individual's underlying trait of interest. The alpha values range from 0 to 1.00, with higher values indicating greater reliability.

5.2. Item Response Theory (IRT)

Item Response Theory is a model that attempts to explain the relationship between latent traits (unobservable attributes) and their manifestations (observed outcomes or performance). With regards to maths tests, it models the response of each student of a given maths ability to each item in the test. IRT assumes that as the trait level increases, so does the probability of a correct response (see Edelen & Reeve, 2007).

The item characteristic curve (ICC) is the basis of IRT. It yields a trace line (s-shaped) that is described by the difficulty and the slope (discrimination) of the item. Figure 1 below provides an illustration of an ICC. The dots on the curve represent the mean quintiles of students' estimated maths ability; the higher up on the curve a student is, the higher the probability of getting item X correctly and the higher the students' estimated maths ability (theta). The point at which a student has a 50% chance of getting item X correctly is the estimated difficulty of item X (-0.138) and the median student's estimated maths ability (theta). The ICC provides a graphical analysis of the item parameters described below and has been used in the analysis described here to select common items across the three countries.

Figure 1: *Example Item Characteristic Curve (ICC)*



Item parameters in IRT are:

- Discrimination (a): The slope parameter is also known as the discrimination of the item, it represents the slope of the ICC at the difficulty level. This slope also indicates the extent to which the item is related to the underlying construct and the ability of a test item to distinguish between individuals of differing ability/knowledge. A steeper slope indicates a closer relationship to the construct and therefore it is more discriminating.
- Difficulty (b): The difficulty level (also known as location parameter) is simply defined as the point on the ICC at which the probability of a positive response to the item is 50%. The location parameter is usually between -2 and 2, with a mean of around zero. The higher the location parameter, the more difficult the item, and a respondent must have a higher ability (the measured construct) to answer that item correctly.
- Guessing (c): an estimate of the probability with which a student with no knowledge of the item can obtain a correct response.

There are three types of IRT models:

- One parameter logistic (1-PL) model is the simplest form of IRT models, which predicts the probability of giving the correct response to an item as a function of the respondent's ability and the difficulty of an item. The discrimination parameter in a 1-PL model is fixed for all items.

- Two parameter logistic (2-PL) model predicts the probability of giving the correct response to an item as a function of the respondent's ability, item difficulty and the discrimination of the item. The discrimination parameter is allowed to vary between items.
- Three parameter logistic (3-PL) assumes that the probability of an individual getting an item right is dependent on the three factors above; item discrimination, item difficulty and guessing.

For the purpose of this analysis, the IRT two parameter (2PL) model was used[3]. The following sections describe the pilot data preparation for analysis, the analysis conducted for item selection, and the results of the analysis using the models described above.

6. Pilot Data Preparation and Analysis

The test data from all three countries were imported into STATA and all cross-country test items were matched using item identification numbers. The cross-country items were renamed to be consistent with the item ID and all the student's responses were scored. A dichotomous variable was created, where the codes used were 0 and 1 for incorrect and correct responses respectively.

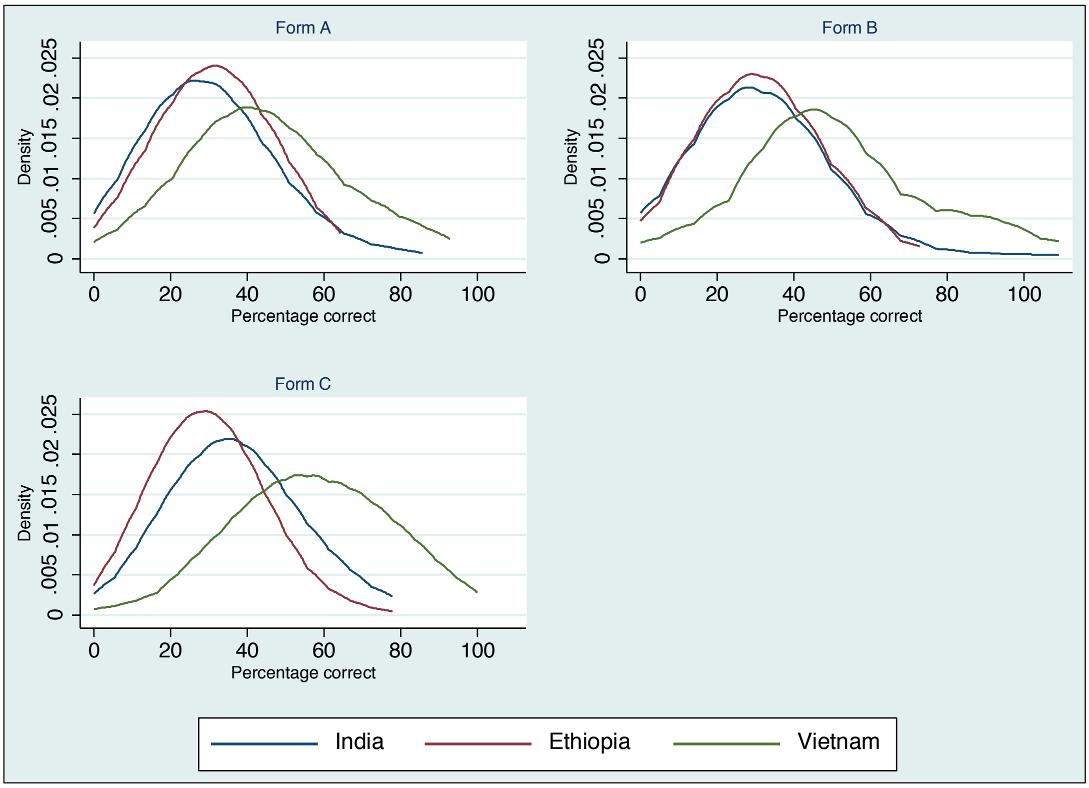
This analysis focused on the common items administered across the three countries for the purpose of cross-country analysis. A form-wise analysis was conducted on each of the common items administered in each country i.e. identical items from within the three forms for Maths and two forms for English in each country were analysed together. While separate analysis was conducted for individual country items, the cross-country analysis using IRT enabled the selection of items that performed well across the three countries and thus provides scope for a cross-country comparison – and later investigation of student performance on a common scale across countries.

6.1. Analysis of pilot Maths test results

Firstly, scores on the common items were summed and the mean scores (percentage correct) were estimated for each country on each form. Figure 2 below provides a first descriptive view of the overall scores on the cross-country items.

[3] Due to sample size requirements for estimating a 3PL IRT model, the IRT 2PL model was used in this analysis. Additionally, the design of the tests included distractor options (see distractor analysis below) in the multiple-choice options, which are intended to reduce the probability of guessing the correct response to an item. In light of this test design, the 2PL model was felt to be more suitable.

Figure 2: *Distribution of cross-country overall mean scores for the three maths forms*

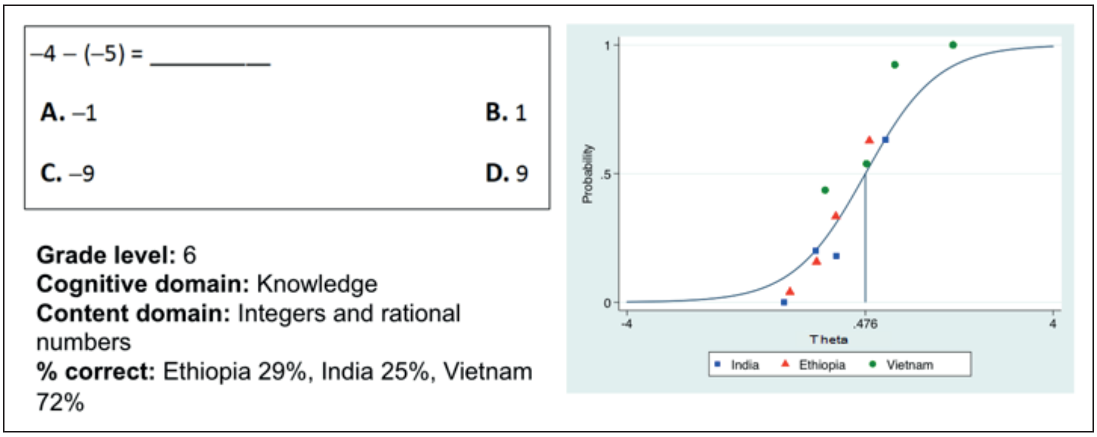


From the graphs above, the cross-country items across the three forms have a similar distribution. The curves in Ethiopia and India are more similar than that of Vietnam. The scores for the samples in Ethiopia and India are narrower and slightly more skewed to the left than the scores for the sample in Vietnam. The graphs show that students in the Ethiopia and India samples scored lower on the cross-country items than their counterparts in Vietnam. Although the students in all three countries are of a similar age, there are differences in their school grades, curriculum and school systems; these differences were also considered in the selection of these items.

Figures 3, 4 and 5 below provide examples of items which assess the three different cognitive domains within our mathematical assessment framework (knowledge, application and reasoning), along with results from the pilot tests in each country. These items are examples of those which functioned well, and which were therefore included as cross-country anchor items for Wave 1 of the 2016-17 school surveys.

The functioning of the items shown below broadly suggest that the ‘reasoning’ item is harder than the ‘knowledge’ and ‘application’ item in Ethiopia and Vietnam, although the Grade 6 ‘knowledge’ item in Figure 5 proved difficult in Ethiopia and India. 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 in Maths, Indian students generally in the middle, and Ethiopian students towards the lower end. Importantly, for each of these items, there is overlap in students’ performance across the three countries. This means that the items work well as anchor items that allow the three tests to be put on the same scale.

Figure 3: *Example of a cross-country Maths item assessing 'knowledge'*



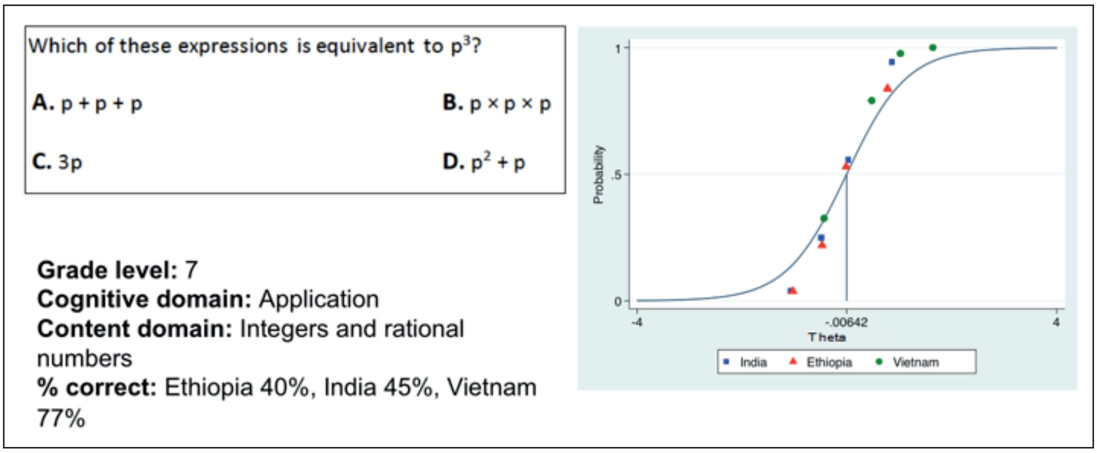
–4 – (–5) = ________

A. –1 **B.** 1

C. –9 **D.** 9

Grade level: 6
Cognitive domain: Knowledge
Content domain: Integers and rational numbers
% correct: Ethiopia 29%, India 25%, Vietnam 72%

Figure 4: *Example of a cross-country Maths item assessing 'application'*



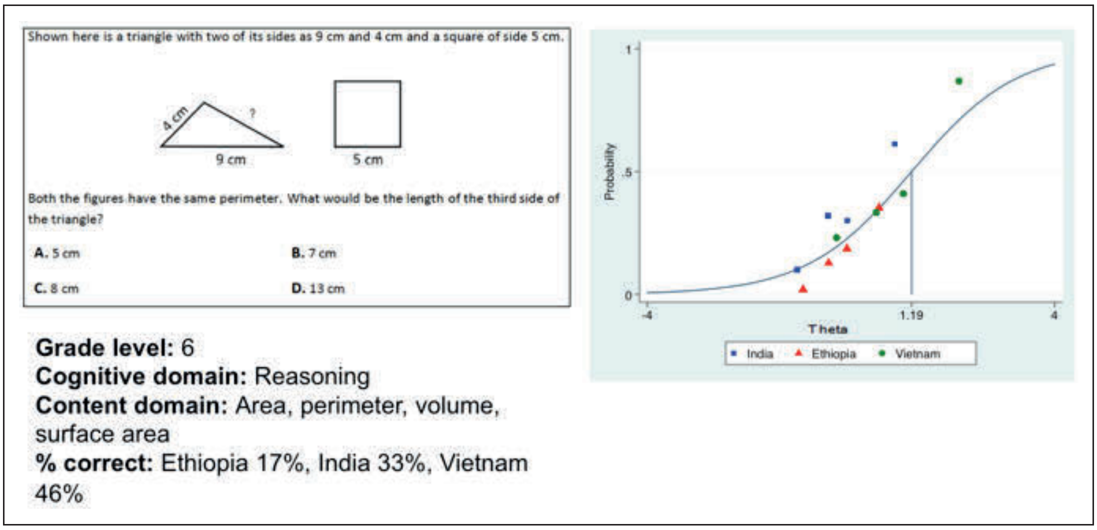
Which of these expressions is equivalent to p^3?

A. p + p + p **B.** p × p × p

C. 3p **D.** $p^2 + p$

Grade level: 7
Cognitive domain: Application
Content domain: Integers and rational numbers
% correct: Ethiopia 40%, India 45%, Vietnam 77%

Figure 5: *Example of a cross-country Maths item assessing 'reasoning'*



Shown here is a triangle with two of its sides as 9 cm and 4 cm and a square of side 5 cm.

Both the figures have the same perimeter. What would be the length of the third side of the triangle?

A. 5 cm **B.** 7 cm

C. 8 cm **D.** 13 cm

Grade level: 6
Cognitive domain: Reasoning
Content domain: Area, perimeter, volume, surface area
% correct: Ethiopia 17%, India 33%, Vietnam 46%

Along with item-level analysis of pilot data, the final selection of items for the Wave 1 and Wave 2 Maths tests in Ethiopia, India and Vietnam has also been led by different priorities within each country. For example, Young Lives' primary school survey findings (James & Rolleston 2015) and pilot data for the present 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 Maths test in Wave 1 is made up of knowledge items, and the test also includes lower grade level items. By contrast, existing Young Lives data (Rolleston et al 2013) and PISA 2012 results point to generally high levels of mathematical knowledge among students in Vietnam. 65% of the items on the Vietnam Wave 1 Maths test therefore assess Grade 10 students' ability to apply their mathematical knowledge or to use mathematical reasoning skills in less familiar contexts. In Wave 2, items were also selected with reference to analysis of data from Wave 1, in addition to the pilot data analysis.

6.2. Analysis of pilot English test results

Figure 6: *Distribution of cross country overall mean scores for English forms*

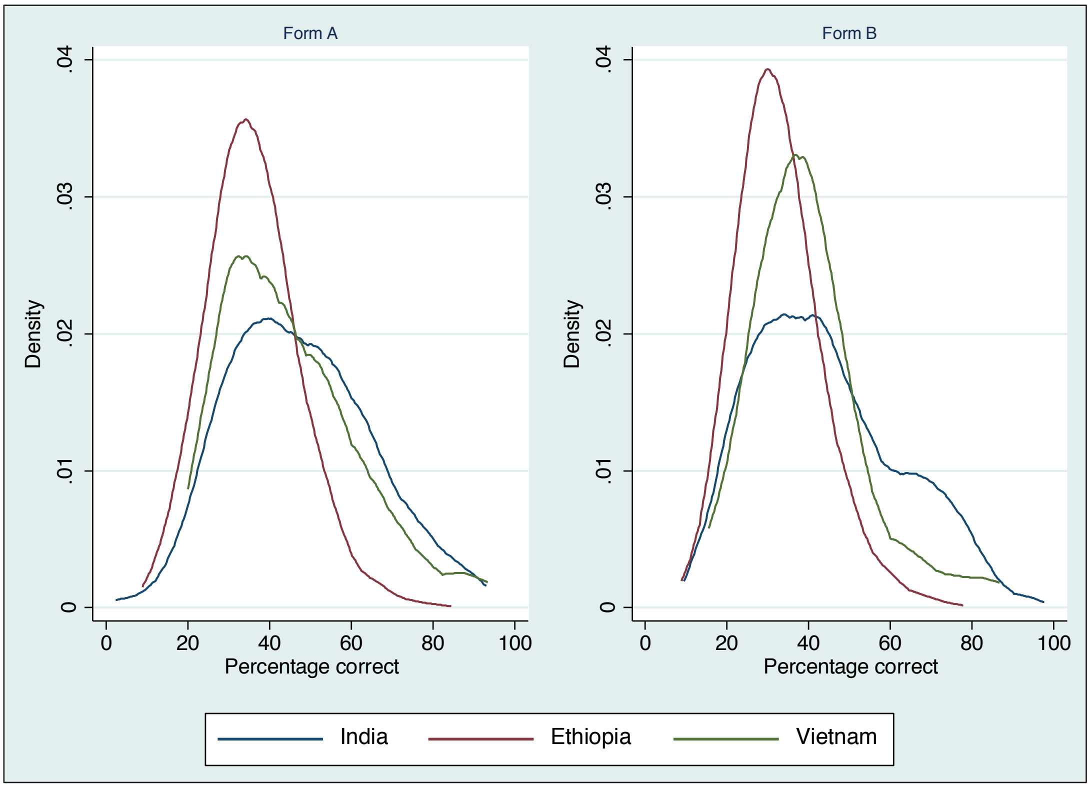
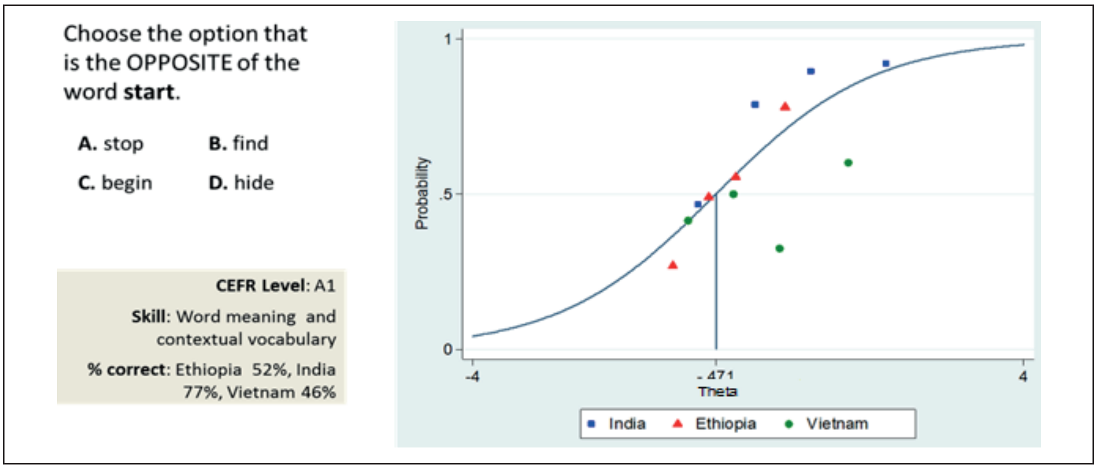


Figure 6 above represents the distribution of overall scores in the piloted English test for the common items administered across the three countries. The scores in the two forms have a similar distribution, with the student sample in India having higher scores in the English test on both forms. The curves for the students in Ethiopia are narrower than both Vietnam and India which suggests that students' scores are more similar in Ethiopia than India and Vietnam. Figures 7, 8 and 9 provide examples of items which assess three of the four skill domains within our English assessment framework (word meaning and contextual vocabulary, sentence construction, and reading comprehension), along with results from the pilots in each country.

Figure 7: *Example of an English item assessing 'word meaning and contextual vocabulary'*



Choose the option that is the OPPOSITE of the word **start**.

A. stop **B.** find

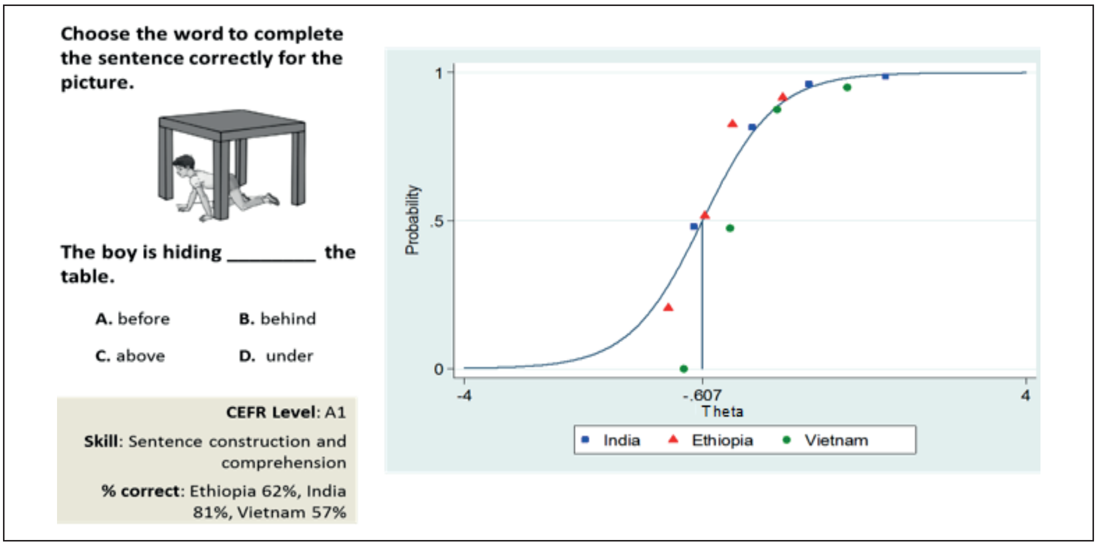
C. begin **D.** hide

CEFR Level: A1

Skill: Word meaning and contextual vocabulary

% correct: Ethiopia 52%, India 77%, Vietnam 46%

Figure 8: *Example of an English item assessing 'sentence construction and comprehension'*



Choose the word to complete the sentence correctly for the picture.

The boy is hiding ________ the table.

A. before **B.** behind

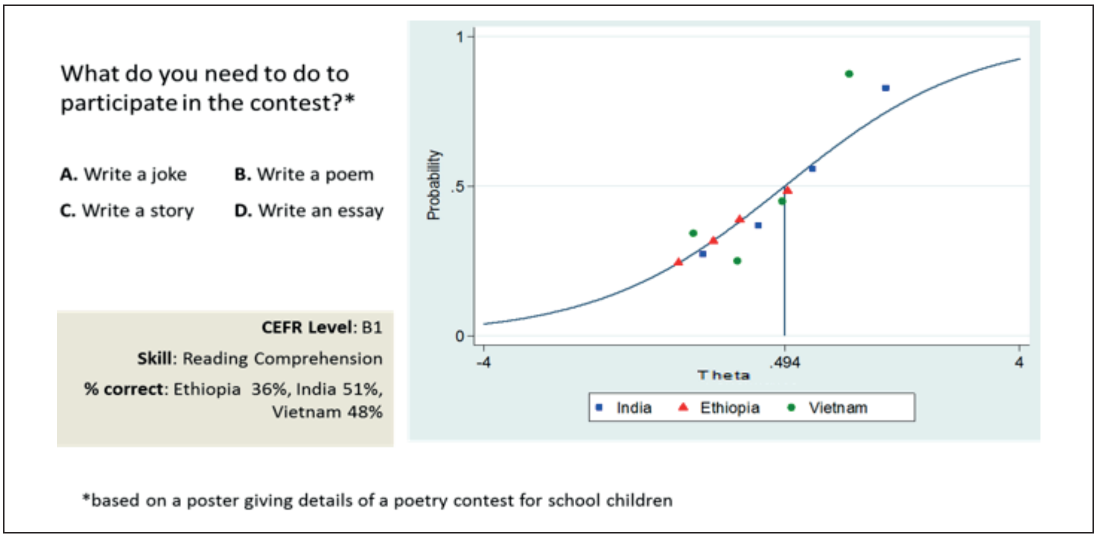
C. above **D.** under

CEFR Level: A1

Skill: Sentence construction and comprehension

% correct: Ethiopia 62%, India 81%, Vietnam 57%

Figure 9: *Example of an English item assessing 'reading comprehension'*



What do you need to do to participate in the contest?*

A. Write a joke **B.** Write a poem

C. Write a story **D.** Write an essay

CEFR Level: B1

Skill: Reading Comprehension

% correct: Ethiopia 36%, India 51%, Vietnam 48%

*based on a poster giving details of a poetry contest for school children

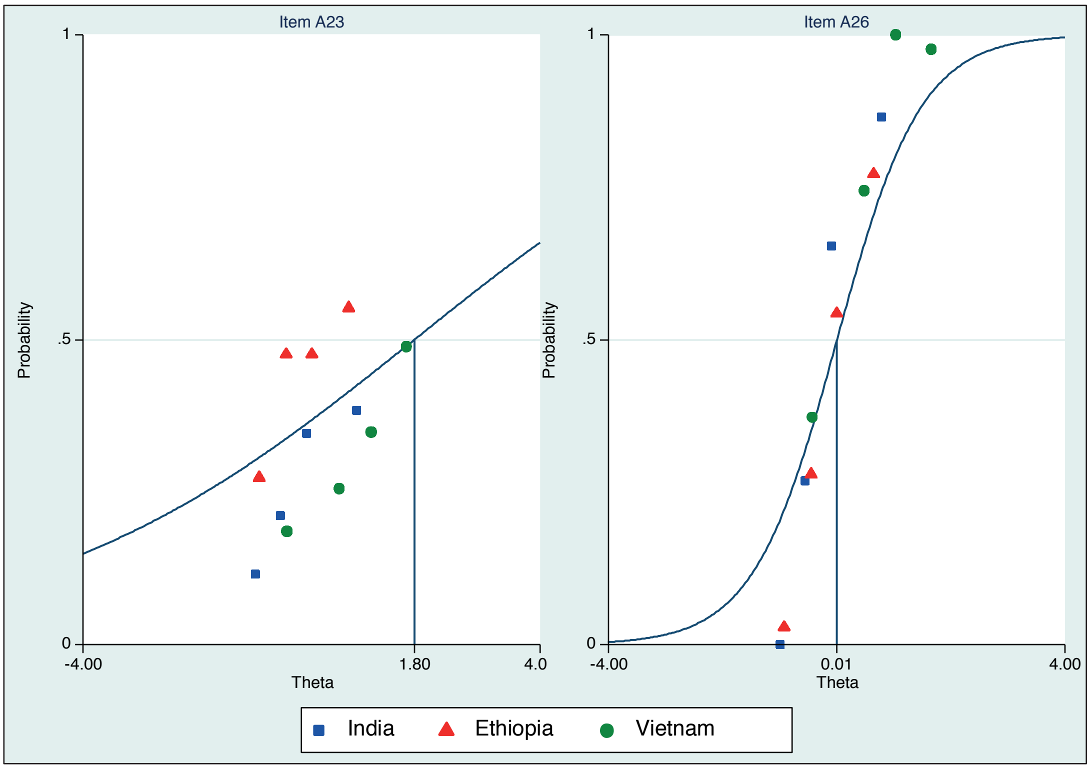
These items are examples of those which functioned well, and which were therefore included as cross-country anchor items for the Wave 1 English tests. Common items across the three countries were identified in three of the four skill domains. There were no items in the 'Word identification' skill domain which suitable for inclusion across all three countries at Wave 1; these items were mostly at lower CEFR levels and pilot data suggested that they were too easy to be included in the test in India. Some of these items have been retained in the Wave 1 English tests in Ethiopia and Vietnam but excluded from the test in India.

Data from the piloting of the English tests suggested that items at CEFR level A1-A2 functioned well in Vietnam and Ethiopia, while in India the items that functioned best were at CEFR levels A2-B1. The functioning of the items in Figures 7, 8 and 9 suggests that the 'word meaning and contextual vocabulary' item was harder than the 'sentence construction and comprehension' item in all three countries. In Ethiopia and India the 'reading comprehension' item was the hardest of the three shown. Importantly, for each of these items, there is an overlap in students' performance across the three countries, which indicates that they function well to support cross-country analysis.

Although our interest in all three countries is to assess functional English across the four skill areas, the balance of skills included in the final English tests varied according to the different priorities and country contexts, in addition to pilot item functioning in each country. For example, in Ethiopia and Vietnam, existing literature and pilot data suggested that levels of English are likely to be between CEFR levels A1-B1, and as a result the Wave 1 English tests in these countries contained a larger proportion of word identification questions (25% and 20% of the items respectively) which were at a lower CEFR level. Meanwhile in India, pilot findings suggested that English proficiency levels were slightly higher than in the other two countries (around A2-C1), and existing literature confirmed that students in India would be more exposed to English outside the classroom than in our other study countries (Graddol, 2010). As a result, a higher proportion of items on the Wave 1 English test in India assessed word meaning and contextual vocabulary (24%), with a smaller number of lower level word identification items included (8%).

6.3. Interpretation of Item Characteristic Curves (ICC)

Figure 10: *Example Item characteristic curves (ICC) for two maths test items*



In Figure 10, the two ICC curves provide an overview of IRT parameters for two maths items: one item that was excluded from the final test (item A23) and one item that was retained for the final test (item A26). Each dot on the ICC represents the mean of quartiles for the estimated theta (predicted Maths ability score), with different colours and shapes representing the three countries. The ICC curve for item A23 has a poor fit to the IRT model; the item difficulty is slightly high at 1.80 and students in the top ability quartile have around 50% chance of selecting the right response to item A26. The slope of the curve (defined by the item's discrimination) is almost flat, which indicates that item A23 discriminates poorly between students of differing maths abilities.

The ICC for item A26 has a steep and positive discrimination, which shows that the item discriminates well between students in different ability groups - i.e. the higher a student's ability, the higher the probability of a positive response on the item. The graphs thus show how difficult the items are and how much they discriminate between the ability quartiles in each of the three countries. Following a review of the ICC curves for all the cross-country items administered in both Maths and English, the decisions to retain items was based on the fit of the pilot data to the two parameter IRT model[4], analysis of CTT item difficulty and distractor analysis (discussed below).

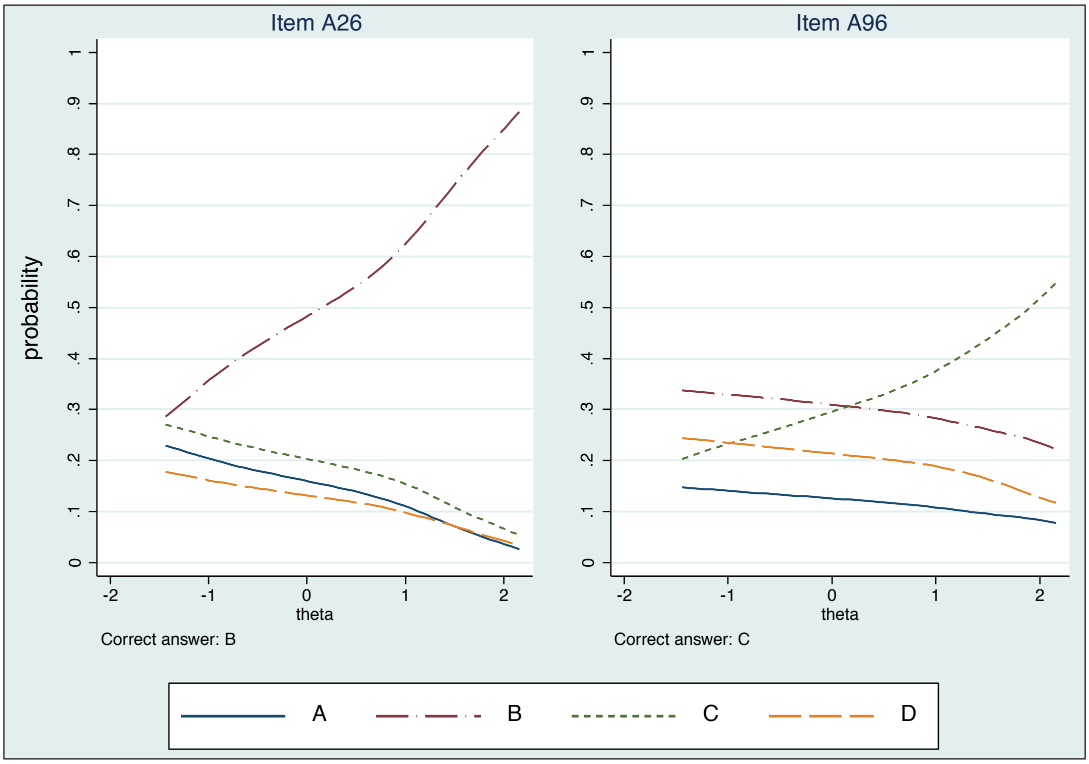
[4] See Appendix 1 for all cross-country item characteristic curves.

6.4. Distractor Analysis

As the items administered were multiple-choice, distractor graphs were produced to allow us to review how students responded to the correct option versus the three other available options, given their estimated ability levels. When analysing distractor graphs we expect that as ability increases, students become more likely to choose the correct answer. Alternative answer options distract most students from the correct option, leaving only the higher ability students to select the correct option.

Figure 11 below uses the cross-country pooled responses to items A26 and A29 for illustration purposes. On item A26, even students at the lowest estimated ability level (theta) were more likely to select the correct response (B) than any other option. By contrast, on item A96, students at the lower end of the estimated ability scale were more likely to select a wrong option than those on the higher end of the ability scale. Option B distracts students from C (the correct option) until the point where their estimated ability is about the mean; from this point, they become more likely to select the correct option. The distractor graphs were disaggregated at the country level and reviewed during the selection process to ensure we identified any items with distractors that functioned differently across the countries. In some cases, this highlighted issues related to translation or formatting which needed to be addressed.

Figure 11: *Distractor graphs in pooled analysis of item A26 and Item A96*



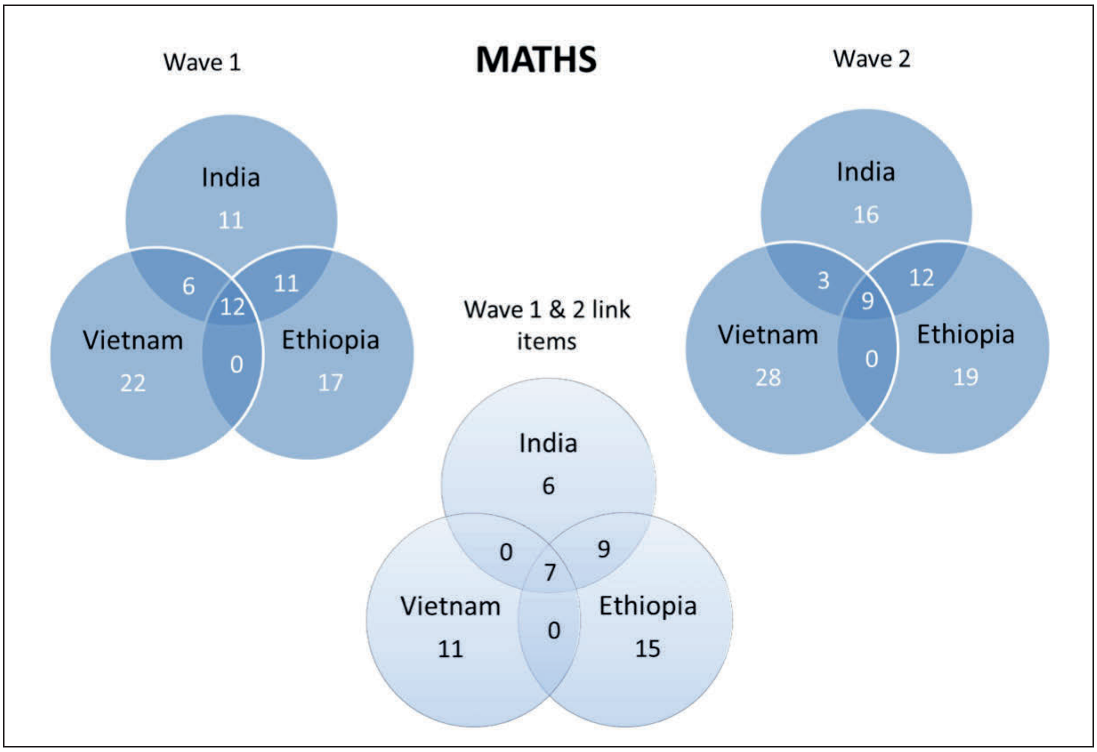
The item information function was also reviewed in during the selection process. The item information function (IIF) is plotted after estimating the parameters of an IRT model. It is used to describe the precision of an item in a test, and to examine the range of abilities for which the test items capture information. More reliable items tend to measure the construct of interest around the estimated difficulty parameter with greater precision. Figure A4 in Appendix 1 provides the graphs illustrating the IIF for each of the retained items form the cross-country items.

7. Final Test Form Development

This process of item selection led to the development of three, 40-item multiple-choice Maths tests to be administered at the start of the year – one for each of the countries within the survey. There are 12 cross-country common items on each Wave 1 Maths test, which were retained from the analysis described above; pilot data for the other unique items administered in each country were analysed, reviewed and selected using a similar process to make up the complete test form of 40 items.

As mentioned earlier, in line with the school effectiveness design of the survey, Math tests were administered at the beginning of the school year (Wave 1) and at the end of the school year (Wave 2). Following Wave 1, three 40-item Maths tests were developed for Wave 2 (one for each country), with items selected based on analysis of pilot and Wave 1 data. Items that functioned well in terms of the IRT model fit, potential to show progress using data from the Wave 1 survey and a balance of difficulty, cognitive and content domains were more likely to be included in the Wave 2 tests. Some of the cross-country common items from Wave 1 were retained and administered in Wave 2 in order to compare learning progress across the three countries. Each Wave 2 Maths test contains 9 cross-country common items. Of these, 7 items are common between Wave 1 and Wave 2 for all three countries. There are also additional link items between two of the countries across waves, which will also allow us to compare learning progress (see Figure 12).

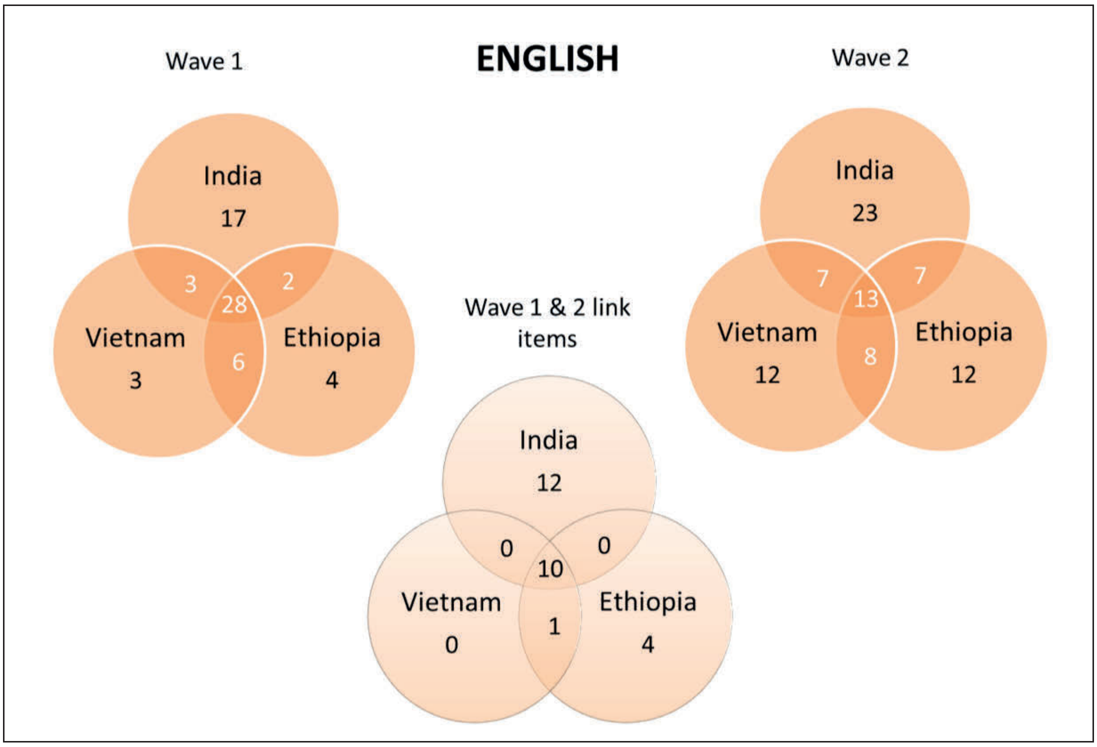
Figure 12: *Cross-country and cross-wave link items for Maths*



The item selection process for the English test led to the development of three, multiple-choice English tests to be administered at the start of the year – one for each of the countries within the survey. The test in India was longer (50 items) than in Ethiopia and Vietnam (40 items); pilot results and existing literature indicated that there would be a greater range of ability levels within the India sample which would be best captured through a longer test form. For Wave 1, there were 28 common English items used across the three countries, with a small number of additional common items between two of the three.

As with Maths, the Wave 2 English tests for the end of the school year were developed using pilot data and Wave 1 data to identify items to include in Wave 2. Three multiple-choice English tests were developed (one for each country); as in Wave 1, the Wave 2 test in India was longer (50 items) than in Ethiopia or Vietnam (40 items each). In Wave 2, 13 English items were common across the three countries. Of these, 10 were common across both Wave 1 and Wave 2, and across all three countries. There are also a smaller number of items common for one or two countries and across both waves (Figure 13).

Figure 13: *Cross-country and cross-wave link items for English*



Forthcoming country reports for Ethiopia, India and Vietnam will present descriptive findings from the 2016-17 Young Lives school surveys in each country, while a forthcoming cross-country report will present descriptive cross-country analysis from the 2016-17 surveys, including a comparison of student progress across the three countries supported by the design outlined in this technical note.

References

Council of Europe (2001) *Common European Framework for Reference Of Languages: Learning, Teaching, Assessment.* Strasbourg, France: Language Policy Unit. [Online] http://www.coe.int/t/dg4/linguistic/Cadre1_en.asp [Accessed 27/07/2016]

Edelen, M. O., & Reeve, B. B. (2007). 'Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement'. *Quality of Life Research*, *16*(1), 5.

Graddol, D. (2010) *English Next: India*. London, UK: British Council. [online] http://issuu.com/britishcouncilindia/docs/english_next_india_-_david_graddol?e=1710046/8032873 [Accessed 27/07/2016]

Grønmo , L. S., M. Lindquist, A. Arora & I. V. S Mullis (2015) *TIMSS 2015 Mathematics Framework*. Boston, MA: TIMSS & PIRLS International Study Centre.

Iyer, P. (2016) *The design of the 2016-17 Young Lives School Survey in Vietnam*. Technical Note 38. Oxford: Young Lives.

Iyer, P. & Moore, R. (2017) 'Measuring learning quality in Ethiopia, India and Vietnam: from primary to secondary school effectiveness'. *Compare: A Journal of Comparative and International Education*. DOI: http://dx.doi.org/10.1080/03057925.2017.1322492.

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.

Krishnan, V. (2013) "The early child development instrument (EDI): An item analysis using classical test theory (CTT) on Alberta's data." Early Child Development Mapping (ECMap) Project Alberta, Community-University Partnership (CUP), Faculty of Extension, University of Alberta, Edmonton, Alberta.

Moore, R. (2016) *The design of the 2016-17 Young Lives School Survey in India*. Technical Note 37. Oxford: Young Lives.

OFQUAL (2011) *Functional Skills Criteria for English. Entry 1, Entry 2, Entry 3, Level 1 and Level 2.* Coventry, UK: OFQUAL.

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.

Appendix 1

Figure A1: *Item Characteristic Curves (ICC) for Maths cross-country items: Form A*

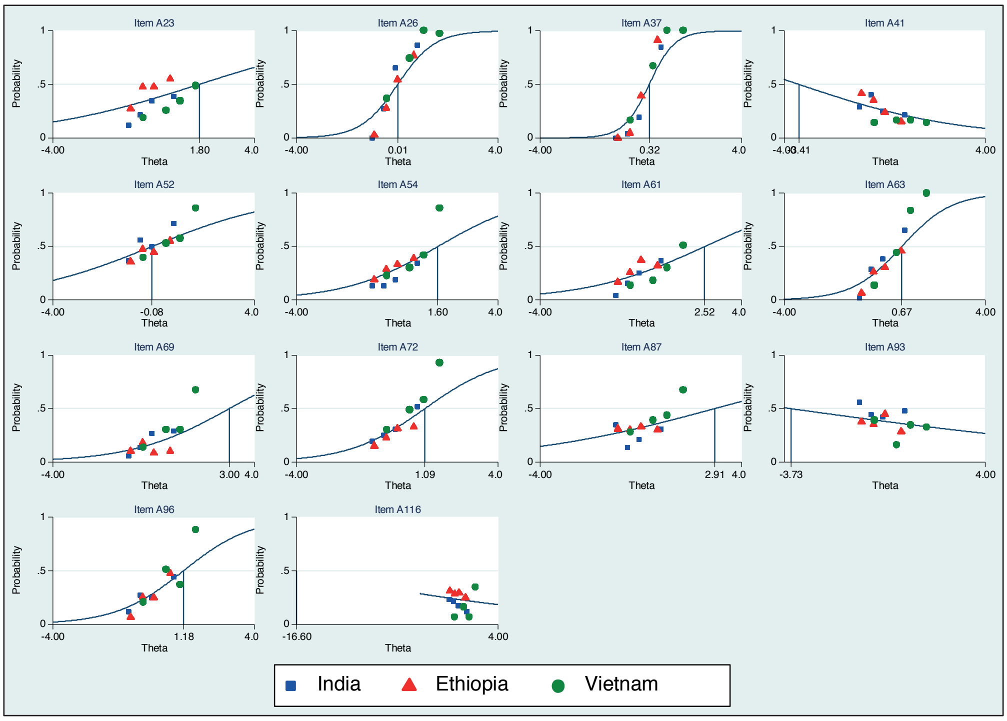


Figure A2: *Item Characteristic Curves (ICC) for Maths cross-country items: Form B*

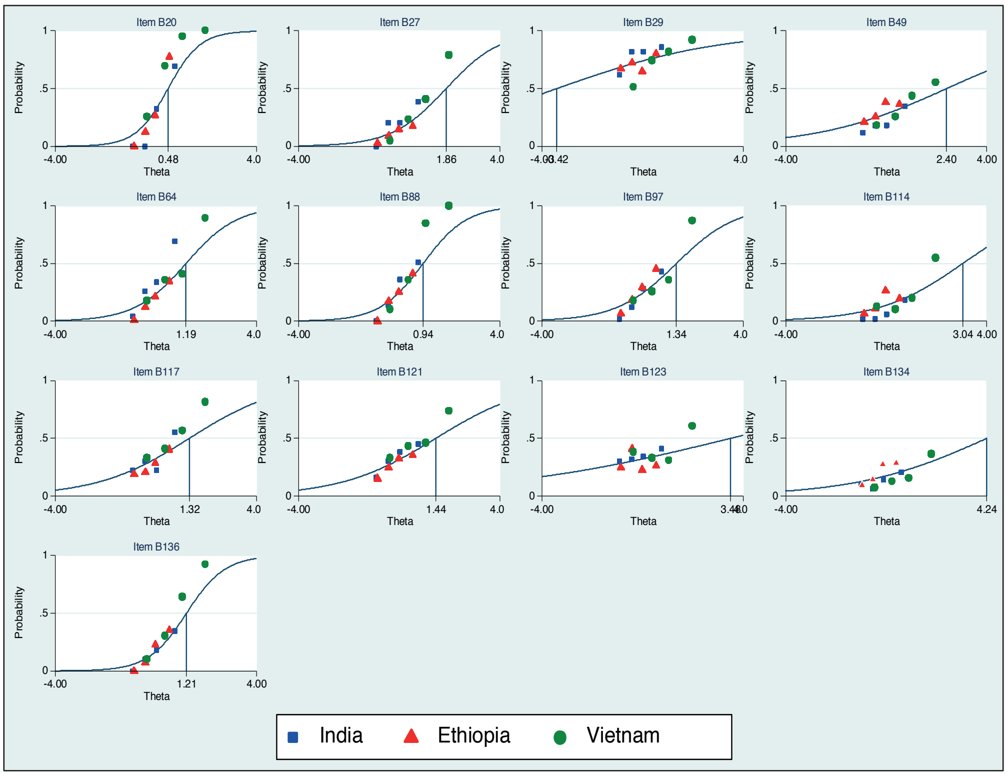


Figure A3: *Item Characteristic Curves (ICC) for Maths cross-country items: Form C*

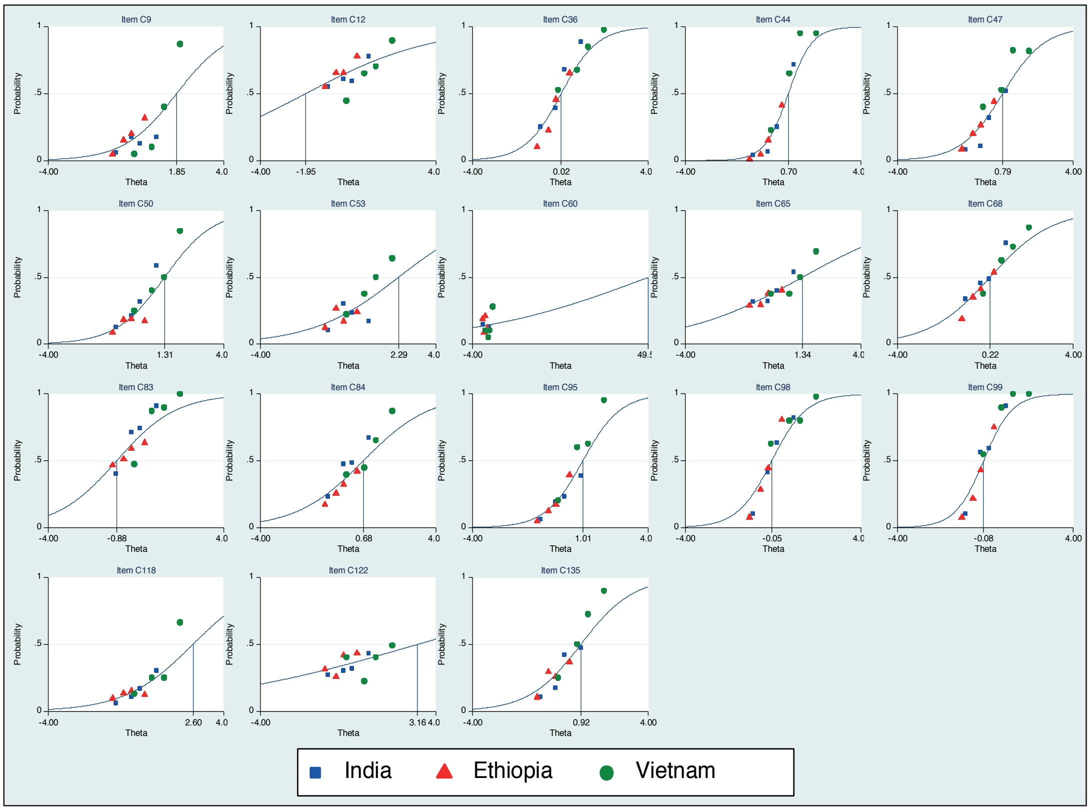


Figure A4: *Item Characteristic Curves (ICC) for English cross-country items: Form A1*

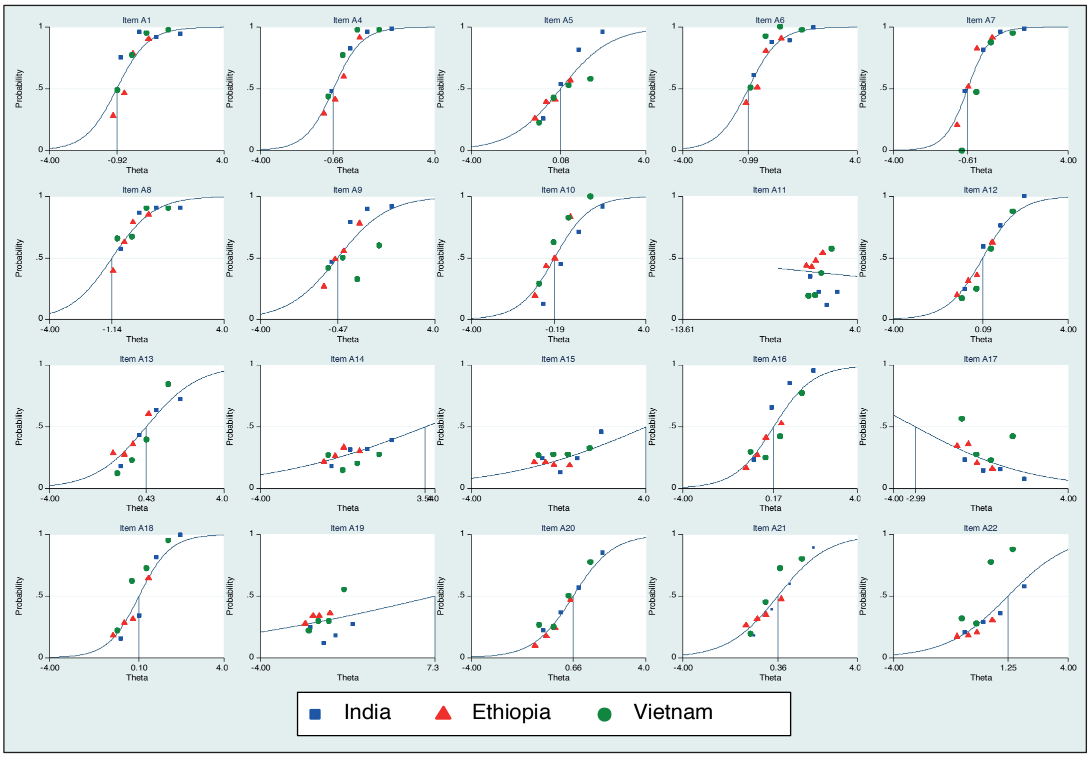


Figure A5: *Item Characteristic Curves (ICC) for English cross-country items: Form A2*

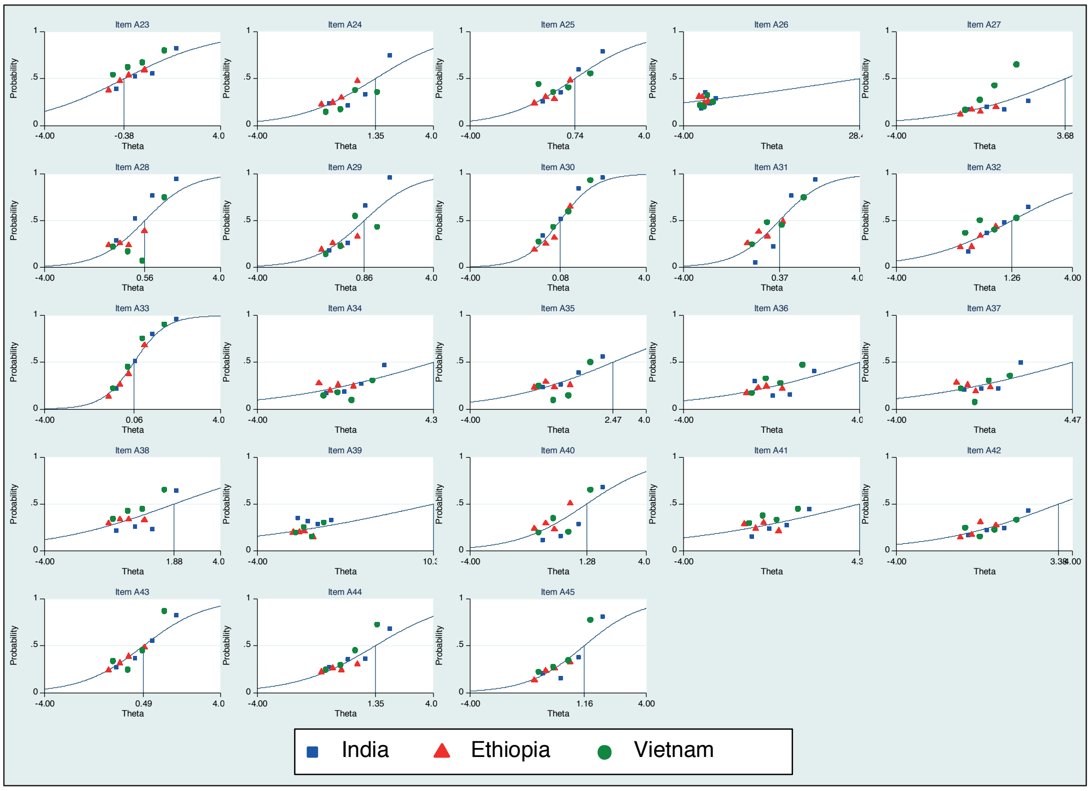


Figure A6: *Item Characteristic Curves (ICC) for English cross-country items: Form B1*

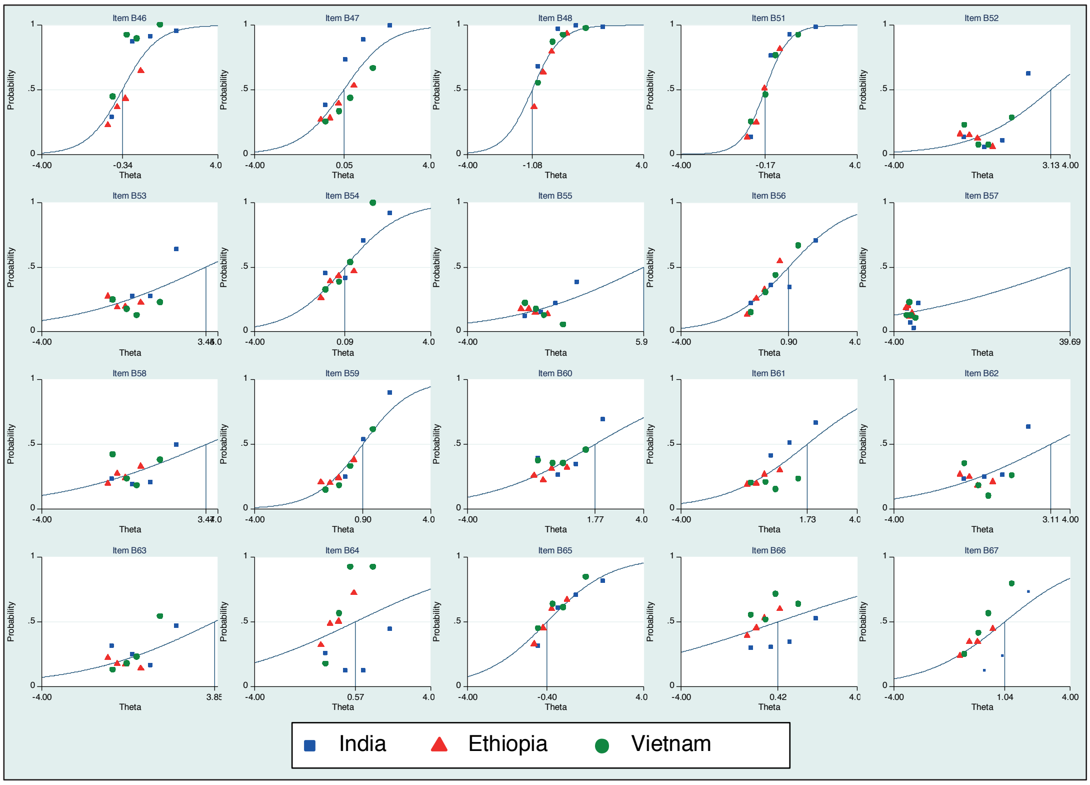


Figure A7: *Item Characteristic Curves (ICC) for English cross-country items: Form B2*

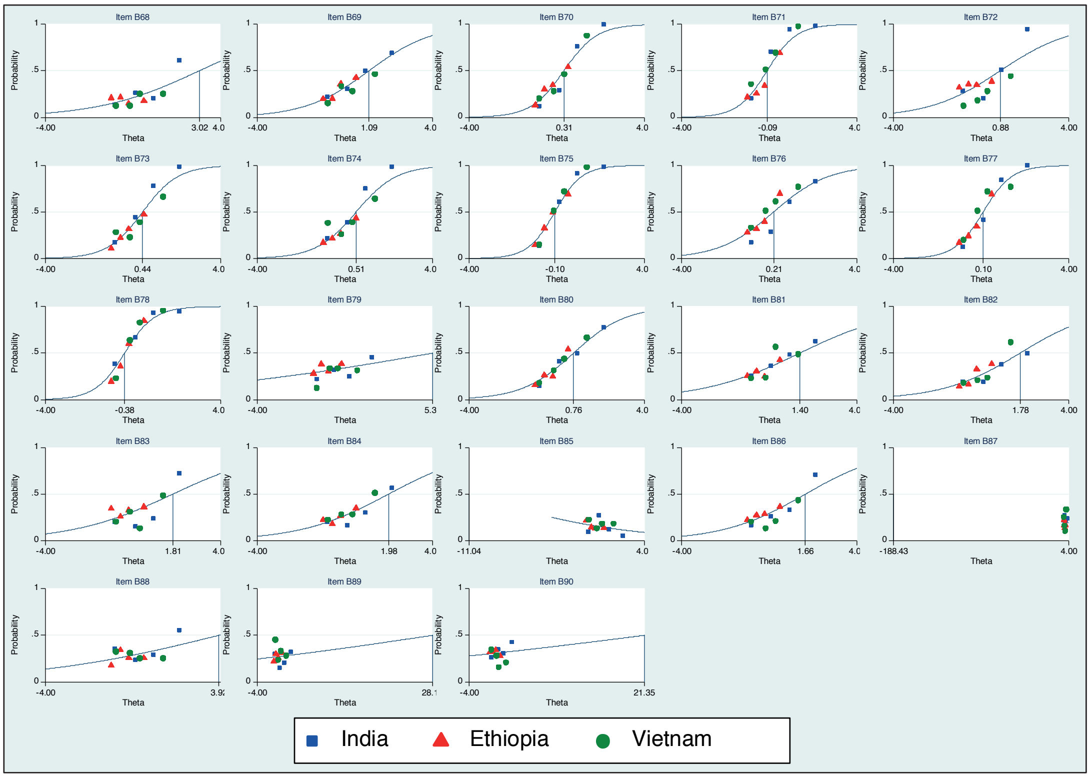
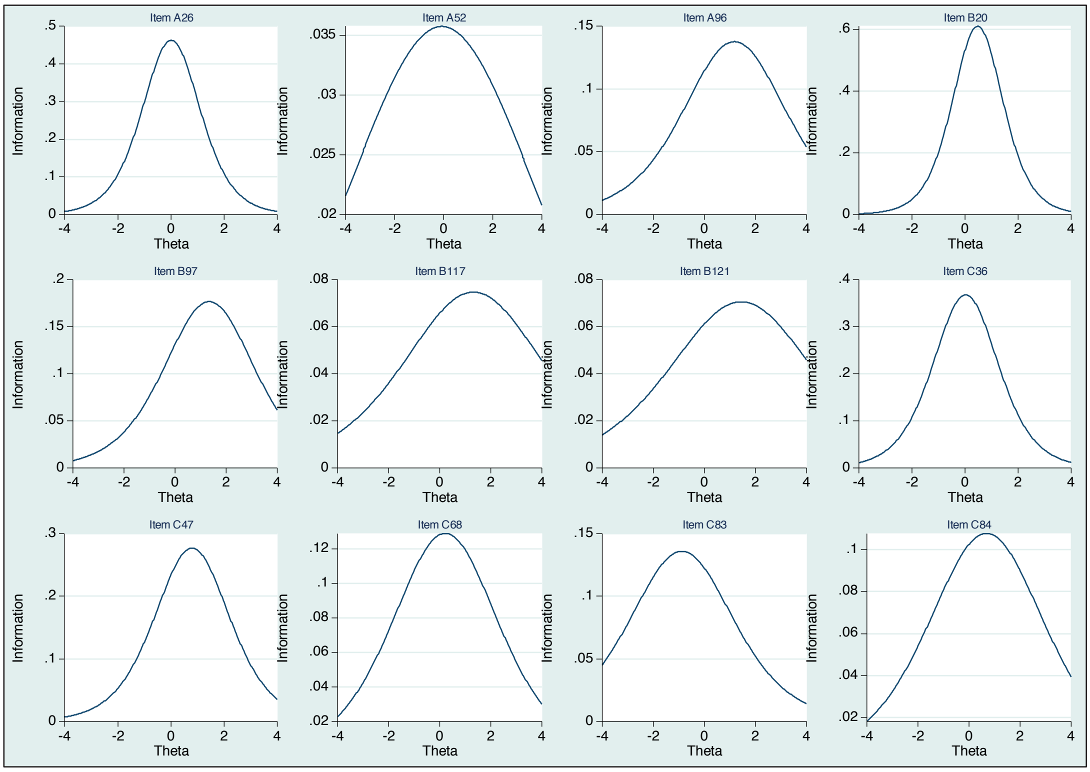


Figure A8: *Example Item Information Function (IIF) graphs for retained items in Maths*



Young Lives School Surveys, 2016–17: The Design and Development of Cross-Country Maths and English Tests in Ethiopia, India and Vietnam

An International Study of Childhood Poverty

About Young Lives

Young Lives is an international study of childhood poverty, involving 12,000 children in 4 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 4 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