The Design of Achievement Tests in the Young Lives Ethiopia School Survey Round 2

Zoe James

1.0 Introduction

Young Lives is an international longitudinal study of childhood poverty in Ethiopia, India (the states of Andhra Pradesh and Telangana), Peru and Vietnam. It combines data collection at the household-level (carried out in 2002, 2006, 2009, 2013/14 and planned for 2016), with longitudinal qualitative research and a newly introduced nested school survey (since 2010).

In Ethiopia, Young Lives has been following 3,000 children in two age cohorts since 2002 (1,000 children age 7-8 in 2002 and 2,000 children age 6-18 months in 2002). These children are spread across twenty purposively selected 'sentinel sites' which are broadly illustrative of the diversity of the country (Outes-Leon & Sanchez 2008). In addition to the longitudinal household survey and qualitative work, Young Lives has conducted two school surveys in Ethiopia.

The first of these school surveys, conducted in 2010, sought to add school level data to the household panel, following a subsample of Young Lives children to their schools. The second school survey, conducted in the 2012-13 academic year, sought to provide evidence on the school, class, teacher and pupil level factors that help or hinder children's learning of core curricular domains over the course of grade 4 and grade 5. To this end a key feature of its design was assessments of children's competency in maths and reading comprehension, linked to the Ministry of Education's Minimum Learning Competencies and building on existing assessments of learning in Ethiopia e.g. the USAID-funded Early Grade Reading Assessment (USAID 2010), as well as Young Lives' household surveys. These tests took place at both the beginning and end of the 2012-13 school year (wave 1 and wave 2), with the aim of enabling value-added analysis.

The second Ethiopia school survey included all children (both Young Lives younger cohort children and non-Young Lives children) studying in all grade four and grade five classes in all schools located within the geographic boundaries of each sentinel site. In addition to the twenty core Young Lives sites, nine additional sites in the Somali and Afar regions were added to enhance the survey coverage and include two so-called "emerging regions".

Relative to the other Young Lives study countries, Ethiopia has a complicated linguistic environment, with over 80 languages in use. Since the 1994 Education and Training Policy , Ethiopia has implemented a mother tongue education programme, in which teaching in mother tongue nationality languages at the primary level is compulsory. Children in the Young Lives Ethiopia school survey therefore have any one of at least ten different mother tongues, and learn in grades 4 and 5 in some eight different mediums of instruction. This poses a complicated scenario in which to implement meaningful large scale norm-referenced tests which ideally aim to compare all children, across linguistic groups, on a common scale.

2.0 Design and Piloting of Wave 1 and Wave 2 Tests

This section outlines the stages involved in the design and piloting of the tests and the procedures for administration in the field at both wave 1 and wave 2 of the survey.

2.1 Wave 1

2.1.1 Design

Following a review by the team of existing school-based assessments in Ethiopia (most notably the 2010 Early Grade Reading Assessment (EGRA) (USAID 2010) and the 2004 and 2008 grade 4 Ethiopian National Learning Assessments (NLA) (NOE (National Organisation for Examiners Addis Ababa) 2008b; NOE (National Organisation for Examiners Addis Ababa) 2004b; NOE (National Organisation for Examiners Addis Ababa) 2004a; NOE (National Organisation for Examiners Addis Ababa) 2008a), a 'test and retest' design involving the administration of tests linked to the Ministry of Education's (MOE) Minimum Learning Competencies (MLCs) in the core domains of maths and reading (with a focus on comprehension), was arrived at. These domains are both central to what is taught in primary schools in Ethiopia, and are comparable to constructs assessed in previous rounds of the Young Lives household survey. According to the 'test and retest' design, the survey tests were conducted at both the beginning and end of the school year (wave 1 and wave 2), with the aim of enabling value-added analysis of school effectiveness.

A multiple choice format was selected largely out of necessity, since the anticipated sample size (estimated before fieldwork at ~13,000) meant that tests suited to individual administration (like EGRA), or tests which, following completion, required detailed marking

of children's written responses, were beyond the resourcing parameters of the survey. A last decision involved the type of test to be developed: i) norm-referenced, or ii) criterion referenced. The decision on this issue was taken based on the interpretation of the scores for later analysis. When scores are norm-referenced, relative score interpretations are of primary interest. A score for a child is ranked within a distribution of scores or compared to the average performance of test takers for several reference populations. Criterion-referenced scores on the other hand convey an absolute level of competence in some defined criterion domain. It was decided that norm-referenced scores were more suited to the multi-purpose nature of the Young Lives study, in which assessment data is used both independently, as well as in conjunction with the household survey.

2.1.2 Maths

In the first instance the team reviewed the MLCs for grades 1-5, alongside textbooks. A bank of items were then developed, each matching to a sub-competency of an MLC and reflecting questions which might reasonably be expected to reflect the content of lessons across the regions included in the survey. Where appropriate, items which had functioned well in household round 3 and other school surveys were included. Whilst the survey was targeted at grades 4 and 5, many items related to target competencies for grades 1-3 were included, following a review of the achievement levels of the younger cohort in maths and reading at round 3 of the household survey which suggested children in grades 4 and 5 may perform below the curricular expectations for those grades.

Consultants were then recruited in each of the languages of instruction in which the survey was to be conducted (Amharic, Oromiffa, Tigrigna, Sidama, Wolayta, Hadiyya, Afar, Somali)[1]. A key requirement for these consultants was strong working and up-to-date knowledge of the language in question and familiarity with the current use of language in schools. To this end, a number of the consultants were sourced from staff at regional teacher training colleges, or were themselves teachers or former-teachers.

For the purposes of piloting, two rotated form tests (with common items and similar levels of difficulty) were developed to enable as many items as possible to be piloted. Items were selected for inclusion in the two pilot tests to ensure the test had a good spread of items both

[1] The ideal situation would have been to source two consultants in each language and compare their tests etc, but this wasn't possible within our resource constraints

in terms of difficulty from grade1-grade5 and in terms of domain, so that the test included items that assessed number operations, measurement and so on, to the extent that these were reflected in the MLCs. The aim after piloting was to then use this extended bank of pilot items to generate a single test in each domain for use in the final survey. Whilst this initial development work took place in English, this version was then adapted into Amharic and was reviewed, prior to the beginning of the adaptation process in the other languages. Consultants were then provided with both the English and Amharic versions of the two tests as well as detailed instructions and support on how to adapt the tests into their language. Whilst items which were entirely numeric needed little adaptation except to ensure that the notation was correct for the region (for example, the symbol used to denote division), any items which contained words and written instructions were adapted to ensure the correct conveyance of meaning and difficulty level, rather than exact translation of the English or Amharic phrasing.

2.1.3 Reading comprehension (mother tongue)

A slightly different process was involved in the development of the reading comprehension tests, owing to the diversity of mediums of instruction in which the tests were administered. Recognising these differences, and the non-comparability of simply translated words both in terms of item difficulty level and familiarity across diverse contexts, the team instead developed independent tests in each language in a process of both 'adaptation' and 'assembly' (Hambleton et al. 2005). These followed a pre-defined structure linked to competencies identified as key in the review of the MLCs and textbooks. A common format was then laid out in prototype English and Amharic versions split into four sections: the first matching words and pictures, the second sentences and pictures, the third asking children to fill in the blanks with the correct word, and the fourth a reading comprehension passage.

Consultants were asked to develop a test in their language, using the prototype version as a guide of structure and approximate content. This ensured a common format across languages, and the assessment of common domains or competencies across the different languages. Consultants were asked to think carefully about the level of difficulty of each item and section in English and Amharic, and to replicate a similar level of difficulty in their own language version. In essence this was not simply an exercise in translation of a single test, but rather one in which different tests were produced following a common format under the understanding that translation would in any case have produced qualitatively different tests.

The reading comprehension section of the test drew on the stimuli material developed for the 2010 Early Grade Reading Assessment, for those languages which were common to both surveys (Amharic, Oromiffa, Tigrigna, Sidama and Somali). Where a stimulus passage had not been developed for EGRA in the Young Lives survey language (Hadiyya, Wolayta), consultants used the Amharic and English prototype as a guide to develop a new passage and associated questions. Questions aimed to assess three different skills, namely a) word identification and selection, b) sentence application and interpretation and c) passage comprehension, interpretation and evaluation.

2.2 Pilot testing

Piloting took place in six regions: Afar, Amhara, Oromia, SNNP, Somali and Tigray in the week commencing 1st October 2012. This involved piloting in five of the seven survey languages, with Sidama selected in SNNP[2]. In Afar, whilst a test had been developed in Afar language (since it was understood that it was in use at least in ABE schools), piloting revealed that in practice Amharic was almost always the medium of instruction. Pilot data was therefore collected for five of the survey languages: Amharic, Oromiffa, Sidama, Somali and Tigrigna. The number of schools, classes and children involved in the piloting is detailed below.

Table 1: Summary of coverage of survey pilot

Region	**Language of instruction**	**Number of schools**	**Number of classes**	**Number of pupils**
Afar	Amharic	1	2	20
Amhara	Amharic	1	2	63
Oromia	Oromifa	1	2	109
SNNP (Sidama zone)	Sidama	1	2	86
Somali	Somali	1	2	60
Tigrigna	Tigrigna	1	2	81

[2] Note that primary schools in Young Lives sites in SNNP teach in Amharic, Sidama, Hadiya and Wolayta but Sidama was selected for the purposes of the pilot since both Young Lives and EGRA data show notably low performance among children learning in this language, making it particularly important to ensure the test is appropriate. It was not possible to pilot in all SNNP MOI for resourcing reasons and this of course poses limitations on the reliability of the data.

Schools were selected in sites similar to nearby Young Lives sites, with the proviso that they were not going to be included in the final survey fieldwork. In each site a single school was selected, and the pilot tests conducted with one grade 4 and one grade 5 class. Since there were two rotated forms of the maths test, half of each class was asked to take form A, and the other half form B. During piloting teachers from both grades 4 and 5 were asked to review the survey tests and comment on their validity for use in each region, both in terms of content and coverage, and language use. Their feedback was incorporated in the revision of items.

Data were then entered and analysed. Techniques from classical test theory (CTT) were used to examine individual item functioning, and to identify items that were insufficiently difficult (over 75% of children answered them correctly) or excessively difficult (fewer than 25% of children answered them correctly) so that they poorly discriminated. The topic coverage and grade level (test balance) of these items were carefully reviewed prior to deciding whether they would be included or removed, to ensure the test maintained a balanced coverage of key domains, and that items at different grade levels were retained, including those in grade 4 and 5 which children found harder, to enable room for respondents to demonstrate progress over the grade 4-5 year. Since a key aim of the design of these tests was to enable linking across language groups, the pilot data was also examined for Differential Item Functioning (DIF) across language groups and where DIF was identified in a test item, a 'distractor analysis' using techniques from CTT was conducted in which the probability that children in each language group selected each of options A, B, C or D was examined. Where significant differences existed between languages, test consultants were asked to review and revise items to refine the difficulty level of either the correct answer or the distractor(s) to ensure greater consistency of item difficulty across language groups. A full list of test items included in Wave 1 and their relationship to school grades and the MLCs can be found in Appendix 1.

2.3 Wave 2

2.3.1 Design

The design of the tests included in the second wave of the school survey, otherwise referred to as the 'retest', was somewhat simpler. This was because a key tenet of the overall design was that the tests administered at both the beginning and end of the school years should be able to be linked using a 'common item' approach. In light of the resource challenges posed by redeveloping and piloting multiple items in multiple languages for the retest, it was

therefore decided that as many items as possible should be replicated between the first and second waves of the test, with some new items being added which corresponded to grade 5 MLCs in maths and reading comprehension. Ultimately, the aim of the retest was to link to wave 1, to provide sufficient variation in ability between children, and to assess competencies which children would be reasonably expected to have learnt during the course of grade 4 and grade 5 i.e. to assess progress during the academic year.

The selection of items for replication followed detailed analysis of the first entry of the data from wave 1 of the survey. Techniques from both CTT and IRT (Item Response Theory) were used to examine each item, with particular attention being paid to DIF across language groups. Items which the majority of children had answered correctly and/or which demonstrated significant DIF across languages at the start of the survey were dropped, and were replaced with items which had been piloted at wave 1 of the school survey or round 4 of the household survey (piloting for which happened to coincide with the development of these second wave school survey assessments), and which assessed competencies on the Grade 5 curriculum. Whilst some of the harder items, particularly in maths, appeared to function poorly insofar as the percentage of children able to answer them correctly was small, these were sometimes retained in wave 2 since they covered more difficult grade 4 or grade 5 competencies and their functioning at wave 1 was thought to relate to the fact that the competency they assessed had not yet been sufficiently covered in class rather than to a problem with the item itself. Their replication at wave 2 of the survey enables us to look at progress on these higher level competencies during a single academic year. In total, 6 items were removed from each of the wave 1 tests and replaced with harder items, some of which link to household survey round 4. A full list of the items included in wave 2, how they relate to wave 1, school grades and the MLCs can be found in Appendix 1.

The inclusion of items from round 4 of the household survey not only helped to increase the difficulty level of the test, but should hopefully enable the school survey assessment scores to be equated with the household survey assessment scores using a 'common item' approach.

2.3.2 Procedures for administration of tests in the field

Procedures for administration of the tests were consistent between wave 1 and wave 2 of the survey, and full details can be found in the survey manuals which are available online with

the survey documentation. It was very important that the conditions and rules for administration enabled children to perform to the best of their ability.

For grade 4 classes, tests were administered in the language of instruction of the class. For grade 5 classes, tests were administered in the language of instruction of that class in grade 4. This was because in some regions the medium of instruction changes in grade 5, but since children had had as yet limited exposure to the new medium of instruction, it was decided that conducting assessment in the language of the previous year would better enable children to answer to the best of their ability, as well as enabling comparison between grades 4 and 5.

In all cases, tests were administered in a whole class environment. Where desks were arranged in a way which may have impeded children's ability to work independently, they were re-arranged. At the start of each test one fieldworker explained the instructions and wrote an example of how to complete a multiple choice question on the blackboard. Children were given one period or 45 minutes to complete the test, and the fieldworkers wrote the start and end time of the test on the blackboard. During the assessment the fieldworkers circulated the room, clarifying doubts and encouraging children where necessary. Children were instructed to attempt all questions.

3.0 Post-hoc review of wave 1 and wave 2 tests

Initial review of the wave 1 and wave 2 test data revealed some response patterns which looked inconsistent across languages in the reading comprehension. One possible reason for this was errors in the answer key, another was errors in the test content and/or translation. Whilst thorough piloting, test review by teachers, and statistical analysis had taken place at the design stage, the complexity of working across languages might be expected to introduce this error, and it was felt that it was important to investigate it more thoroughly a) to gather a 'correct' answer key, and b) to learn where errors might have been introduced so that any problem items could be removed/ corrected, and more generally to facilitate institutional learning for future assessment design in this context. A post-hoc review of the instruments in each language was therefore conducted between January and March 2014.

Two teachers currently teaching in each region (in SNNP in each zone) were asked to review the wave 1 and wave 2 maths and reading comprehension tests. Teachers were then

interviewed, and the interviewer completed a questionnaire in which teachers commented on: the correct answer(s), the clarity of the language, the appropriateness of the mathematical notation, the cultural appropriateness of any pictures and of the passage and questions, the grade in which the competency under consideration would be taught, and the grade in which teachers would expect children to be able to answer an item correctly. Where errors were identified, or the two teachers disagreed and could not reach agreement, the reasons for the error and disagreement were recorded in a free-form text box to aid interpretation of the results. The same interviewer conducted the fieldwork in all of the regions and zones, allowing for significant learning as the process progressed. For example, it became apparent that having an additional teacher from higher (non-self-contained) grades was helpful, so the majority of interviews actually involved three teachers of mathematics and reading comprehension from both the self-contained (grades1-4) and non-self-contained (Grade5-8) systems. The full interview protocol is included in Appendix 2

Data were reviewed to construct new answer keys for the reading comprehension tests. It was revealed that particular problems exist in the Sidama and Wolayta translations, and this should be borne in mind in analyzing the data since spelling and grammatical errors are particularly pervasive in these tests. Items with critical errors in any language, including spelling or meaning errors, or totally unfamiliar pictures were removed/ dropped. A revised language-wise answer key can be found in Appendix 3. It should be noted that this process has created different total scores for each language, and the reading comprehension tests should be treated separately by language.

Fewer critical errors were found in the maths test and so it was decided to keep the answer key the same for that test, which can be found next to the corresponding test items in Appendix 1.

Bibliography

Hambleton, R.K., Merenda, P.F. & Spielberger, C.D. eds., 2005. *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*, Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

NOE (National Organisation for Examiners Addis Ababa), 2004a. *Ethiopian Second National Learning Assessment of Grade 4 Students*, Addis Ababa: National Organisation for Examinations.

NOE (National Organisation for Examiners Addis Ababa), 2004b. *Ethiopian Second National Learning Assessment of Grade 8 Students*, Addis Ababa: National Organisation for Examinations.

NOE (National Organisation for Examiners Addis Ababa), 2008a. *Ethiopian Third National Learning Assessment of Grade Eight Students*, Addis Ababa: National Organisation for Examinations.

NOE (National Organisation for Examiners Addis Ababa), 2008b. *Ethiopian Third National Learning Assessment of Grade Four Students*, Addis Ababa: National Organisation for Examinations.

Outes-Leon, I. & Sanchez, A., 2008. *An Assessment of the Young Lives Sampling Approach in Ethiopia. Technical Note 1*, Oxford, UK: Young Lives.

USAID, 2010. *Ethiopia Early Grade Reading Assessment: Data Analytic Report*, Addis Ababa: USAID Ethiopia.

Appendix 1. School Survey Test specifications

Table 1. Maths test items administered in Waves 1 & 2 & answer key

	Wave 1 Question	Target Grade	MLC Competency	Correct Answer?	Keep?	Wave 2 Item	Grade	MLC Competency	Correct Answer?
1	How many dots are there?	<1	counting	D	N	Which of these is equal to 342?	3	"addition" "add whole numbers up to 10,000"	C
2	Put numbers in ascending order: 19, 6, 2, 11	1	"numbers - whole"	A	Y		1		
3	Which is a triangle?	1	"Shapes"	A	N	Which of these is the name for 9740?	3	"Whole Numbers" "read and write whole numbers up to 10,000"	B
4	2+3=___	1	Numbers - addition	A	Y		1		
5	9x2=____	1	Numbers - multip	A	Y		1		
6	15+12-3=____	2	(Numbers - add/sub	C	Y		2		
7	How many minutes in 1 hour?	2	Time	B	N	It takes Chris 4 minutes to wash a window. He wants to know how many minutes it will take him to wash 8 windows at this rate. He should?	2	"Multiplication" "solve word problems involving multiplication using 1 digit numbers and 10"	A
8	Tamiru has 5 Birr. His mother takes 4 Birr. How many Birr does Tamiru have left?	3	Numbers - addition Money	D	Y		3		
9	9Birr-6Birr=_____Birr	2	Money	A	Y				

10	Which is half of 6?	3	Numbers - fractions	B	Y		3		
11	Which number is closest to 900,000?	4	Numbers - whole	C	Y		4		
12	85x5=____	2	Numbers -multip	A	Y		2		
13	2.34+7.65=_____	4	Numbers - decimals	C	Y		4		
14	How many cents in 1 Birr?	1	Money	A	N	A cake was cut into 8 pieces of equal size. John ate 3 pieces of the cake. What fraction of the cake did John eat?	4	Fractions "Identify fractions as parts of a whole"	B
15	What part is shaded?	2	Numbers - fractions	A	Y		2		
16	Which difference is closest to 300000?	4	Numbers - sub	B	Y		4		
17	2488/8=____	3	Numbers - division	B	Y		3		
18	2kg=____g	3	measurement-weight	D	Y		3		
19	30m=_____cm	4	measurement - weight	D	N	A piece of rope 204cm long is cut into 4 equal pieces. Which of these gives the length of each piece in cm?	4	"numbers-division" "solve word problems involving division of whole numbers up to 1000000 by 1 digit numbers"	D
20	What is the value of the number 2 in 928?	2	numbers - whole"	A	Y		2		
21	What is average of 10, 12, 18, 24?	4	Patterns + graphs	B	Y		4		
22	Calculate the perimeter of the rectangle:	5	Measurement - perimeter	C	Y		5		
23	4.465-1.286=____	5	Numbers - sub/dec	C	Y		5		
24	Fill in the appropriate number in the sequence: 1, 3, __, 27	N/A	Non-curricular	A	N	Maria has 6 red boxes	NA	NON CURRIC R4	D
25	What is the area of the square below:	5	Area	C	Y		5		

Table 2. Reading Comprehension test items administered in Waves 1 & 2

WAVE 1					WAVE 2	
	Wave 1 Question	Keep in wave 2?	If Keep Q No. in Wave 2			Wave 2 Question
1	Picture word	N			1	Picture word
2	Picture word	N			2	Picture word
3	Picture word	N			3	Sentence pictures
4	Picture word	Y	1		4	Sentence pictures
5	Picture word	Y	2		5	Simple cloze
6	Sentence pictures	Y	3		6	Simple cloze
7	Sentence pictures	N			7	Simple cloze
8	Sentence pictures	Y	4		8	Simple cloze
9	Sentence pictures	N			9	Simple reading
10	Simple cloze	Y	5		10	Simple reading
11	Simple cloze	Y	6		11	Simple reading
12	Simple cloze	N			12	Simple reading
13	Simple cloze	Y	7		13	Simple reading
14	Simple cloze	Y	8		14	Simple reading
15	Simple reading	Y	9		15	Intermediate cloze
16	Simple reading	Y	10		16	Intermediate cloze
17	Simple reading	Y	11		17	Intermediate cloze
18	Simple reading	Y	12		18	Intermediate cloze

19	Simple reading	Y	13		19	Intermediate cloze
20	Simple reading	Y	14		20	Intermediate reading
21	Intermediate cloze	Y	15		21	Intermediate reading
22	Intermediate cloze	Y	16		22	Intermediate reading
23	Intermediate cloze	Y	17		23	Intermediate reading
24	Intermediate cloze	Y	18		24	Intermediate reading
25	Intermediate cloze	Y	19		25	Intermediate reading

Appendix 2: Post-hoc test review protocol

Young Lives School Survey Assessment Review 2014

Please complete the following for each language and assessment. Note that all information will be used anonymously and kept securely.

Date of interview:	
Region/ Language:	
Assessment (maths or literacy):	
Teacher 1 name:	
Teacher 1 school:	
Teacher 1 grade:	
Teacher 1 specialisation (if appropriate):	
Teacher 2 name:	
Teacher 2 school:	
Teacher 2 grade:	
Teacher 2 specialisation (if appropriate):	

The below table provides some guidance notes about the kinds of insight needed to think about each question.

GUIDANCE NOTES FOR QUESTIONS 2-8	
2. What is the correct answer?	*If more than one question is correct circle all possible correct answers and provide details of why more than one option is correct in the comments section*
3. Rate the clarity of language used in specified parts of the test item **3.1 Passage/ text** (for literacy questions only) **3.2 Question** (for those maths questions with no words this will be NA) **3.3 Multiple Choice options**	*Think about whether the language used in the passage is clear or confusing. If language is confusing, please provide details in the comments section, such as 'the language used here would be interpreted as follows….X' or 'this is an uncommon way to express this'*
4. Rate the appropriateness of the mathematical notation/ expression (for maths questions only)	*If any notation is inappropriate or unfamiliar please provide details in the comments section. For example, 'this type of operation is more usually expressed like this'*
5. Rate the cultural appropriateness of the pictures (for some literacy questions only)	*If any pictures are not culturally appropriate please explain why in the comments section. For example 'this picture would be unfamiliar and would confuse children'*
6. Rate the cultural appropriateness of the passage and questions (for some literacy questions only)	*If the passage and/or questions are not culturally appropriate please explain why in the comments section.For example, 'the scenario in this passage is*

	unfamiliar or confusing'
7. In what grade would you expect to teach the competency assessed by this question? *Note that here we want to know in what grade the curriculum expects you to teach this topic.*	*If this question is not on the Ethiopian curriculum, or does not make sufficient sense to relate to a grade, please circle these options.*
8. In what grade would you expect children to be able to answer this item correctly? *Note that here we want to know in what grade children actually master this concept, which may differ to the grade in which you are expected to teach it.*	*If this question is not on the Ethiopian curriculum, or does not make sufficient sense to relate to a grade, please circle these options.*

1. Please discuss with teachers about the different methods they use for evaluating students' learning and progress in this subject. This might include true/ false quizzes, short answer tests, long answer tests where pupils have to write longer responses, multiple choice tests, or alternative methods. Please record the main methods used below and include a discussion of the appropriateness of a multiple choice test format for enabling an evaluation of pupils' best performance.

Main methods used for evaluating pupils learning and progress in this subject in grades 4 and 5:

Please answer each of the following questions in relation to every item in the assessment tool. Answer by circling the most appropriate response category(ies). Do this in relation to the Wave 1 test first, and then move on to the new questions included in Wave 2 at the end.

<table>
<tr><td colspan="4">WAVE: ____ QUESTION: ____</td></tr>
<tr><td colspan="2">2. What is the correct answer?</td><td colspan="2">A | B | C | D | None</td></tr>
<tr><td rowspan="9">3. Rate the clarity of language used in specified parts of the test item, using the following criteria:</td><td rowspan="3">3.1 Passage/ text
(for literacy questions only)</td><td>Excellent</td><td>Good</td></tr>
<tr><td>Reasonable</td><td>Poor</td></tr>
<tr><td>NA</td><td></td></tr>
<tr><td rowspan="3">3.2 Question
(for some maths questions this will be NA)</td><td>Excellent</td><td>Good</td></tr>
<tr><td>Reasonable</td><td>Poor</td></tr>
<tr><td>NA</td><td></td></tr>
<tr><td rowspan="3">3.3 Multiple Choice options</td><td>Excellent</td><td>Good</td></tr>
<tr><td>Reasonable</td><td>Poor</td></tr>
<tr><td>NA</td><td></td></tr>
<tr><td colspan="2" rowspan="3">4. Rate the appropriateness of the mathematical notation/ expression (for maths questions only)</td><td>Excellent</td><td>Good</td></tr>
<tr><td>Reasonable</td><td>Poor</td></tr>
<tr><td>NA</td><td></td></tr>
<tr><td colspan="2" rowspan="3">5. Rate the cultural appropriateness of the pictures
(for some literacy questions only)</td><td colspan="2">Very culturally appropriate</td></tr>
<tr><td colspan="2">Somewhat culturally appropriate</td></tr>
<tr><td colspan="2">Not culturally appropriate</td></tr>
<tr><td colspan="2" rowspan="3">6. Rate the cultural appropriateness of the passage and questions (for some literacy questions only)</td><td colspan="2">Very culturally appropriate</td></tr>
<tr><td colspan="2">Somewhat culturally appropriate</td></tr>
<tr><td colspan="2">Not culturally appropriate</td></tr>
<tr><td colspan="2" rowspan="3">7. In what grade would you expect to teach the competency assessed by this question? Note that here we want to know in what grade the curriculum <u>expects</u> you to teach this topic.</td><td colspan="2">Grade _____ (insert grade)</td></tr>
<tr><td colspan="2">Not on the curriculum</td></tr>
<tr><td colspan="2">Question does not make sense</td></tr>
<tr><td colspan="2" rowspan="3">8. In what grade would you expect children to be able to answer this item correctly? Note that here we want to know in what grade children actually master this concept, which may differ to the grade in which you are expected to teach it.</td><td colspan="2">Grade _____ (insert grade)</td></tr>
<tr><td colspan="2">Not on the curriculum</td></tr>
<tr><td colspan="2">Question does not make sense</td></tr>
<tr><td colspan="4">9. Please use this space to record explanations of and comments about the answers given above.</td></tr>
</table>

Appendix 3: Revised language-wise answer keys for the wave 1 and wave 2 reading comprehension tests following post-hoc review

Wave 1 item	Correct answer						
	Amharic	**Oromiffa**	**Tigrigna**	**Somali**	**Sidama**	**Hadiya**	**Wolayta**
1	B	B	B	B	B	B	B
2	C	C	C	C	A	C	C
3	B	B	B	B	B	B	B
4	B	B	B	B	B	B	B
5	C	C	C	C	B	C	C
6	C	C	C	C	B	C	C
7	A	A	A	A	DROP	A	A
8	C	C	C	C	B	C	C
9	B	B	DROP	B	DROP	DROP	B
10	D	D	D	A	C	D	D
11	B	B	B	A	B	B	B
12	A	A	A	A	C	A	A
13	C	C	C	A	A	C	DROP
14	D	D	D	A	B	D	D
15	C	C	C	C	B	C	C
16	A	A	A	A	D	A	A
17	C	C	C	A	C	C	C
18	B	B	B	B	B	B	B
19	B	B	B	B	B	B	B
20	DROP	A	A	A	A	A	A
21	B	B	B	A	D	B	DROP
22	DROP	B	B	B	A	B	C
23	A	A	A	D	A	A	A
24	DROP	A	A	C	B	A	DROP
25	C	C	C	C	B	C	B

Wave 2 item	Correct answer						
	Amharic	**Oromiffa**	**Tigrigna**	**Somali**	**Sidama**	**Hadiya**	**Wolayta**
1	B	B	B	B	B	B	B
2	C	C	C	C	B	C	C
3	C	C	C	C	B	C	C
4	C	C	C	C	B	C	C
5	D	D	D	A	C	D	D
6	B	B	B	A	B	B	B
7	C	C	C	A	A	C	DROP
8	D	D	D	A	B	D	D
9	C	C	C	C	B	C	C
10	A	A	A	A	D	A	A
11	C	C	C	A	C	C	C
12	B	B	B	B	B	B	B
13	B	B	B	B	B	B	B
14	DROP	A	A	A	A	A	A
15	B	B	B	A	D	B	DROP
16	DROP	B	B	B	A	B	C
17	A	A	A	D	A	A	A
18	DROP	A	A	C	B	A	DROP
19	C	C	C	C	B	C	B
20	A	A	A	A	A	A	A
21	A	A	A	A	A	A	A
22	D	DROP	B	B	DROP	B	B
23	DROP	DROP	C	D	A	DROP	C
24	A	B	B	DROP	DROP	B	B
25	C	DROP	C	C	C	C	C