YOUNG LIVES TECHNICAL NOTE **NO. 15**
January 2008

Psychometric characteristics of cognitive development and achievement instruments in Round 2 of Young Lives

Santiago Cueto
Juan Leon
Gabriela Guerrero
Ismael Muñoz

Young Lives
An International Study of Childhood Poverty

Contents

1. Introduction

Young Lives (YL) is a long-term international research project about childhood poverty based at the University of Oxford that integrates cutting edge research with local, national and international policy analysis engagement. The project seeks to a) improve understanding of the causes and consequences of childhood poverty and to examine how policies affect children's well-being and b) inform the development and implementation of future policies and practices that will reduce childhood poverty. YL is tracking the development of 12,000 children in Ethiopia, India (state of Andhra Pradesh only), Peru and Vietnam through quantitative and qualitative research over a 15-year period (2000-2015). YL has been following two groups of children per country since the beginning of the project: 2,000 children in each country who were born in 2000/01; and 1,000 children in each country who were born in 1994/95.

Up to now, the project has carried out two rounds of data collection. The first round took place in 2002 and the second round began in 2006 and was completed in 2007. The data collected in both rounds includes information on children's and their families' access to key services, work patterns and social relationships, as well as core economic indicators such as assets. There is also an assessment of children's nutritional and educational measures.[1]

Getting estimates of the cognitive abilities and achievements over time of children that are participating in the YL project is important as these variables may be considered both outcomes (proxy for the individual's skills) and predictors of later outcomes. For instance, a recent paper has established the association of cognitive abilities in early life and later outcomes in education, health and income.[2] With regard to achievement, administering and reporting results of standardised tests for students has become a common practice in recent years in many developing countries, as these are widely regarded as indicators of success in schooling and/or acquisition of basic skills or knowledge for adult life.[3] However, accurately measuring the cognitive development and achievement of the children in each cohort and getting meaningful scores may prove to be a complex endeavour.

In 2006, pilots of several cognitive development and achievement tests were carried out in each country prior to Round 2 of YL. As a result of these, it was decided to administer the following tests to the YL children: the Peabody Picture Vocabulary Test (PPVT) and the Cognitive Developmental Assessment (CDA). These were administered to assess children's verbal and quantitative ability respectively in the younger cohort (aged between 4.5 and 5.5 years old at the time of Round 2). The PPVT, plus two reading and writing items from Round 1 and a mathematics achievement test, were administered to assess children's verbal and quantitative abilities respectively in the older cohort (aged between 11.5 and 12.5 years old at the time of Round 2).

The main concern in administering and using the results of these tests is that their reliability and validity is established before using the data for research. This is because these tests for the most part were not developed for the specific contexts in which they were used in YL.

1 For further information about the project, please refer to the Young Lives presentation document available at www.younglives.org.uk.

2 See Grantham-McGregor et al. (2007).

3 See, for example, http://www.pisa.oecd.org for a description of the Programme for International Student Assessment, carried out periodically among OECD and several developing nations.

Hence, the main goal of the analysis presented in this paper is to establish the reliability and validity of each of the tests administered for each cohort within each country. In this process, the items with the best psychometric properties are different across countries. Therefore, we do not carry out international comparisons of results in this paper but suggest that the test results should be used for analysis within countries (even within countries, different language groups rely on different combinations of items to establish their ability). However, the general construct measured by each test is the same across countries. Hence, a comparison of the relationships between achievement and other variables across countries is a possibility.

In order to get indicators of reliability and validity, we used current standards on psychometrics to guide the analyses.[4] The psychometric characteristics of each test were estimated through several methods. The reliability analysis was developed to see how consistent the scores are for the children. In other words, the reliability index tells us how accurate and stable the scores are. We used both Classical Test Theory (CTT) and Item Response Theory (IRT) methods to estimate reliability indicators. The validity analysis had the objective of evaluating the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests (in this case research). To address this analysis, we estimated the correlation between each test score and some variables such as age and educational level to check if they were supported by previous empirical evidence reported in the literature (i.e. on average, children in higher grades of school should get higher results than children in lower grades or out of school, and parental education should be positively correlated with scores on the tests).

The paper is organised into six sections, including this introduction. The second section presents the framework for the analysis, introducing definitions and theories of validity and reliability. Section three includes a description of each of the cognitive development and achievement instruments used in the YL project. The fourth section presents information on the data and methods used in this paper, and the results on the psychometric characteristics of the tests are presented in section five. Finally, in section six, some considerations about the use of the test scores are presented, as well as recommendations for the assessment of cognitive development and achievement in Round 3 of the YL project.

2. Validity and reliability of cognitive development and achievement instruments

As mentioned before, the main objective of the present study is to analyse the psychometric properties of each of the instruments administered for each cohort within each country, i.e. to establish the validity and reliability of the instruments.

The tests used in YL to measure cognitive development and achievement are norm-referenced. When scores are norm-referenced, relative score interpretations are of primary interest. A score for an individual is ranked within one or more distributions of scores or

4 Mainly the *Standards for Educational and Psychological Testing* published in 1999 by the American Educational Research Association, American Psychological Association and National Council on Measurement in Education.

compared to the average performance of test takers for several reference populations (e.g., based on age, grade, etc.) In contrast, when scores are criterion-referenced, absolute score interpretations are of primary interest. The test score directly conveys a level of competence in some defined criterion domain (AERA, APA and NCME 1999). Although criterion-reference assessments are interesting and useful in education, the primary interest regarding the cognitive development and achievement assessment in YL was to maximise the true variance between individuals, so that their ability characteristics could be correlated accurately with other individual, family and social background variables.

This section of the paper contains information on current standards of psychometrics regarding validity, reliability and fairness in testing. The information presented here is based on the Standards for Educational and Psychological Testing published in 1999 by the American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME).

2.1 Validity

According to AERA, APA and NCME (1999), 'validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations.' (p. 9).

In the past, validity theory considered there were different types of validity, e.g. content, construct, predictive, etc. Nowadays the theory has evolved and defined validity as a unitary concept. There are various sources of evidence that may shed some light on different aspects of validity, but they do not constitute different types of validity.

The main sources of evidence that might be used to evaluate the validity of an instrument (AERA, APA and NCME 1999) are:

- evidence based on test content, which is an analysis of the relationship between a test's content and the construct it is intended to measure
- evidence based on response processes, which requires theoretical and empirical analyses of the response processes of test takers in order to provide evidence of the fit between the construct and the nature of performance given by examinees
- evidence based on internal structure, which is an analysis of the degree to which the relationships between test items and test components conform to the construct assessed by the test
- evidence based on relations to other variables external to the test (such as the scores of other tests measuring the same construct or group membership variables), which requires an analysis of the degree to which these relationships are consistent with the construct underlying the test
- evidence based on consequences of testing, which proposes the incorporation of the intended and unintended consequences of test use into the concept of validity

It is important to notice that a test cannot be qualified as valid in absolute terms, i.e. valid for all purposes or in all situations or populations. Each use or interpretation of the test requires validation. This is why the objective of this paper is to establish the validity and reliability of the PPVT, CDA and the mathematics achievement test *for the YL sample and purposes*, even though these instruments' psychometric properties have been established previously by other studies. In general terms the results of these tests will be used for research purposes, not other uses such as individual diagnoses or recommendations for treatment.

According to this recent understanding of validity, validation is the joint responsibility of the test developer and the test user. While the test developer is responsible for providing relevant evidence and a rationale in support of the intended test use, the test user is ultimately responsible for evaluating the evidence in the particular context in which the test is to be used (AERA, APA and NCME 1999).

2.2 Reliability

According to AERA, APA and NCME (1999), reliability refers to the consistency of an instrument when the testing procedure is repeated on a population of individuals or groups. The assumption behind this definition is that individuals and groups exhibit some degree of stability in their behaviour. Nevertheless, successive samples of behaviour from the same person are rarely identical in all relevant aspects. Because of this variation, an individual obtained score will always reflect at least a small amount of measurement error.

Errors of measurement are generally viewed as random and unpredictable. They are distinguished from systematic errors which affect the performance of individuals in a consistent way. The sources of error can be broadly categorised as those internal to the examinees and those external to them. Fluctuations in the level of an examinees' motivation and interest or attention are examples of internal factors that may lead to score inconsistencies. Differences among testing sites in their freedom from distractions, random effect of scorer subjectivity, and variation in scorer standards are examples of external factors. Random measurement errors cannot be removed from observed scores. However, their aggregate magnitude can be summarised in several ways (AERA, APA and NCME 1999).

There are two main theories that deal with the issue of reliability and measurement errors: Classical Test Theory and Item Response Theory. Each has different assumptions that are explained briefly in the following sub-sections.

Classical Test Theory

In Classical Test Theory (CTT), the observed score of the test-taker has two components: the true score and the error of measurement as shown in the following equation:

$X = X_t + E$

X: The observed score
X_t: The true score
E: Error of measurement

According to this theory, the measurement error (E) has three main sources: the item itself (e.g., construction of the item); the person taking the test (e.g., motivation, health, etc.); and the environment or context (e.g., adequacy of the place where the test is administered). The main idea behind this theory is to minimise this error of measurement in order to have a more precise estimate of the true score of the test-taker. For this reason, CTT uses three indicators to analyse and improve the psychometric qualities of the items within a test: *difficulty* (p value of each item), *discrimination* (difference in an item between low and high scorers in the test), and *non-response rate* (an assessment of the understanding of the items). Additionally, there are two overall indicators of the quality of the test: reliability (inter-item correlations) and the standard error of measurement (SEM).

Difficulty index is the proportion of people who answer the item correctly. Thus, the difficulty index is calculated as follows:

$p_i = c_i / N$

p_i: proportion of children who answer item i correctly
c_i: number of children who answer item i correctly
N: number of children evaluated

This index ranges from 0 to 1. Values close to 0 mean only a few people answered the item correctly; values close to 1 means the item was answered correctly by most of the individuals. Hence, if the purpose of a test is to have a wide variety in total scores, items with values close to 0 or 1 have to be reviewed or may as well be eliminated, since they provide relatively little information for discriminating between test-takers. Finally, the value of the difficulty index can be classified into five categories, the first and the fifth being the ones that require special attention (Crocker and Algina 1986). The categories are: *extremely easy (.75 – 1), easy (.55 - .74), moderate (.45 - .54), difficult (.25 - .44) and extremely difficult (0 - .24)*.

The *discrimination index* is used to estimate the extent to which an item helps to discriminate between people with high and low performance in a given test. A common formulation for estimating discrimination is this index:

$D_i = P_{i(H)} – P_{i(L)}$

D_i: discrimination index for item i
$P_{i(H)}$: proportion of children in the higher tercile on the total score for the test who answer item i correctly
$P_{i(L)}$: proportion of children in the lower tercile on the total score for the test who answer item i correctly

This index ranges from -1 to 1, where positive values mean that the item is discriminating in favour of the high achievers and negative values means that the item is discriminating against high achievers. According to the value of the index, the discrimination power of any item can be categorised as follow: *extremely high (.40 – 1), high (.30 - .39), moderate (.20 - .29), low (0 - .19) and to discard (< 0)*. The items that present problems are those located in the last two categories (low or to discard) (Matlock-Hetzel 1997; Crocker and Algina 1986).

The *non-response rate* is the proportion of people who do not answer the item. Thus the non-response rate is:

$nr_i = 1 – p_i - q_i$

nr_i: proportion of children who do not answer the item i
p_i: proportion of children who answer the item i correctly
q_i: proportion of children who answer the item i incorrectly

According to the percentage of people who did not answer the item, the non-response rate can be categorised as follows: *adequate (0 - .15), acceptable (.16 - .20), tolerable (.21 - .29) and to discard (.30 – 1)*. In this way, items with non-response rates above .30 have to be discarded or reviewed because most of the examinees may have found the item problematic (e.g. not understandable or too difficult) (Matlock-Hetzel 1997; Oosterhof 1990; Crocker and Algina 1986).

Reliability index is an overall indicator of the consistency of the latent construct being measured, or the degree to which an instrument produces the same measurements when the measuring procedure is repeated several times. There are different methods of estimating the reliability index, the best known being test-retest and internal consistency. Thus, the formula for the reliability index is:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_x^2}\right]$$

α = reliability index for the test
k = number of items in the test
σ^2_i = the variance of the item i
σ^2_x = the variance of the test

This index or indicator ranges from 0 to 1. Values close to 0 mean that there is a low correlation among the items used in the test and values close to 1 mean that the items used are measuring the same latent construct.[5] According to the literature, values above .60 indicate an acceptable reliability for research purposes.

Standard error of measurement (SEM) is an overall indicator of the quality of the test. SEM indicates the distribution of the scores around the true score and the formula to estimate this is:

$SEM = (1 - \alpha_{tt})^{1/2}\sigma t$

α_{tt} = reliability index of the test
σ_t = standard deviation of the scores in the test

As with the previous formula, the SEM is constant for every level of ability or score obtained by the children. The amount of error determines the reliability of a test score: the higher the error, the lower the reliability (Koretz 2008: 145).

The difficulty, discrimination and non-response indicators briefly explained above provide indications of the adequacy of the items in a given test and allow us to identify the characteristics of the item. Depending on the results, some items will be eliminated from the database (see criteria below). The results of the overall reliability analyses are presented in section five of this report and speak to the accuracy and utility of the test as a whole.

Item Response Theory

Item Response Theory (IRT), in contrast with CTT, is more focused on the item than the test level. For the purposes of this paper, we have used a special case of IRT for the reliability analyses, called the one-parameter model or Rasch Model.

The Rasch analysis is centred on estimating the probability of success of the test-taker on a specific item. The model estimates the probability of answering the item correctly as a logistic function of the difference between the individual's ability and the item difficulty (see equation below). Thus, the Rasch model allows us to create an interval scale of scores for both the item's difficulty and individual's ability, and these scores are scaled in logits (Baker 2004).

5 This formula is used to calculate Cronbach's alpha.

$$p_i(\theta) = \frac{1}{1 + e^{-(\theta - b_i)}}$$

$P_i(\theta)$: is the probability to answer item i correctly
θ: is the test-taker ability
b_i: is the item difficulty index of the item i
e: is the exponential function

The use of Rasch analysis contributes to the normalisation of test curves.

Just as CTT has four indicators to assess the adequacy of the items, Rasch Models have statistics to evaluate the fit of the item into the model. The idea underlying these statistics is that correct answers in more difficult items are accomplished by people with higher ability. At the same time, these people will have a greater probability of attaining higher scores on easier items than on more difficult ones.

Rasch models use the mean square (MNSQ) fit statistics to identify item and person ratings that deviate from expectations. The MNSQ fit statistics value is the ratio of the observed variance (variance attributable to the observed variable) and the expected variance (variance estimated by the 1PL model). A ratio of 1 indicates that the observed variance equals expected variance; ergo, the error of measurement is almost zero or nonexistent. When the MNSQ fit statistics value is greater than 1.0, there is more variation in the observed variable than the 1PL model predicted. For example, 1.70 indicates 70 per cent more variation. When the fit statistics value is less than 1.0, there is less variation in the observed variable than the 1PL model predicted.

There are two types of MNSQ fit statistics that have to be taken into account at the moment of analysing the item fit: the *outfit* and the *infit* statistics.

Outfit mean square is a chi square statistic that measures the unexpected observations on items that are very easy or very hard for a given individual. In other words, this statistic reflects the fact that there are individuals with higher ability giving incorrect answers to items that should have been easy for them, and vice-versa. Problems with this indicator should be considered, although they do not represent a serious threat to the reliability of the test (Linacre 2002).

Infit mean square is a chi square statistic that measures unexpected patterns of observations by persons on items that are roughly targeted at them. In other words, this statistic evaluates how well the observations fit the IRT model or the size of the residuals in the estimated model. Problems with this indicator indicate a threat to the reliability and also to the validity of the test (Linacre 2002).

These statistics can be categorised as follows according the value of the *infit* or *outfit*: *off-variable noise is greater than useful information (>2), noticeable off-variable noise (1.5 – 1.99), productive of measurement (.5 – 1.5) and overly predictable (<.5)* (Linacre 2008). In sum, *infit* and *outfit* values between .5 and 1.5 indicate that there is a good fit of the items.

The statistics briefly explained here provide an indication of the fit of the item into the model. Additionally, Rasch Models consider two indicators of the overall precision of the scores estimated. The first indicator is the *reliability person index* and the second one is the *standard error of measurement* (SEM).

Reliability person index is an indicator of the proportion of the sample variance that is not due to measurement error or the ratio between true variance and observed variance, where true variance is the difference between observed and error variance.

$\alpha_{tt} = (\sigma^2_o - \sigma^2_E)/ \sigma^2_o$.

σ_o = Standard deviation of the children measures
σ_E = Average of the standard measurement error for each child

This index ranges from 0 to 1 and values above 0.5 indicate there is an adequate estimation of the children's ability (Linacre 2008).

SEM in the Rasch Model is the root square of the inverse of the Test Information Function and it can be expressed using the following formula: [6]

$SEM = 1 / (I(\theta))^{1/2}$

In contrast to CTT, SEM in the Rasch Model is a function of the ability of the children, i.e. the SEM is not constant and varies at each level of ability (Hambleton et al. 1991). Nevertheless, the interpretation of the indicator is the same as in CTT: the higher the error, the lower the reliability. In this report, we present the average SEM across children as a measure of overall precision of the Rasch scores calculated by the model. The results of these reliability analyses are presented in section five of this report.

2.3 Fairness in testing and test use

There are different ways of conceptualising fairness in testing according to the *Standards* developed by AERA, APA and NCME (1999). The Standards recognise at least four main ways in which the term fairness is used. The first two characterisations relate fairness to the absence of bias and to equitable treatment of all examinees in the testing process. The third characterisation of test fairness addresses the equality of testing outcomes for examinee subgroups defined by race, ethnicity, gender, disability and other characteristics. However, the idea that fairness requires equality in overall passing rates for different subgroups has been discarded by most testing professionals. Finally, the fourth definition of fairness refers to equity in opportunity to learn the material covered in an achievement test. In what follows, we will briefly describe the first two definitions of fairness as well as their implications for the analyses presented in this paper. We are focusing only on these two definitions because there is a relative consensus in the testing literature that when a test is free of bias and the examinees have received fair treatment in the testing process, then the conditions of fairness have been met.

In relation to the first definition of fairness, bias arises 'when deficiencies in a test itself or the manner in which it is used result in different meanings for scores earned by members of different identifiable subgroups (AERA, APA and NCME p. 74). In other words, bias refers to construct-irrelevant components that result in systematically lower or higher scores for identifiable groups of examinees. When evidence of such deficiencies is found at the level of item response patterns for member of different groups, the term *differential item functioning* (DIF) is used (AERA, APA and NCME 1999). For the purposes of this paper, we have used statistical procedures for identifying items on the tests that function differently across subgroups of examinees based on gender and language. The objective of the analysis was to check if examinees with similar overall ability differ on average in their responses to any particular item. The results of these analyses as well as the recommendations to deal with gender and language biases are presented in section five of this report. Specifically to estimate DIF we used the Mantel-Haenszel method, using the total raw score as the matching criterion.

6 The Test Information Function is the addition or sum of the *n* item Information Functions in the test. This item Information Function is the probability of getting the item right multiplied by the probability of getting it wrong. Thus, the formula is: $I(\theta_i) = p_i(\theta)q_i(\theta)$.

The second definition of fairness that we will discuss in this paper emphasises fairness as equitable treatment in the testing process. According to this view, 'fairness requires that all examinees be given a comparable opportunity to demonstrate their standing on the construct the test is intended to measure. Just treatment also includes such factors as appropriate testing conditions and equal opportunity to become familiar with the test format, practice materials, and so forth' (AERA, APA and NCME 1999: 75).

In order to guarantee fairness in the testing process we took into account the following considerations. First, during the pilot process we sought to establish the cultural relevance of the test administered, given that experts in each country were asked to verify the relevance of each item and then translate it. Second, we carefully standardised the administration procedures to provide all examinees with appropriate testing conditions and also with the same opportunity to demonstrate their ability. We asked the examiners to evaluate the appropriateness of testing conditions and to report that information in the questionnaires (all tests were administered individually). During the process of cleaning the databases, we have considered as invalid all the observations where examiners reported inadequate conditions that seriously compromise the performance of the child.

Finally, it is important to remember here that, given some of the above mentioned considerations, we do not carry out international comparisons of results in this paper but suggest that the test results should be used for analysis within countries (and even within countries, it may be that different language groups require different combinations of items to establish their ability, as the bias analysis described above may suggest). However, as stated before, given that the construct measured by each test is the same across countries, comparisons of associations between the abilities measured and other variables across countries could be established.

3. Measuring cognitive development and achievement in Young Lives

3.1 Round 1

To establish the background of cognitive development and achievement measurements, reading, writing and numeracy assessment items were administered to the children from the older cohort in Round 1 of YL.

- The reading item required children to read three letters ('T, A, H'), one word ('hat'), and one sentence ('the sun is hot').[7] The reading item was scored as follows: 1 point if children could read the sentence, 0.66 point if they could read the word, 0.33 point if they could read the letters and 0 points if they could not read anything or did not respond.

7 These reading items were used in India, Ethiopia and Vietnam. In the case of Peru, the reading item required the children to read the letters 'N, A, P'; the word '*pan*' (which is Spanish for 'bread'); and the sentence '*El pan es rico*' (which is Spanish for 'the bread is tasty').

- The writing item asked the children to write a simple sentence ('I like dogs') which was spoken out loud by the examiner. This item was scored as follows: 1 point if children could write the sentence without difficulty or errors, 0.5 point if they could write with difficulty or errors, and 0 points if they could not write anything or did not respond.
- The numeracy (maths) item required children to solve a basic multiplication (2 x 4). The answer given by children was scored one point if children answered correctly and 0 points if they answered incorrectly or did not respond.

During the planning of Round 2, it was decided to keep these items to ensure comparability of information, even though it was expected that they would be answered by most respondents given their ages and school experience. The numeracy assessment item was included in the Maths Achievement Test that we will describe below, and the other items were included in the child questionnaire.

In Round 1, the Raven's Colored Progressive Matrices (CPM) Test was administered to the older cohort to test intellectual abilities. The test requires that the child completes a pattern of figures so that they make sense. The test has three subscales, each with 12 items. Subscales A and B measure aspects related to the cognitive processes of the children while subscale AB measures the intellectual capacity of the children.[8] The test is supposed to be relatively free of cultural bias; however this statement has been challenged by some studies (Court 1983). The score was the number of correct responses to the items. The administration of this test proved to be difficult (especially in rural areas in Ethiopia). Field workers reported difficulties in explaining the tasks and also excessive time taken to administer the test to many children. Given these observations, plus discouraging results of pilot tests for Round 2 in Peru, it was decided to drop this test from YL for Round 2, although the Children or Adult version of the Raven could be administered again in future rounds of YL.

3.2 Round 2

In 2006, pilot tests of several cognitive development and achievement tests were carried out in each country with the purpose of studying the reliability, validity and feasibility of administering some of these tests during the second round of YL, planned for the last trimester of 2006. Children participating in these pilot studies were about the same age as children in YL would have been when participating in Round 2 data collection but were chosen in areas different from those where YL is working (all were children not included in the YL sample). The instruments used in the pilot test included the tests that were administered during the first round of YL, plus some new assessments that were considered for the second round.[9]

Based on the pilot tests results, the previous experience from Round 1, and the international evidence on cognitive development and achievement assessment, it was decided to administer the tests listed in Table 1 in the second round.

8 For further information on the Raven's CPM, please refer to the publisher's web page at: http://harcourtassessment.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=015-4686-743&Mode=summary

9 The full country reports for the pilot tests specifying the sample, the instruments administered and the results are available from Santiago Cueto (scueto@grade.org.pe).

Table 1. *Cognitive abilities and achievement instruments administered in Round 2 of YL*

	Younger cohort	Older cohort
Verbal	Peabody Picture Vocabulary Test (PPVT)	PPVT and reading and writing items from Round 1
Quantitative	Cognitive Developmental Assessment (CDA)	Mathematics achievement test (including the maths item from Round 1)

Peabody Picture Vocabulary Test (PPVT)

The PPVT is a widely-used test of receptive vocabulary. Its main objective is to measure vocabulary acquisition in persons from 2.5 years old to adulthood. The test is individually administered, un-timed, norm-referenced and orally administered. The task of the test taker is to select the picture that best represents the meaning of a stimulus word presented orally by the examiner.

The PPVT was originally developed in 1959 by Dunn and Dunn. Since then, it has been updated and improved several times (PPVT-R 1981; PPVT-III 1997; and PPVT-IV 2007). PPVT's validity and possible cultural bias have been studied repeatedly during the past decades not only by the authors but also by other researchers. Several studies have found that both the PPVT-R and the PPVT-III show a positive strong correlation with some commonly-used intelligence measures, such as the Wechsler and the McCarthy Scales (Campbell et al. 2001; Gray et al. 1999; Campbell 1998). Regarding the existence of a cultural bias, evidence is not conclusive since mixed results have been reported in the literature. On the one hand, Williams and Wang (1997) and Washington and Craig (1999) found no item bias in the PPVT using a sample of African American pre-schoolers. Similar results were found by Restrepo et al. 2006). On the other hand, studies developed by Ukrainetz (2000 and 2002), Stockman (2000) and Champion et al (2003), also in African American pre-school children, found there was item bias. Although the test was originally developed for English speakers, a Spanish version has been developed and normed.

In the second round of the YL study, the PPVT-III (Dunn and Dunn 1997) was used to evaluate both cohorts in India, Ethiopia and Vietnam; however, in Peru, the Spanish version of the PPVT-R (Dunn et. al. 1986) was used to evaluate both cohorts.[10] In order to use the PPVT, all the necessary copies for each country were bought directly from the publisher.

Even though both versions used to evaluate the children participating in YL measure the same construct and follow the same principles, there are some differences between them, especially in relation to the number of items and how they are arranged. The PPVT-R (Hispanic version) has only one form containing 125 items. The PPVT-III is available in two parallel forms designated as Form III-A and Form III-B. Each form contains 204 items grouped into 17 sets of 12 items each. The form used in YL was Form III-A.

In both versions of the PPVT used in the second round of the YL study, the items are arranged in order of increasing difficulty. Not all the items in the test are administered to a given child but only those within his or her *critical range*. The examiner must select the appropriate Start Item according to the child's chronological age and continue administering the test until the child reaches a ceiling, i.e., those items extremely easy or extremely hard for

10 For further information on the PPVT, please refer to the PPVT-III Examiner's Manual (Dunn and Dunn, 1997) and the PPVT-R Examiner's Manual (Dunn et. al., 1986).

the child are not administered but only those within his or her *critical range*. This requires that the examiner correctly establishes the Basal Item Set and Ceiling Item Set for the individual. The PPVT has basal and ceiling rules and the basal set must always be established first. In the case of the PPVT-R, the basal is formed by the highest eight consecutive correct responses and the ceiling is formed by the lowest eight consecutive responses containing six errors. In the case of the PPVT-III, the basal set rule is one or no errors in a set of 12 items and the ceiling set rule is eight or more errors in a set.

The PPVT offers raw scores as well as standard scores. Raw scores are calculated by subtracting the total number of errors from the ceiling item. In the case of the PPVT-R, the examiner must only count the errors between the highest basal and the lowest ceiling. In the case of the PPVT-III, the examiner must count the total number of errors made by the examinee from the basal set through the ceiling set. The test manual includes tables for the conversion of the raw scores into standard scores. However, we have not used standard scores for Ethiopia, India or Vietnam here because the standardisation samples of both PPVTs used in YL have different characteristics from the project's sample. The validity and reliability analyses presented in section 5 were done using the raw scores in all cases.

The PPVT was administered to both cohorts of children. The test was translated into each country's main languages by the local team and verified by a local expert before the pilot study conducted prior to the second round of data collection.[11]

Cognitive Developmental Assessment (CDA)

The CDA was developed by the International Evaluation Association (IEA) during the second phase of the Pre-Primary Project in order to assess the effect of attending a pre-school centre in the cognitive development of 4 year old children. IEA granted YL study authorisation to use this instrument.

The test has several subtests:[12]

- Spatial relations: This area consists of two sections. The first requires the child to perform an action in response to the test item. Spatial notions such as: *on, under, behind, in front* or *beside* are assessed by indications like: '*Put the toy on the chair*'. The second section requires the child to indicate which one of a set of pictures fits the description provided. Concepts like *through*, *around, between, up, toward,* etc. are evaluated. An example of this type of question is: '*Point to the jar that is between the spoons'*.
- Quantity: This subscale requires children to indicate which one of a set of pictures fits the description provided by the examiner. Notions such as *a few, most, half, many, equal, a pair,* etc. are assessed with statements such as: *'Point to the plate that has a few cupcakes'*.
- Time: This subscale requires the child to point out which of a set of pictures best represents the concept provided by the examiner. Measures in this area involve knowing what *day of the week* it is, if its *night*, *morning* or being aware of such concepts as *before*, *starting*, etc. Similar indications to the ones presented above are used in this subtest, for example: *'Show me which child has finished drinking'.*

11 The test pictures remained the same.

12 Further information on the Pre-Primary project and the CDA is available in: Preprimary Project (2005) Draft III. *Sampling, Instrumentation and Data Collection.*

For the second round of YL it was decided to administer only the quantitative sub-scale of the CDA, since the pilot studies showed that the reliability of the time sub-scale was low and the spatial relations subscale took too long to administer.

In the quantity subscale, the task was for children to pick an image from a selection of three or four that best reflected the concept verbalised by the examiner (e.g. few, most, nothing, etc.) This subscale had fifteen items and all had to be administered to the child. Each correct answer was scored 1 point, with 0 points for wrong answers or no response, amounting to a maximum total score of 15 on the CDA quantity subscale.

The CDA was administered only to the younger cohort. The test was translated into each country's main languages by the local team and verified by a local expert before the pilot study conducted prior to the second round of data collection.

Achievement Test

1) Verbal

The reading and writing items from Round 1 (described in the previous section) were also administered in Round 2 of YL to the children from the older cohort.

2) Mathematics

The maths test used for the second round of YL had 10 items, scored 1 for correct and 0 for blank or incorrect. Most of the items included in the test were selected from the publicly released items of the Trends in International Mathematics and Science Study developed by the IEA in 2003.[13] The items were originally developed to assess fourth and eighth graders. Items with different difficulty levels were selected in order to discriminate between higher and lower achievers. In addition to these items, the numeracy item from Round 1 (solving a basic multiplication task) was also included.[14]

The items were selected for the topics of number and number sense only. This was due to the fact that other topics (i.e. geometry, measurement, data and algebra) might not be covered in depth by students in school and it might be unfair to include them in the evaluation of non-schooled children or dropouts. Number items, however, are directly related to basic skills necessary in any modern society. The format of the items was either multiple choice or short answer.

The mathematics achievement test was administered only to the older cohort. The test was translated into each country's main languages by the local team and verified by a local expert before the pilot study conducted prior to the second round of data collection.

It is important to notice that the mathematics achievement test used for the pilot studies had 27 items: the single item from Round 1; four additional items with no text, just numbers, requesting children to add, subtract and divide figures of up to three digits; and 22 items taken from TIMSS. The items were arranged in two booklets for the pilot study. They included the same items but presented them in a different order, in order to test for the effect of this (the difficulty of the items was established using the international results from TIMSS). The first booklet had the items randomly organized, while the second booklet had the items organised according their level of difficulty. The pilot results showed the second booklet had higher reliability and took less time for the children to complete. The number of items

13 Further information about TIMSS and the released items is available at: http://timss.bc.edu/timss2003i/released.html.

14 The only exception is India where the original item 2 times 4 was changed in round two for another basic multiplication (2 times 7).

included in the test was then reduced to 10, taking into account the variance in item difficulty and in exercise type.

4. Data and Methods

4.1 Sample

YL has been following two groups of children since the beginning of the project: 2,000 children in each country who were born in 2000/01, and 1,000 children in each country who were born in 1994/95.[15] Table 2 summarises the sample size for each country and cohort available for the analysis presented below.[16]

Table 2: *Number of children by country and cohort of study*

	Younger cohort	Older cohort
Ethiopia	1912	979
India (state of Andhra Pradesh only)	1940	993
Peru	1963	685
Vietnam	1970	990

Tables 3 and 4 present the characteristics of the sample by country for the younger and older cohort respectively. Children in the younger cohort sample are five years old on average. In all the countries, the number of boys and girls included in the sample is balanced. The sample is predominantly rural in India, Ethiopia and Vietnam; however, in the case of Peru the sample is more balanced between urban and rural children. Every country's sample includes children who speak a minority language but the percentage is higher in Ethiopia, where more than 50 per cent of the sample speaks a minority language. Regarding household size, the average number of members per household in India, Ethiopia and Peru is around six; in Vietnam the households seem to be smaller (4.7). Finally, parents' educational attainment is low in the four countries but especially in Ethiopia and India, and less so in Peru. Additionally, the educational level of mothers is lower than the educational level of fathers.

Table 3: *Sample characteristics by country, younger cohort (Standard Deviation)*

	Ethiopia	India	Peru	Vietnam
Mean age of children (years)	5.2	5.4	5.3	5.3
	(0.32)	(0.32)	(0.39)	(0.31)
Percentage female	47.1	46.8	49.0	48.8
	(49.93)	(49.91)	(50.00)	(50.00)
Percentage of children who speak a minority language	55.8	7.8	11.4	22.5
	(49.68)	(26.90)	(31.74)	(41.76)
Percentage of children who live in an urban area	39.3	25.6	55.3	20.6
	(48.85)	(43.65)	(49.73)	(40.48)
Mean number of members in the household	6.0	5.5	5.5	4.7
	(2.08)	(2.22)	(2.08)	(1.51)
Percentage of mothers with complete secondary education or more	5.6	6.1	37.0	13.6
	(23.09)	(23.85)	(48.29)	(34.24)
Percentage of fathers with complete secondary education or more	12.2	15.0	47.0	18.8
	(32.78)	(35.69)	(49.92)	(39.11)

15 With the exception of Peru, where the original number of children for the older cohort was 714.

16 For a description of sampling procedures in each country see http://www.younglives.org.uk/countries.

Children in the sample for the older cohort are 12 years old on average. In this cohort, the sample is also balanced between girls and boys. Most of the children in the sample come from rural areas, except in the case of Peru where more than 70 per cent of the children live in an urban area. On average, children from India, Peru and Vietnam were attending the sixth grade of basic education at the time of data collection; however, children from Ethiopia were mostly in fourth grade. The percentage of children who speak a minority language is low in all the countries, with the exception of Ethiopia where the percentage is over 50 per cent. Educational attainment is also low among parents from the older cohort, especially in Ethiopia and India.

Table 4. *Sample characteristics by country, older cohort (Standard Deviation)*

	Ethiopia	India	Peru	Vietnam
Mean age of children (years)	12.1	12.4	12.3	12.3
	(0.32)	(0.36)	(0.52)	(0.33)
Percentage female	49.4	52.0	46.1	49.9
	(50.02)	(49.99)	(49.89)	(50.03)
Mean last grade attended in school	4.2	6.7	5.9	6.6
	(1.64)	(1.02)	(1.07)	(0.77)
Percentage of children who speak a minority language	54.0	4.2	6.3	17.2
	(49.87)	(20.14)	(24.27)	(37.73)
Percentage of children who live in an urban area	40.4	25.1	73.3	20.6
	(49.10)	(43.38)	(337.48)	(40.48)
Mean number of members in the household	6.5	5.2	5.6	4.9
	(2.05)	(1.83)	(1.99)	(1.38)
Percentage of mothers with complete secondary education or more	4.5	4.0	33.0	13.7
	(20.72)	(19.67)	(47.05)	(34.44)
Percentage of fathers with complete secondary education or more	8.5	11.0	46.3	19.4
	(27.86)	(31.28)	(49.90)	(39.56)

4.2 Data collection

The second round of data collection was begun late 2006 and completed early 2007. A considerable effort was made in all the countries to develop standardised survey operations procedures. All the procedures to be followed by the examiners during the fieldwork were established in the *Examiner's Manual*. This manual had a specific section regarding the administration of cognitive development and achievement instruments where the guidelines for the administration and scoring of the three instruments used in the second round of data collection were presented.

All the fieldworkers who participated in the study had experience in the administration of surveys. They were experienced fieldworkers from the first round of the YL or other similar projects. They were trained before the starting of the field work on the specific procedures of the administration and scoring of these tests by the specialists within each country research team. Several role-playings were carried out during training sessions in order to establish the correct procedures to administer the tests.

Throughout the training process, the importance of standardising the administration procedures in order to provide all examinees with appropriate testing conditions was emphasised. Fieldworkers were instructed to administer the test to the children in their preferred language. For that purpose, every test used in Round 2 was carefully translated into each country's main languages by the local team and verified by a local expert. In case the examiner did not speak the preferred language of the examinee, the use of an interpreter was authorised. The interpreter was asked to use the translated version of the test provided by the examiner and follow his or her directions carefully.

Because of the longitudinal nature of the YL project, children were evaluated individually at home in both rounds. Fieldworkers were instructed to find an adequate space to administer the tests within the house. They were asked to evaluate the appropriateness of testing conditions and report that information in the questionnaires. Tables 5 and 6 summarise the cases of inadequate conditions as reported by fieldworkers during the administration of the three tests by country and cohort. The two inadequate testing conditions reported in the following two tables (i.e. insufficient light and inadequate vision or hearing) compromise the measurement of the child's abilities and therefore those observations were considered invalid for the subsequent analyses. As shown in the tables, the number of observations invalidated was small. It should be noticed however that fieldworkers also reported other inadequate conditions, such as not having a flat surface to work on during the examination or being in an environment with too much noise or too many distractions. However, we decided not to consider the tests taken under these conditions as invalid, since these conditions are very common in the poor urban and rural contexts where most of the Young Lives children live and in which they are therefore used to working in.

Table 5: *Inadequate conditions during the administration of the PPVT according to examiners' report*

Country:	**India**	**India**	**Ethiopia**	**Ethiopia**	**Vietnam**	**Vietnam**	**Peru**	**Peru**
Cohort age:	**5**	**12**	**5**	**12**	**5**	**12**	**5**	**12**
Total	23	1	4	0	3	2	2	1
Insufficient light	3	0	0	0	2	2	1	1
Child showed evidence of uncorrected vision or hearing loss	20	1	4	0	1	0	1	0

Table 6: *Inadequate conditions during the administration of the CDA and the mathematics achievement test according to examiners' report*

	CDA (younger cohort)				**Maths (older cohort)**			
	India	**Ethiopia**	**Vietnam**	**Peru**	**India**	**Ethiopia**	**Vietnam**	**Peru**
Total	20	24	3	1	3	1	1	3
Insufficient light	2	5	2	1	0	0	1	3
Child showed evidence of uncorrected vision or hearing loss	18	19	1	0	3	1	0	0

In addition to assessing and reporting testing conditions, examiners were also asked to register the time taken by the child to complete each test administered. Table 7 contains information regarding the administration time for each test by country and cohort. The administration time for the CDA was lower in Peru, although we are not sure why. Given that Peruvian children were more likely to come from urban areas than children in the other countries, we estimated the difference in time between Peruvian urban and rural children, but found no significant differences. The fieldworkers reported that children found this test easy and fun to take, and that it was similar to others they had taken in preschool or school.

Regarding the administration time of the PPVT, it is important to bear in mind that the version of the test used in Peru (PPVT-R in Spanish) was different from the version used in the other three countries (PPVT-III). This version was shorter (125 items in total compared to 204 items in the PPVT-III) and could explain the relatively shorter time for test administration in Peru.

Table 7: *Administration time in minutes (median and standard deviation) by instrument and country**

	Ethiopia	**India**	**Peru**	**Vietnam**
CDA	15	10	5	10
	(9,61)	(6,60)	(3,82)	(6,59)
Maths	20	15	14	16
	(12,05)	(8,38)	(7,36)	(11,70)
PPVT 5	25	20	11	17
	(11,90)	(11,37)	(4,99)	(7,70)
PPVT 12	31	35	12	20
	(12,81)	(11,76)	(4,77)	(7,36)

*Note: Administration time for the CDA and maths test of less than three minutes or more than 70 minutes were removed from this particular analysis as they were considered unlikely and thus probably coding mistakes. In the case of the PPVT, administration time of less than four minutes and more than 60 minutes were also removed.

Regarding the scoring process, all the tests were scored by the fieldworkers following the procedures established in the Examiner's Manual. Most of the tests used in YL were objective so children's scores did not depend on scorers' discretion, with the only exception of the reading and writing items where the scoring instructions were quite simple and straightforward. Therefore, inter-scorer reliability was not an issue in the context of the YL study.

Finally, during data collection, the work of the examiners was permanently overseen by supervisors specifically trained for that task who mastered all the procedures. All the survey operations were coordinated by the General Fieldwork Coordinator appointed by each research team. Additionally, principal researchers from each country personally supervised the administration in some of the sites. This confirmed that the study was conducted according to what was planned.

4.3 Database cleaning

The data sets provided by the four country teams were reviewed and cleaned before estimating the reliability and validity of the three achievement and development instruments. The following general steps were implemented during the data cleaning process:

- *Identification variable (ID) cleaning*. In YL, data about children and their homes and families appear in several data files, e.g. cognitive development and achievement instruments, child questionnaire and household questionnaire. Checking that the ID of the children was correctly introduced in the different data sets was crucial since this is the main variable used to merge the data sets.
- *Children's age validation*. In the context of a cognitive development and achievement assessment, the correct calculation of children age is very important, particularly in tests like the PPVT where the child's chronological age determines his or her starting point in the test. For that reason, the age of the children was validated by contrasting their recorded age with their date of birth.
- *Resolving inconsistencies in the data*. Cross-tabulation of variables was undertaken to check for consistency among the responses given. All such inconsistencies that were detected were flagged, and the national research team leader was asked to correct the data. Those cases that could not be corrected and where data made no sense were removed from the data set.

The cleaning process of the PPVT data set requires an especial mention. Out of the three cognitive and development instruments, the PPVT was the most difficult to clean because of the rules for test administration. Unlike the other tests, in the PPVT not all the items are administered to a given child but only those within his or her critical range, i.e., those items extremely easy or extremely hard for the individual are not administered. This requires that the examiner select the appropriate Start Item (according to the child's chronological age) and correctly establish the Basal Item Set and Ceiling Item Set for the individual.

All the observations in the data set were reviewed to verify that Basal and Ceiling Sets were correctly established for each child. When any of these conditions was not met, the observation was removed from the data set. Observations were also removed when the administration conditions were not appropriate because the light in the room was insufficient or the child showed evidence of uncorrected vision or hearing loss. Table 8 summarises the steps followed to clean the PPVT data sets. However, there were some other errors that could be corrected during the cleaning process. These errors have also been reported at the end of Table 8. In these cases, children were administered more PPVT items than required, i.e. the test was not discontinued when the children reached the Ceiling Item Set. This usually happened because the examiner did not mark an incorrect answer as an error and therefore failed in recognising that the child had already made enough errors to reach a ceiling. In such cases the additional items were not considered for score calculations.

Table 8: *Summary of valid and invalid observations of the PPVT data sets from the four countries*

Country:	**India**	**India**	**Ethiopia**	**Ethiopia**	**Vietnam**	**Vietnam**	**Peru**	**Peru**
Cohort:	**5**	**12**	**5**	**12**	**5**	**12**	**5**	**12**
INITIAL OBSERVATIONS	1940	993	1912	979	1969	990	1963	685
Children who did not take the test because they had serious mental or physical disabilities or could not respond to the sample items and instructions correctly (*)	16	0	37	2	18	1	40	1
Children who were discontinued in the test before time (i.e. never formed a ceiling set)	43	13	4	4	104	11	14	3
Children who never formed a basal set or started in the incorrect item	7	7	6	20	97	31	3	7
Children tested in inadequate conditions (insufficient light)	3	0	0	0	2	2	1	1
Children tested in inadequate conditions (child showed evidence of uncorrected vision or hearing loss)	20	1	4	0	1	0	1	0
Children who did not have accurate information regarding the language used to take the test(**)	0	1	0	0	0	0	1	1
TOTAL OBSERVATIONS REMOVED FROM DATA SET	89	22	51	26	222	45	60	13
FINAL OBSERVATIONS	1851	971	1861	953	1747	945	1903	672
Children who were not discontinued in the test on time (***)	39	22	9	2	95	14	55	28

NOTES:

* This is not an administration error. However, these observations had to be removed from the data set because the PPVT should not be administered to children with serious mental or physical disabilities that impaired their ability to understand the task or those who were not able to successfully complete the training items.

** This is not an administration error. However, these observations had to be removed from the data set because in the case of the PPVT the analysis had to be done separately for each language.

*** Although this constitutes an administration error, it was possible to correct the error with the information available and it was not necessary to remove the observations from the data set.

After cleaning all the databases following the procedures described above, the total number of observations available for analysis by cohort and country was:

Table 9. *Number of children by cohort and country available for analysis*

	Younger cohort		Older cohort	
	CDA	**PPVT**	**Maths**	**PPVT**
Ethiopia	1888	1861	951	953
India	1927	1851	983	971
Peru	1950	1903	683	672
Vietnam	1906	1747	981	945

Finally, Table 10 summarises the information of the total number of observations available for analysis by cohort, country and language. For the CDA and Mathematics tests, the numbers reported are those given by the parent when asked about the maternal tongue of the child; for the PPVT, the number reported is the language used by the child for answering the test. The match is not perfect, as many bilingual children who had a given language reported as their main tongue by one of their parents used a different one to respond to the PPVT. For the DIF analysis, we used maternal tongue for the CDA and maths tests and language used during administration for the PPVT (we did not have the language used for test administration for the CDA and maths).

Table 10. *Number of children by cohort, country and language available for analysis after cleaning the datasets*

	Younger cohort		Older cohort	
	CDA	**PPVT**	**Maths**	**PPVT**
Ethiopia				
Amarigna	842	836	436	439
Oromifa	308	292	159	155
Tigrigna	378	376	198	196
Other(s)	360	357	158	163
India				
Telugu	1778	1660	946	928
Kannada	68	66	8	4
Other(s)	81	125	29	39
Peru				
Spanish	1732	1686	640	641
Quechua	216	182	43	31
Other(s)	2	35	-	-
Vietnam				
Tieng Viet Nam	1726	1609	921	937
H'mong	105	87	31	5
Other(s)	75	51	29	3

5. Results

This section is divided into two parts. We first present the psychometric properties of the tests used in the second round of YL: CDA, PPVT and the mathematics achievement test. We then present the results for the common achievement items from Rounds 1 and 2.

The validity and reliability analyses presented here were done following both Classical Test Theory (CTT) and Item Response Theory (IRT). The statistical package STATA 10 was used to perform the analyses according to CTT. In the case of the IRT analyses, we used the specialised software WINSTEPS 3.65 to estimate children's abilities and item difficulty for each test as well as the goodness of fit of the items.

5.1 Psychometric characteristics of the tests used in Round 2

The psychometric analyses reported in this section were done separately for each cohort within each country. Following both CTT and IRT, we estimated the psychometric properties of each of the items of the tests used in Round 2. Additionally, we identified the items with differential item functioning across subgroups of examinees based on gender and language, adjusting for total ability.

We then defined 5 criteria as indicators of a poor result for an item:

A. Item-Test correlation lower than 0.10

B. Infit out of the range 0.5 to 1.5

C. Outfit out of the range 0.5 to 1.5

D. The difference by gender is significant at 5 per cent (gender bias), controlling for overall ability

E. The difference by language is significant at 5 per cent (language bias), controlling for overall ability

When a given item from a test met criteria A, B and/or D, we decided to exclude the item from the analysis for the specific country and cohort where the problem with the item was found. We decided not to exclude from the analysis items flagged with criteria C because the outfit out of range does not represent a serious threat to the reliability of the test (Linacre 2002). Finally, regarding the last criterion (language bias), we decided that when one or more of the items from a test was flagged for this reason, the recommendation would be to perform the analysis for each linguistic group separately. Indeed this was the case for all the linguistic groups we could analyse (where the number of subjects allowed for comparisons). The tables containing item statistics for each of the tests can be found in Annex 1.[17] The psychometric analyses presented below use the final configuration of each of the tests after removing those items that did not meet the criteria previously defined.

17 Annexes can be found in the document, Psychometric characteristics of cognitive development and achievement instruments in Round 2 of Young Lives: Annexes, available from http://www.younglives.org.uk/publications/technical-notes.

5.2 Distributions of raw and Rasch scores

The best known of all theoretical probability distributions is the normal distribution, also known as the bell-shaped distribution, which is symmetrical around its mean value. Two characteristics are generally used to describe a distribution: skewness (lack of symmetry) and kurtosis (tallness or flatness). In a normal distribution skewness is zero and kurtosis is 3. When a distribution is not normal, there might be problems with either one of these indicators or both. There are two types of skewness: positive skew, when the right tail is longer and therefore the distribution is concentrated on the left side (right-skewed); and negative skew, when the left tail is longer and therefore the distribution is concentrated on the right side (left-skewed). In terms of its kurtosis, a normal distribution is *mesokurtic*, while a distribution which is not normal can be *leptokurtic* when it has a more acute peak around the mean (kurtosis values are greater than 3), or *platykurtic* when it has a smaller peak around the mean (kurtosis values are smaller than 3; Gujarati 2003).

As we mentioned before, all the tests used in YL are norm-referenced, therefore it is expected that they will have a normal distribution. Using the final scores of each test, we graphed the distribution of raw and Rasch scores for each of the tests by cohort and language. The decision to include both raw and Rasch scores was based on the fact that frequently the use of Rasch analysis contributes to the normalisation of test curves.

Table 11 summarises the characteristics of the distributions of raw and Rasch scores for each instrument by country and language. The graphs of all the distributions are included in Annex 2.

Table 11. *Characteristics of the distributions of raw and Rasch scores for each instrument by country and language*

		CDA		**Maths Test**		**PPVT younger cohort**		**PPVT older cohort**	
		Raw Score	**Rasch Score**	**Raw Score**	**Rasch Score**	**Raw Score**	**Rasch Score**	**Raw Score**	**Rasch Score**
Ethiopia	Amarigna	NS / PK	PS / LK	NS / PK	ND	PS / LK	PS / LK	PK	NS / PK
	Oromifa	NS	LK	PK	PS	N.A.	N.A.	N.A.	N.A.
	Tigrigna	PK	ND	NS	ND	N.A.	N.A.	N.A.	N.A.
India	Telugu	NS	LK	NS	NS	PS / LK	PS / LK	NS	NS / LK
	Kannada	ND	ND	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
Peru	Spanish	NS / LK	NS / LK	NS / LK	NS	PS / PK	PK	NS / LK	NS / LK
	Quechua	NS	LK	ND	ND	N.A.	N.A.	N.A.	N.A.
Vietnam	Tieng Viet Nam	NS/LK	LK	NS / LK	NS	P.S / LK	PS / PK	NS / LK	NS / LK
	H'mong	PS	PS / LK	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.

Notes:
ND: Normal Distribution

Where the distribution was not normal, one or more of the following were flagged:
NS: Negatively skewed
PS: Positively skewed
LK: Leptokurtic
PK: Platykurtic
N.A.: Not applicable (not enough subjects for the analysis)

A few comments on the characteristics of the distributions:

- Regarding the CDA
  - The use of Rasch scores contributes to the normalisation of test curves in the cases of Ethiopia (Tigrigna) and India (Kannada).
  - In every other case, the distribution is leptokurtic, i.e. it has a greater percentage of scores closer to the mean. This means that this particular test is only allowing us to differentiate between extremes (higher and lower achievers) but is not providing us with extensive information about those children located around the mean.
- Regarding the mathematics achievement test
  - The use of Rasch scores contributes to the normalisation of test curves in the cases of Ethiopia (Amarigna and Tigrigna) and Peru (Quechua).
  - In the other cases, the main problem is that the distribution is negatively skewed, i.e. it has a greater percentage of scores concentrated in the higher end of the distribution. This means there were problems in measuring children's ability because the items were extremely easy and a ceiling effect was observed for some or many children. This is particularly the case in India (Telugu) and Vietnam (Tieng Viet Nam).
- Regarding the PPVT – Younger Cohort
  - There is not a normal distribution of test scores in any country.
  - In general terms, the variable is biased and the distribution is positively skewed, i.e. it has a greater percentage of scores concentrated in the lower end of the distribution.
- Regarding the PPVT – Older Cohort
  - There is not a normal distribution of test scores in any country.
  - In general terms, the distributions of tests scores were found to be negatively skewed and leptokurtic, suggesting this test was too easy for many children, who were concentrated around the mean, which was located near the higher end of the distribution.

5.3 Reliability indexes and standard error of measurement (SEM) according to CTT and IRT

Using the corrected raw scores, we calculated the reliability indexes of each of the tests according to Classical Test Theory (CTT) and Item Response Theory (IRT). In the case of CTT, a reliability coefficient of 0.6 or more was considered acceptable for research purposes. In the case of IRT, a person reliability index above 0.5 was considered an adequate estimation of the individual's ability, meaning that the test could discriminate between higher and lower achievers (Linacre 2008).

For every test used in Round 2 of YL, separate reliability indexes were calculated for each language within a country, as long as the language had at least 100 observations. The only exception to this criterion was the PPVT. Since the test has 204 items for three countries, the number of observations required was larger (at least four observations for each item of the

test in order to have stability in the analysis). For this reason, in the case of PPVT, reliability indexes are only reported for the predominant languages.[18]

In addition to providing the reliability indexes above mentioned, we also reported the average standard error of measurement (SEM) across children as an indicator of overall precision of the scores calculated. As we mentioned before, there is an inverse relationship between the SEM and the reliability index: the higher the error, the lower the reliability. Tables 12 to 14 contain information about the reliability indexes and the SEM for each of the three tests used in the second round of YL.[19]

Table 12 contains information regarding the reliability indexes for the CDA. In general terms, the reliability coefficients of the CDA are acceptable according to CTT (Cronbach's Alpha above 0.6) and/or IRT (person reliability index above 0.5). The only serious exception is the case of Peru (Quechua) where both reliability coefficients are low according to the standards set. In the other cases, at least one of the reliability coefficients is above the standards set.

Table 12. *Reliability indexes and standard error of measurement (SEM) for the CDA by country and language according to Classic Test Theory (CTT) and Item Response Theory (IRT)*

Country	Languages	Items deleted	CTT		IRT[2]	
			Reliability[1]	SEM[3]	Reliability	SEM[3]
Ethiopia	Amarigna	Item 1	0.72	1.59	0.66	35.9
	Oromifa	Item 1	0.58	1.66	0.60	32.8
	Tigrigna	Item 1	0.68	1.67	0.67	32.6
India	Telugu	Item 12	0.63	1.58	0.59	36.0
	Kannada	Item 12	0.56	1.54	0.54	36.3
Peru	Spanish	Item 1, 9 and 14	0.61	1.36	0.56	42.2
	Quechua	Item 1, 9 and 14	0.43	1.50	0.45	36.7
Vietnam	Tieng Viet Nam	Item 6	0.65	1.37	0.61	41.7
	H'mong	Item 6	0.48	1.52	0.51	34.4

1. Using Cronbach's Alpha coefficient
2. The Rasch scores were fixed with a mean of 300 and standard deviation of 50
3. Annex 3 and Annex 4 present additional results for SEM in CTT and IRT respectively

Table 13 contains information regarding the reliability indexes for the mathematics achievement test. In general terms, the reliability coefficients for this test are acceptable both according to CTT (Cronbach's Alpha above 0.6) and IRT (person reliability index above 0.5). This suggests that the correlation between items is high and they are all measuring the same construct. The only exceptions are Peru (Quechua) and Vietnam (Tieng Viet Nam) where only one of the reliability coefficients is above the standards set.

18 This means that Rasch scores were only calculated for children who spoke a language that had a sufficient number of subjects. For other cases no Rasch score is calculated but a corrected raw score is available for research purposes.

19 Further information about the estimation of SEM according to CTT and IRT can be found in Annexes 3 and 4 respectively.

Table 13. *Reliability indexes and standard error of measurement (SEM) for the maths test by country and language according to Classic Test Theory (CTT) and Item Response Theory (IRT)*

Country	Languages	Items deleted	CTT		IRT[2]	
			Reliability[1]	SEM	Reliability	SEM
Ethiopia	Amarigna	Item 9	0.73	1.18	0.69	51.4
	Oromifa	Item 9	0.75	1.20	0.71	51.0
	Tigrigna	Item 9	0.70	1.17	0.65	50.7
India	Telugu	Item 8 and 9	0.82	0.96	0.61	62.3
Peru	Spanish	Item 1 and 8	0.66	1.01	0.54	62.0
	Quechua	Item 1 and 8	0.59	1.10	0.63	54.3
Vietnam	Tieng Viet Nam	Item 7	0.70	0.89	0.34	74.4

1. Using Cronbach's Alpha coefficient
2. The Rasch scores were fixed with a mean of 300 and standard deviation of 50.

Finally, Table 14 presents the reliability indexes for the PPVT according to CTT and IRT. It is important to notice that in the case of the PPVT, we did not use Cronbach's Alpha coefficient to estimate the reliability according to CTT. Instead, we used split-half reliability coefficients, another measure of internal consistency, to estimate the reliability of the test. The split-half method correlates the scores on odd items versus even items, giving credit below the basal and assuming errors above the ceiling. Then, the Spearman-Brown prophecy formula is used to estimate the reliability coefficients for the full length test.[20,21] Regarding IRT, the person reliability index was estimated as in the previous two tests.

As it can be seen in Table 14, the reliability indexes for the PPVT according to CTT are quite high and split-half reliability coefficients are above 0.90 for every country and cohort. Reliability indexes according to IRT are also acceptable according to the standards previously set, since person reliability indexes are above 0.96 for every country and cohort.[22]

20 In the case of the PPVT, separate analyses were done for the younger and older cohort. In order to calculate the reliability index of each cohort, we first calculated the reliability indexes of each age group included in the cohort (e.g. 4-year-old group and 5-year-old group in the younger cohort). Then the global reliability index for the cohort was obtained by calculating the simple mean of the different reliability indexes for each age group.

21 It could be argued that giving credit below the basal and assuming errors above the ceiling might artificially inflate the reliability of the test (even though this is exactly what is done when calculating the raw score as per test instructions). The following method is an alternative to estimate the reliability of the PPVT according to CTT without making this assumption. Using also the split-half method, one could use a Rasch model to estimate the ability of the children in each half of the test (odd items versus even items). The two abilities are then correlated and finally the Spearman-Brown prophecy formula is used to estimate the reliability coefficients for the full length test. When we calculated the reliability of the PPVT following this method, the reliability coefficients were also high and above 0.90 for all countries and both cohorts.

22 The items deleted in the PPVT datasets due to inadequate psychometric properties by country and cohort of study can be found in Annex 5.

Table 14. *Reliability indexes and standard error of measurement (SEM) for the PPVT by country according to Classic Test Theory (CTT) and Item Response Theory (IRT)*

	CTT		IRT[2]	
	Reliability[1]	SEM	Reliability	SEM
5 years				
Ethiopia (Amarigna)	0.90	4.6	0.96	20.4
India (Telugu)	0.97	3.7	0.96	18.9
Peru (Spanish)	0.95	4.0	0.98	21.5
Vietnam (Tieng Viet Nam)	0.96	3.6	0.97	17.9
12 years				
Ethiopia (Amarigna)	0.98	3.7	0.98	14.9
India (Telugu)	0.97	4.2	0.96	14.9
Peru (Spanish)	0.95	3.5	0.98	22.4
Vietnam (Tieng Viet Nam)	0.98	3.6	0.97	16.2

1. Estimated using the split half method (with Spearman Brown correction)
2. The Rasch scores were fixed for each language with a mean of 300 and standard deviation of 50

5.4 Correlations between test scores

The correlations between raw and Rasch scores were calculated separately for each language (with at least 100 observations) within each country.[23] All the correlations are almost perfect and positive as expected, given that Rasch scores are monotonically 1-1 with the raw scores; the only difference is the normalisation. The correlation coefficients in all the cases are above 0.90 and statistically significant.

Table 15. *Correlation between raw and Rasch scores by country and main languages*

	CDA	Maths	PPVT – Younger Cohort	PPVT – Older Cohort
Ethiopia				
Amarigna	0.97 **	0.97 **	0.97 **	0.99 **
Oromifa	0.98 **	0.98 **	N.A	N.A
Tigrigna	0.99 **	0.98 **	N.A	N.A
India				
Telugu	0.97 **	0.93 **	0.97 **	0.98 **
Kannada	0.99 **	N.A	N.A	N.A
Peru				
Spanish	0.98 **	0.98 **	0.99 **	0.99 **
Quechua	0.98 **	0.99 **	N.A	N.A
Vietnam				
Tieng Viet Nam	0.98 **	0.98 **	0.98 **	0.99 **
H'mong	0.99 **	N.A	N.A	N.A

Note: Pearson correlation was used in all the cases
**p<.05, *p<10 • N.A: Not applicable

23 The number of observations available for analysis by country and language is specified in Table 10 of this report.

Table 16 shows the correlations between the two tests administered in each cohort. The correlations were calculated for the full sample within each country and also for the predominant language of the country.[24] As shown, the correlation coefficients in all the cases are high, positive and statistically significant as would be expected from previous literature. The results suggest that children who perform relatively highly in one test also tend to have high results in the other.

Table 16. *Correlation between test scores (Rasch scores) by cohort, country and language*

	Younger Cohort (PPVT – CDA)		Older Cohort (PPVT – maths)	
Ethiopia				
Full sample	0.53	***	0.42	***
Amarigna	0.63	***	0.43	***
India				
Full sample	0.53	***	0.59	***
Telugu	0.55	***	0.59	***
Peru				
Full sample	0.58	***	0.60	***
Spanish	0.59	***	0.59	***
Vietnam				
Full sample	0.56	***	0.62	***
Tieng Viet Nam	0.48	***	0.53	***

Note: Pearson correlation was used in all the cases ***p<.01, **p<.05, *p<10

Table 17 shows the correlation between the original raw score of each test and the corrected raw score (after removing those items that did not meet the criteria defined at the beginning of this section). In all the cases, the correlation coefficients are high, positive and statistically significant.

Table 17. *Correlation between original raw score and corrected raw score by country and main languages*

	CDA		Maths		PPVT – Younger Cohort		PPVT – Older Cohort	
Ethiopia								
Amarigna	0.99	***	0.98	***	0.98	***	0.99	***
Oromifa	0.98	***	0.98	***	N.A		N.A	
Tigrigna	0.99	***	0.98	***	N.A		N.A	
India								
Telugu	0.99	***	0.97	***	0.99	***	0.99	***
Kannada	0.99	***	N.A		N.A		N.A	
Peru								
Spanish	0.96	***	0.97	***	0.99	***	0.99	***
Quechua	0.94	***	0.96	***	N.A		N.A	
Vietnam								
Tieng Viet Nam	0.98	***	0.97	***	0.99	***	0.99	***
H'mong	0.97	***	N.A		N.A		N.A	

Note: Pearson correlation was used in all the cases ***p<.01, **p<.05, *p<10 N.A: Not applicable

24 We were only able to calculate the correlation between tests for the predominant language because in the case of the PPVT none of the other languages had enough observations to conduct separate analysis by language.

As mentioned before, the PPVT offers raw scores as well as standard scores. The test manual includes tables to convert the raw scores into standard scores. However, we have not used standard scores for Ethiopia, India or Vietnam because the standardisation samples for the original PPVT were for the English language.

In the case of Peru, standard scores are available because the version of the PPVT used in this country had been standardised for a Latin-American sample. Table 18 presents the correlation coefficients between PPVT's original raw score, corrected raw score and standard score. The correlation between these three scores is positive and very high.

Table 18: *Correlation between PPVT's original raw score, standard score and corrected raw score for the Peru – Spanish sample*

	Younger Cohort		Older Cohort	
	Original Raw score	**Standard score**	**Original Raw score**	**Standard score**
Standard score	0.97	1	0.97	1
Corrected Raw score	0.99	0.97	0.99	0.96

Note: Pearson correlation was used in all the cases

5.5 Correlations between test scores and demographic and educational variables

In this section we present validity evidence based on relations to other variables to test if predicted associations are confirmed. Specifically we expected that, on average, children of more educated parents should get higher results than children of parents with a lower educational level. Also, children in higher grades should have higher scores. In these analyses we present unadjusted results, since the purpose is to confirm associations and not provide explanations (which would require controlling for covariates).

Before presenting the correlations of test scores and educational variables, it is important to have descriptive statistics of the children and their parents' educational level (see Tables 19 and 20). In the case of the younger cohort, most of the children from India, Peru and Vietnam have pre-school education (regardless of their language); however in Ethiopia the percentage of children with pre-school education is considerably lower, especially among children speaking Oromifa (10.7 per cent) and Tigrigna (2.4 per cent). Parents' educational attainment is low in the four countries. Except in the cases of Peru (Spanish) and Vietnam (Tieng Viet Nam), the percentage of parents with no education is considerably higher, especially in Ethiopia and India. Additionally, the educational level of mothers is lower than the educational level of fathers.

In the case of the older cohort, practically all the children in the sample are enrolled at school, regardless of their language. On average, children from India, Peru and Vietnam were in the sixth grade of basic education at the time of data collection; however, children from Ethiopia were mostly in the fourth grade. In relation to parents' educational level, the educational attainment of parents from the older cohort is also low, especially in Ethiopia and India.

Table 19. *Parents' and children's education by language and country, younger cohort (standard deviation)*

	Percentage of children with pre-school education	Mother			Father		
		No education	Secondary Complete	Higher education (University)	No education	Secondary Complete	Higher education (University)
Ethiopia							
Amarigna	48.8	34.8	7.4	1.2	17.8	10.9	4.4
n=842	(50.0)	(47.7)	(26.1)	(10.8)	(38.3)	(31.2)	(20.5)
Oromifa	10.7	45.8	1.6	0.3	16.2	5.2	1.0
n=308	(31.0)	(49.9)	(12.7)	(5.7)	(36.9)	(22.2)	(9.8)
Tigrigna	2.4	76.7	0.3	0.3	38.4	1.9	1.3
n=378	(15.3)	(42.3)	(5.1)	(5.1)	(48.7)	(13.5)	(11.4)
India							
Telugu	87.0	52.4	2.9	2.9	33.7	5.8	7.2
n=1778	(33.7)	(50.0)	(16.7)	(16.7)	(47.3)	(23.4)	(25.9)
Kannada	97.1	29.4	1.5	0.0	20.6	2.9	5.9
n=68	(17.0)	(45.9)	(12.1)	(0.0)	(40.7)	(17.0)	(23.7)
Peru							
Spanish	85.5	4.6	21.1	5.5	0.9	28.6	8.8
n=1732	(35.2)	(20.9)	(40.8)	(22.9)	(9.6)	(45.2)	(28.4)
Quechua	73.6	40.3	0.0	0.0	3.7	8.3	0.0
n=216	(44.2)	(49.2)	(0.0)	(0.0)	(18.9)	(27.7)	(0.0)
Vietnam							
Tieng Viet Nam	93.4	3.9	7.1	4.3	2.7	10.8	5.3
n=1726	(24.9)	(19.5)	(25.6)	(20.3)	(16.3)	(31.0)	(22.5)
H'Mong	61.9	91.4	0.0	0.0	65.7	0.0	0.0
n=105	(48.8)	(28.1)	(0.0)	(0.0)	(47.7)	(0.0)	(0.0)

Table 20. *Parents' and children's education by language and country, older cohort (standard deviation)*

	Percentage of children who are enrolled in the school	Mean grade of the children who attend school	Mother			Father		
			No education	Secondary Complete	Higher education (University)	No education	Secondary Complete	Higher education (University)
Ethiopia								
Amarigna	98.2	4.8	36.9	6.4	0.0	21.3	8.0	1.8
n=436	(13.4)	(1.5)	(48.3)	(24.5)	(0.0)	(41.0)	(27.2)	(13.4)
Oromifa	94.3	3.7	37.1	1.3	0.6	18.9	4.4	1.3
n=159	(23.2)	(1.6)	(48.5)	(11.2)	(7.9)	(39.2)	(20.6)	(11.2)
Tigrigna	97.0	4.2	78.8	1.0	0.0	36.9	0.1	0.5
n=198	(17.2)	(1.4)	(41.0)	(10.0)	(0.0)	(48.4)	(7.1)	(7.1)
India								
Telugu	89.7	6.7	61.0	1.1	2.5	43.2	3.8	5.8
n=946	(30.4)	(1.0)	(48.8)	(10.2)	(15.7)	(49.6)	(19.1)	(23.4)
Peru								
Spanish	98.9	6.0	6.3	18.9	4.2	0.9	30.0	8.3
n=640	(10.4)	(1.1)	(24.2)	(39.2)	(20.1)	(9.6)	(45.9)	(27.6)
Quechua	100.0	5.3	62.8	0.0	0.0	14.0	9.3	0.0
n=43	(0.0)	(1.2)	(48.9)	(0.0)	(0.0)	(35.1)	(29.4)	(0.0)
Vietnam								
Tieng Viet Nam	97.4	6.7	5.1	9.1	2.9	3.6	11.4	5.6
n=921	(15.9)	(0.6)	(22.0)	(28.8)	(16.9)	(18.6)	(31.8)	(23.1)

Regarding the CDA, in general the correlation between test scores (both raw and Rasch scores) and parents' education is positive and statistically significant, as would be expected from the literature.

Table 21. *Correlation between CDA score and education variables by country and language*

	Raw scores		Rasch scores	
	Father's education [a]	**Mother's education [a]**	**Father's education [a]**	**Mother's education [a]**
Ethiopia				
Amarigna	0.43 **	0.48 **	0.43 **	0.48 **
Oromifa	0.15 **	0.18 **	0.15 **	0.18 **
Tigrigna	0.06	0.11 **	0.06	0.11 **
India				
Telugu	0.27 **	0.28 **	0.27 **	0.28 **
Kannada	0.22	0.11 **	0.22	0.11 **
Peru				
Spanish	0.31 **	0.35 **	0.31 **	0.35 **
Quechua	0.03	0.11	0.03	0.11
Vietnam				
Tieng Viet Nam	0.22 **	0.23 **	0.22 **	0.23 **
H'mong	0.38 **	0.36 **	0.38 **	0.36 **

Note: Spearman's rank correlation was used in all the cases

[a] Parents' education is a categorical variable with the values of: 0=No education, 1=parent could read and write, 2=Incomplete Primary, 3=Complete Primary, 4=Incomplete Secondary, 5=Complete Secondary, 6=Technical or Vocational education (complete or incomplete), 7=University education (complete or incomplete).

**p<.05, *p<.10

In relation to the Maths test, a positive and statistically significant correlation between test scores (both raw and Rasch scores) and educational variables of the child and his/her parents was found. In the case of Ethiopia, the correlation between test scores and children's grade of study was considerably higher than the correlation between test score and parents' educational level.

Table 22. *Correlation between Maths test score and education variables by country and language*

	Raw scores			Rasch scores		
	Child's grade[a]	Father's education[b]	Mother's education[b]	Child's grade[a]	Father's education[b]	Mother's education[b]
Ethiopia						
Amarigna	0.41 **	0.34 **	0.33 **	0.41 **	0.34 **	0.33 **
Oromifa	0.57 **	0.32 **	0.26 **	0.57 **	0.32 **	0.26 **
Tigrigna	0.35 **	0.10	0.16 **	0.35 **	0.10	0.16 **
India						
Telugu	0.12 **	0.22 **	0.26 **	0.12 **	0.22 **	0.26 **
Kannada	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
Peru						
Spanish	0.37 **	0.39 **	0.37 **	0.37 **	0.39 **	0.37 **
Quechua	0.15	0.35 **	0.50 **	0.15	0.35 **	0.50 **
Vietnam						
Tieng Viet Nam	0.21 **	0.30 **	0.30 **	0.21 **	0.30 **	0.30 **
H'mong	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.

Note: Spearman's rank correlation was used in all the cases.

[a] Defined as the last grade attended by the child.

[b] Parents' education is a categorical variable with the values of: 0=No education, 1=parent could read and write, 2=Incomplete Primary, 3=Complete Primary, 4=Incomplete Secondary, 5=Complete Secondary, 6=Technical or Vocational education (complete or incomplete), 7=University education (complete or incomplete).

**p<.05, *p<10

N.A: Not applicable

In the case of the PPVT, separate correlations between test scores (both raw and Rasch scores) and educational variables were calculated for each cohort. In the younger cohort, a positive and statistically significant correlation was found between PPVT score and parents' educational variables. In all the countries, mothers' educational level had higher correlation values with the PPVT score than fathers' educational level.

In the older cohort, positive and statistically significant correlations were also found between test scores and parents' educational level in every country analysed. Just as in the case of the younger cohort, PPVT scores in general seemed to be more associated with mothers' educational level except in the case of Ethiopia, where the correlation is higher with fathers' educational level. In the case of the older cohort, the correlation between the test score and the child's grade of study was also estimated. This correlation was positive and statistically significant in every country, as would be expected according to the previous literature. The association between PPVT scores and children's education was especially high in Peru, where the correlation coefficient was around 0.5.

Table 23. *Correlation among the scores and education variables by country and language in the PPVT test*

	Raw scores			Rasch scores		
	Child's grade[a]	Father's education[b]	Mother's education[b]	Child's grade[a]	Father's education[b]	Mother's education[b]
	Younger Cohort					
Ethiopia						
Amarigna	N.A.	0.48 **	0.52 **	N.A.	0.48 **	0.52 **
India						
Telugu	N.A.	0.32 **	0.37 **	N.A.	0.32 **	0.37 **
Peru						
Spanish	N.A.	0.50 **	0.56 **	N.A.	0.50 **	0.56 **
Vietnam						
Tieng Viet Nam	N.A.	0.35 **	0.36 **	N.A.	0.35 **	0.36 **
	Older Cohort					
Ethiopia						
Amarigna	0.38 **	0.45 **	0.39 **	0.38 **	0.45 **	0.39 **
India						
Telugu	0.24 **	0.24 **	0.31 **	0.24 **	0.24 **	0.31 **
Peru						
Spanish	0.47 **	0.44 **	0.47 **	0.47 **	0.44 **	0.47 **
Vietnam						
Tieng Viet Nam	0.38 **	0.42 **	0.46 **	0.38 **	0.42 **	0.46 **

Note: Spearman's rank correlation was used in all the cases.

a Defined as the last grade attended by the child.

b Parents' education is a categorical variable with the values of: 0=No education, 1=parent could read and write, 2=Incomplete Primary, 3=Complete Primary, 4=Incomplete Secondary, 5=Complete Secondary, 6=Technical or Vocational education (complete or incomplete), 7=University education (complete or incomplete).

**p<.05, *p<10

N.A: Not applicable due to low variability in children's grades; see analysis presented below.

Since we were not able to correlate test scores with children's educational level for the younger cohort due to low variability in these, we try to approximate the relationship between these variables by comparing the mean scores of children without pre-school, with pre-school and currently attending first grade (although these cases were few considering children in the sample were between 4.5 and 5.5 years old). Tables 24 to 27 present the mean raw and Rasch score in the CDA and the PPVT by educational attainment for each of the four countries. Readers should bear in mind that the following results (as well as the previous ones) only suggest there is an association between education and test scores, not a causal relationship.

As would be expected from previous research, in general children who have had access to formal education perform better in the test than children who have never attended pre-school in their lives. In the cases of Peru (Spanish) and Vietnam (Tieng Viet Nam), the longer children spent in pre-school, the higher the score in the PPVT. In the other cases, children attending pre-school in general perform better than those who do not. However the differences between those who have attended for less than six months and those who have attended for more than six months are not always in favour of the latter.

The only country with a large number of children from the younger cohort attending the first grade of primary education was India. In this case, it was also clear that more educated children perform better than children with less or no education.

Table 24. *Mean raw and Rasch score for the PPVT and CDA by educational attainment in the younger cohort (standard deviations) – Ethiopia*

	First Grade	Pre-school attended:		Never attended a pre-school
		Less than 6 months	More than 6 months	
			Amarigna	
		(n=27)	**(n=384)**	**(n=430)**
PPVT – Raw score	N.A	31.9[a]	30.7[a]	17.0[b]
		(15.4)	(16.9)	(7.6)
PPVT – Rasch Score	N.A	333.1[a]	326.6[a]	273.5[b]
		(48.4)	(51.2)	(31.5)
CDA – Raw Score	N.A	11.1[a]	10.3[a]	7.4[a]
		(2.4)	(2.6)	(2.7)
CDA – Rasch Score	N.A	340.1[a]	323.6[a]	276.6[a]
		(48.6)	(47.1)	(40.3)
			Oromifa	
		(n=8)	**(n=25)**	**(n=275)**
PPVT – Raw score	N.A.	N.A.	N.A.	N.A.
PPVT – Rasch Score	N.A.	N.A.	N.A.	N.A.
CDA – Raw Score	N.A.	9.9[a,b]	9.6[a]	8.0[b]
		(2.2)	(2.4)	(2.5)
CDA – Rasch Score	N.A.	331.6[a,b]	327.3[a]	296.6[b]
		(45.9)	(52.3)	(49.0)
			Tigrigna	
			(n=8)	**(n=369)**
PPVT – Raw score	N.A.	N.A.	N.A.	N.A.
PPVT – Rasch Score	N.A.	N.A.	N.A.	N.A.
CDA – Raw Score	N.A.	N.A	11.4[a]	7.6[b]
			(1.6)	(2.9)
CDA – Rasch Score	N.A.	N.A	368.0[a]	298.2[b]
			(47.1)	(48.8)

Note: Means with different super index are statistically different at 5 per cent (using a t-test for independent samples). Only groups with more than 5 children were considered for the multiple comparisons.
N.A: Not applicable

Table 25. *Mean raw and Rasch score for the PPVT and CDA by educational attainment in the younger cohort (standard deviations) – India*

	First Grade	Pre-school attended:		Never attended a pre-school
		Less than 6 months	More than 6 months	
	Telugu			
	(n=209)	**(n=257)**	**(n=1119)**	**(n=193)**
PPVT - Raw score	41.3[a]	26.2[b]	26.2[b]	22.5[b]
	(29.4)	(19.1)	(19.8)	(18.0)
PPVT – Rasch Score	330.6[a]	296.8[b]	297.3[b]	287.2[b]
	(59.5)	(48.4)	(47.4)	(43.8)
CDA - Raw Score	10.1[a]	9.5[a,b]	9.3[b]	8.8[b]
	(2.4)	(2.7)	(2.6)	(2.6)
CDA - Rasch Score	314.2[a]	300.1[b]	299.1[b]	289.6[b]
	(47.7)	(52.3)	(49.8)	(47.4)
	Kannada			
			(n=62)	
PPVT - Raw score	N.A.	N.A.	N.A.	N.A.
PPVT – Rasch Score	N.A.	N.A.	N.A.	N.A.
CDA - Raw Score	N.A.	N.A	9.7	N.A
			(2.3)	
CDA - Rasch Score	N.A.	N.A	301.7	N.A
			(50.6)	

Note: Means with different super index are statistically different at 5 per cent (using a t-test for independent samples). Only groups with more than 5 children were considered for the multiple comparisons.
N.A: Not applicable

Table 26. *Mean raw and Rasch score for the PPVT and CDA by educational attainment in the younger cohort (standard deviations) – Peru*

	First Grade	Pre-school attended:		Never attended a pre-school
		Less than 6 months	More than 6 months	
			Spanish	
	(n=26)	(n=616)	(n=839)	(n=251)
PPVT - Raw score	33.0[a,b,c]	29.9[b]	34.9[c]	17.6[d]
	(16.6)	(17.4)	(17.8)	(13.0)
PPVT – Rasch Score	306.9[a,b,c]	298.2[b]	311.7[c]	262.3[d]
	(44.6)	(48.3)	(48.4)	(41.0)
CDA - Raw Score	9.0[a]	8.5[a]	8.8[a]	7.1[b]
	(2.3)	(2.1)	(2.0)	(2.3)
CDA - Rasch Score	313.0[a]	301.3[a]	307.6[a]	270.0[b]
	(53.4)	(48.5)	(47.0)	(52.2)
			Quechua	
		(n=103)	(n=53)	(n=57)
PPVT - Raw score	N.A.	N.A.	N.A.	N.A.
PPVT – Rasch Score	N.A.	N.A.	N.A.	N.A.
CDA - Raw Score	N.A	8.0[a]	7.9[a]	7.5[a]
		(2.1)	(2.0)	(1.9)
CDA - Rasch Score	N.A	303.3[a]	301.6[a]	291.8[a]
		(53.4)	(48.3)	(45.6)

Note: Means with different super index are statistically different at 5 per cent (using a t-test for independent samples). Only groups with more than 5 children were considered for the multiple comparisons.

N.A: Not applicable

Table 27. *Mean raw and Rasch score for the PPVT and CDA by educational attainment in the younger cohort (standard deviations) – Vietnam*

	First Grade	Pre-school attended:		Never attended a pre-school
		Less than 6 months	More than 6 months	
		Tieng Viet Nam		
		(n=321)	**(n=1276)**	**(n=114)**
PPVT - Raw score	N.A	32.7[a]	40.9[b]	26.0[c]
		(14.8)	(18.2)	(14.2)
PPVT – Rasch Score	N.A	283.1[a]	307.3[b]	259.3[c]
		(42.9)	(49.5)	(46.0)
CDA - Raw Score	N.A	10.0[a]	10.3[a]	8.2[b]
		(2.6)	(2.1)	(3.1)
CDA - Rasch Score	N.A	299.3[a]	303.6[a]	263.8[b]
		(55.5)	(46.1)	(61.3)
			H'Mong	
			(n=63)	**(n=40)**
PPVT - Raw score	N.A.	N.A.	N.A.	N.A.
PPVT – Rasch Score	N.A.	N.A.	N.A.	N.A.
CDA - Raw Score	N.A.	N.A	6.8[a]	5.7[b]
			(2.3)	(1.7)
CDA - Rasch Score	N.A.	N.A	310.1[a]	283.4[b]
			(53.7)	(40.3)

Note: Means with different super index are statistically different at 5 per cent (using a t-test for independent samples). Only groups with more than 5 children were considered for the multiple comparisons.
N.A: Not applicable

The correlations above show a pattern of higher test results associated with higher parental education and school experience by the child. This constitutes an important source of evidence of the validity of the tests used.

The fact that more education is in general associated with better test scores could be partially explained by the children's age: older children are attending higher grades. Even though there is not much variability in age within each cohort (only 12 months), we calculated the correlation between test scores (both raw and Rasch) and age in months. The results are shown in Tables 28 and 29.

Table 28. *Correlation between raw score and age in months by country and main languages*

	CDA		Maths		PPVT – Younger Cohort		PPVT – Older Cohort	
Ethiopia								
Amarigna	0.12	***	0.01		0.16	***	0.08	*
Oromifa	0.10	*	-0.05		N.A		N.A	
Tigrigna	0.12	**	0.04		N.A		N.A	
India								
Telugu	0.16	***	-0.09	***	0.12	***	-0.06	*
Kannada	0.18		N.A		N.A		N.A	
Peru								
Spanish	0.37	***	0.16	***	0.49	***	0.31	***
Quechua	0.04		0.26		N.A		N.A	
Vietnam								
Tieng Viet Nam	0.22	***	0.07	**	0.21	***	0.20	***
H'mong	0.13		N.A		N.A		N.A	

Note: Pearson correlation was used in all the cases ***p<.01, **p<.05, *p<10 N.A: Not applicable

Table 29. *Correlation between Rasch score and age in months by country and main languages*

	CDA		Maths		PPVT – Younger Cohort		PPVT – Older Cohort	
Ethiopia								
Amarigna	0.13	***	0.01		0.17	***	0.08	*
Oromifa	0.11	**	-0.03		N.A		N.A	
Tigrigna	0.12	**	0.04		N.A		N.A	
India								
Telugu	0.15	***	-0.08	**	0.13	***	-0.04	
Kannada	0.17		N.A		N.A		N.A	
Peru								
Spanish	0.36	***	0.17	***	0.49	***	0.31	***
Quechua	0.04		0.28	*	N.A		N.A	
Vietnam								
Tieng Viet Nam	0.21	***	0.08	**	0.22	***	0.22	***
H'mong	0.14		N.A		N.A		N.A	

Note: Pearson correlation was used in all the cases ***p<.01, **p<.05, *p<10 N.A: Not applicable

In general, the correlations are positive and statistically significant. Some are relatively small - less than 0.2 - but others are quite high, especially for Peru. This could be explained partly by the fact that in these countries older children are attending more advanced grades and this relates to higher results in tests. Further descriptive information about the relationship between the scores, age and education of the children for all four countries can be found in Annex 6.

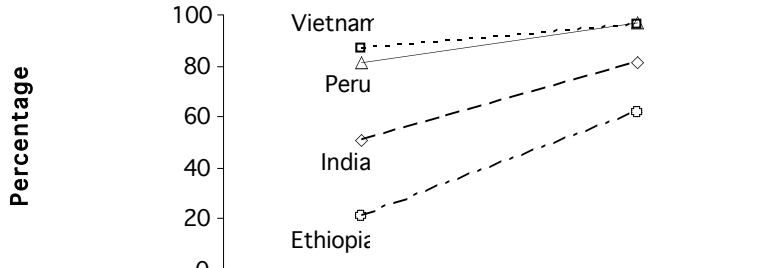
5.6 Comparison of common items from Rounds 1 and 2

As mentioned before, reading, writing and numeracy assessment items administered to the children from the older cohort in Round 1 of YL were also administered in Round 2 to assure comparability of information. The reading item required the children to read three letters, one word and one sentence. The writing item required that each child write a simple sentence which was spoken out loud by the examiner. Finally, the maths item required children to solve a basic multiplication task.[25] It was expected that the results would be higher in Round 2, given increased experience in school for most children. Only children with answers in both rounds are included in the analyses below.

In general terms, results from Round 1 indicated that more than 50 per cent of the children were able to read a sentence, write without difficulty and solve a basic multiplication problem, except in Ethiopia where these percentages were considerably lower at near 20 per cent. Four years later, when children were between 11.5 and 12.5 years old, a clear ceiling effect was observed in most of the sample. Children from Peru and Vietnam were able to read, write and solve the multiplication problem without difficulty (percentages in both countries were near or above 90 per cent for each of the items). In India the percentages were also higher and near 70 per cent of children capable of responding to the three items without difficulty. However, in Ethiopia the percentages were still low compared to the other countries.[26]

Despite the fact that most of the children in the sample were able to solve the three achievement items, it is interesting to notice the different evolution patterns within countries. Ethiopian children improved in all areas after starting low. Indian children could already do maths and improved in the other areas. Peruvian children could already read and improved in the other areas. And finally Vietnamese children scored highly in all areas and went to score very highly. These results suggest there is an enormous country by item by time interaction, i.e. there is not a uniform evolution pattern from Round 1 to Round 2 for all three items and four countries.

Figure 1. *Percentage of children who could read a sentence*

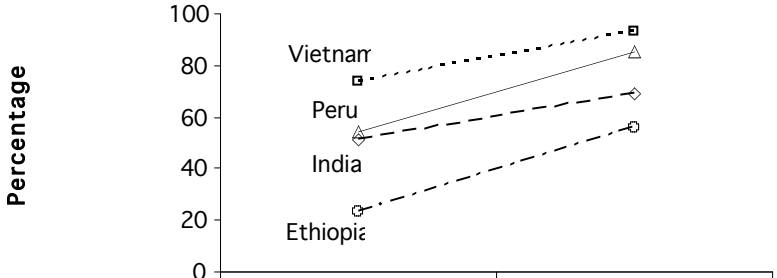


	Round 1	Round 2
Ethiopia	21	62
India	51	81
Peru	81	97
Vietnam	87	96

25 The items and scoring criteria are specified in section three of this report.

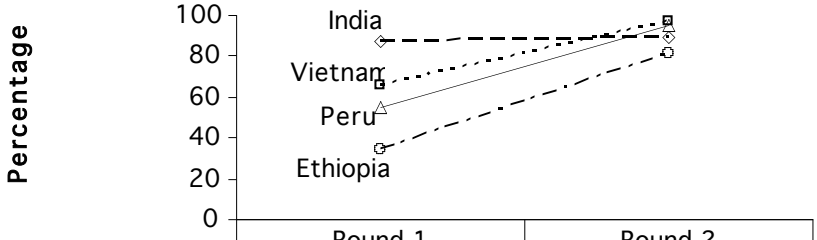
26 Additional tables reporting in further detail children's performance in the reading, writing and numeracy items in both Round 1 and 2 can be found in Annex 7.

Figure 2. *Percentage of children who could write without difficulty*



	Round 1	Round 2
Ethiopia	23	56
India	51	69
Peru	54	85
Vietnam	74	93

Figure 3. *Percentage of children who could multiply 2 times 4 correctly*



	Round 1	Round 2
Ethiopia	34	81
India	87	89
Peru	55	95
Vietnam	66	97

Note: In the case of India the results from Round 1 and 2 are not strict since the item in Round 2 was changed from 2 x 4 to 2 x 7.

As mentioned before, a ceiling effect was observed for all three items in all the countries (with the exception of Ethiopia, where the ceiling effect was observed in the numeracy item only). Additionally, we checked if this effect was uniform across genders and different languages. Tables 30 and 31 contain information about this matter. In the case of Ethiopia, the ceiling effect is observed for one language (Amarigna) but not for the other two (Tigrigna and Oromifa), except in the numeracy item where the ceiling effect is observed in all three languages. Regarding gender, boys and girls performed similarly.

Table 30. *Common items scores of Round 1 and Round 2 by country and language*

		N	Percentage of children who can read a sentence		Percentage of children who can write without difficulty		Percentage of children who can multiply 2 times 4	
			1st round	2nd round	1st round	2nd round	1st round	2nd round
Ethiopia	Amarigna	436	37	73	28	73	42	84
	Tigrigna	198	6	57	20	56	25	88
	Oromifa	159	3	55	19	33	17	73

Note: In the cases of India, Peru and Vietnam it was not possible to conduct the analysis for the other country languages because the number of observations was not enough (less than thirty).

Table 31. *Common items scores of Round 1 and Round 2 by country and gender*

		N	Percentage of children who can read a sentence		Percentage of children who can write without difficulty		Percentage of children who can multiply 2 times 4*	
			1st round	2nd round	1st round	2nd round	1st round	2nd round
Ethiopia	Female	475	22	65	24	55	29	81
	Male	476	21	59	23	58	38	81
India	Female	510	46	80	54	69	85	89
	Male	473	56	82	49	69	90	89
Peru	Female	313	81	97	58	89	52	94
	Male	366	80	96	51	83	58	95
Vietnam	Female	492	88	96	76	94	66	96
	Male	489	86	96	71	92	66	97

*In the case of India the results from Round 1 and Round 2 are not strictly comparable since the item in Round 2 was changed from 2 times 4 to 2 times 7.

We finally checked for the correlation among achievement items in Rounds 1 and 2. The analyses were done separately for each country and the information is presented in Tables 32 to 35. Three things should be noticed about these correlations. First, the correlation among achievement items is positive and statistically significant between items in the first and second round for each of the four participating countries, as would be expected. These correlations indicate that children who master the skill required to solve one item tend to have the necessary skills to respond correctly to the other items.

Second, when we look at the correlations between items from different rounds in the four countries, the score in reading in Round 1 has higher correlation values with reading and writing in Round 2. On the other hand, the score in numeracy in Round 1 seems to be weakly associated with performance in Round 2 (r below 0.3), including performance in the numeracy item.

Finally, correlation coefficients between common items were reduced from the first to the second round in most of the countries, mostly due to low variability in Round 2. Nevertheless they are still positive and statistically significant. This could be explained by the fact that most of the children were able to respond to all the achievement items correctly in Round 2 and this low variability could be causing the reduction in the coefficients from one round to another. The ceiling effect observed on most items, rounds and countries should be taken

into account in the interpretation of these results, given that it should have an effect on reducing the coefficient of the correlations.

Table 32. *Correlation among common items in each round and between rounds – Ethiopia*

		1st round			2nd round		
		Score in Reading	**Score in Writing**	**Score in Numeracy**	**Score in Reading**	**Score in Writing**	**Score in Numeracy**
1st round	Score in Reading	1					
	Score in Writing	0.73	1				
	Score in Numeracy	0.50	0.47	1			
2nd round	Score in Reading	0.32	0.28	0.22	1		
	Score in Writing	0.38	0.34	0.29	0.54	1	
	Score in Numeracy	0.20	0.20	0.18	0.32	0.39	1

Table 33. *Correlation among common items in each round and between rounds – India*

		1st round			2nd round		
		Score in Reading	**Score in Writing**	**Score in Numeracy**	**Score in Reading**	**Score in Writing**	**Score in Numeracy**
1st round	Score in Reading	1					
	Score in Writing	0.46	1				
	Score in Numeracy	0.39	0.26	1			
2nd round	Score in Reading	0.29	0.16	0.19	1		
	Score in Writing	0.36	0.24	0.26	0.56	1	
	Score in Numeracy	0.22	0.19	0.13	0.39	0.37	1

Table 34. *Correlation among common items in each round and between rounds – Peru*

		1st round			2nd round		
		Score in Reading	**Score in Writing**	**Score in Numeracy**	**Score in Reading**	**Score in Writing**	**Score in Numeracy**
1st round	Score in Reading	1					
	Score in Writing	0.55	1				
	Score in Numeracy	0.41	0.41	1			
2nd round	Score in Reading	0.26	0.18	0.14	1		
	Score in Writing	0.35	0.32	0.20	0.30	1	
	Score in Numeracy	0.23	0.16	0.16	0.37	0.29	1

Table 35. *Correlation among common items in each round and between rounds – Vietnam*

		1st round			2nd round		
		Score in Reading	**Score in Writing**	**Score in Numeracy**	**Score in Reading**	**Score in Writing**	**Score in Numeracy**
1st round	Score in Reading	1					
	Score in Writing	0.64	1				
	Score in Numeracy	0.38	0.41	1			
2nd round	Score in Reading	0.34	0.28	0.17	1		
	Score in Writing	0.37	0.29	0.16	0.49	1	
	Score in Numeracy	0.31	0.22	0.13	0.39	0.36	1

6. Final Considerations

In this report we have presented the psychometric characteristics of the three instruments that were used in Round 2 of YL to assess the cognitive development and assessment of children participating in the study. These were the Cognitive Developmental Assessment (CDA), the Peabody Picture Vocabulary test (PPVT) and the mathematics achievement test. The CDA and the PPVT were administered to the younger cohort and the PPVT and the mathematics achievement test were administered to the older cohort.

The psychometric properties of each of these tests have been analysed in this paper. The reliability and validity analyses were done separately for each cohort and country, and also within each country. The analyses were done for each main language used in the country, given that we found differential item functioning by language of the child. All the tests used in Round 2 of YL are norm-referenced, therefore it was expected that they would have a normal distribution. Nevertheless, not all the distributions of raw and Rasch scores by country and language were normal. Some results showed a ceiling effect, where many children scored near the highest possible score. For many of these, the test scores might be an underestimation of their abilities in the construct measured.

The reliability indexes of each of the tests for each country and language were calculated using both Classical Test theory (CTT) and Item Response theory (IRT). The reliability indexes for all the tests were acceptable for most samples according to the standards set (for CTT indexes, coefficients above 0.6 and for IRT indexes, coefficients above 0.5). The high reliability coefficients suggest that for each test the correlation between items is high and they are all measuring the same construct.

Regarding the validity of the instruments used in Round 2 of YL, both the correlations between test scores and education levels of the children's mothers and fathers as well as the comparison of mean test scores of groups of children with different educational levels are mostly positive and significant as expected. It was found in all the cases that more educated children perform better than children with less or none education. This constitutes an important source of evidence of the validity of the tests used in Round 2. However in some cases, age was also associated with scores, even though the variability in ages was relatively small. As age is also associated with grade in school, these two variables could be considered in some analysis of test scores as important confounders.

In order to increase the reliability and validity of the tests, some items with poor indicators (such as those with differential item functioning for boys and girls) were excluded from the scores, generating a new score for each child (both raw and converted into a Rasch scale). We also excluded from the analysis the scores of children who experienced problems in the administration of the test or who had difficulties in answering (e.g. due to problems in vision or hearing). A corrected database was generated with the original raw scores, corrected raw scores and, based on these, a Rasch score for all children.

In addition to the test mentioned above, reading, writing and numeracy assessment items administered to the children from the older cohort in Round 1 of YL were also administered in Round 2 to the same children. The results show that in the second round, most of the children in the sample were able to answer the reading, writing and numeracy items correctly and therefore they were of little use for most of the sample. However, there was a large positive gain between Rounds 1 and 2 for many children in Ethiopia.

In light of the reliability and validity results commented on above, we think the following should be taken into account when considering the inclusion of any of these tests in research or when reusing them to assess cognitive development and achievement in Round 3 of the YL project.

- CDA: in general terms, the psychometric properties of this test are acceptable for all the countries and languages analysed here. Nevertheless, this test would definitely not be useful in the next round of YL because it was originally targeted at preschoolers and by 2009 children in the younger cohort will be between 7.5 and 8.5 years old.
- Mathematics achievement test: in general terms, the psychometric properties of this test are acceptable for all the countries and languages analysed here. However, in most of the cases the distribution of test scores (both raw and Rasch scores) was found to be negatively skewed, showing a clear ceiling effect for many children. This indicates that this particular test would not be useful to assess children in Round 3 of YL, but a new mathematics test could be constructed by keeping some of the more difficult items and introducing new ones.
- PPVT – Younger cohort: the psychometric properties of this test are acceptable for all the countries and languages analysed here. The analyses showed that this test allows for differentiation between children with different achievement levels. It would be possible to use this test again in Round 3 of YL to assess verbal skills in the younger cohort.
- PPVT – Older cohort: the psychometric properties of this test are acceptable for all the countries and languages analysed here. Nevertheless, in all cases the distribution of test scores (both raw and Rasch scores) was found to be negatively skewed and leptokurtic, suggesting this test was too easy for many children who scored around the mean, which was located near the higher end of the distribution.

As mentioned at the beginning of this paper, we do not encourage international comparisons of results but suggest that the test results should be used for analysis within countries. Even within countries, separate analysis for each language may be necessary and appropriate.

The task set for the achievement and development instruments used in Round 2 of Young Lives was indeed difficult: tests should be valid and reliable, and include a normal distribution, for children in all four countries (even though populations were quite different). This had to be accomplished with a relatively small set of items in order to keep time of administration manageable. This report includes information on the extent to which these goals were accomplished, so that researchers interested in using YL data may select which scores to use and know its potential and limitations. For future rounds of YL, however, the number of items may well have to be increased in order to capture the variability of skills amongst children, which should be wider as time goes by.

References

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1999) *Standards for Educational and Psychological Testing*, Washington DC: American Educational Research Association

Baker, F. (2004) *Item Response Theory: Parameter Estimation Techniques* (2nd ed.), New York: M. Dekker

Campbell, J.M., S.K Bell and L.K. Keith (2001) 'Concurrent Validity of the Peabody Picture Vocabulary Test - Third Edition as an Intelligence and Achievement Screener for Low SES African American Children', *Assessment* 8.1: 85-94

Campbell, J.M. (1998) 'Review of the Peabody Picture Vocabulary Test - Third Edition', *Journal of Psychoeducational Assessment* 16.4: 334-8

Champion, T.B., Y.D. Hyter, A. McCabe and L.M. Bland-Stewart (2003) 'A Matter Of Vocabulary: Performances Of Low-Income African American Head Start Children on the Peabody Picture Vocabulary Test–III', *Communication Disorders Quarterly* 24: 121-7

Court, J.H. (1983) 'Sex Differences in Performance on Raven's Progressive Matrices: A Review', *Alberta Journal of Educational Research* 29: 54-74

Crocker, L. and J. Algina (1986) *Introduction to Classical and Modern Test Theory*, New York: Holt, Rinehart and Winston

Dunn, L. and L. Dunn (1997) *Examiner's Manual for the PPVT-III. Form IIIA and IIIB,* Minnesota: AGS

Dunn, L., E. Padilla, D. Lugo and L. Dunn (1986) *Manual del Examinador para el Test de Vocabulario en Imágenes Peabody (Peabody Picture Vocabulary Test) – Adaptación Hispanoamericana (Hispanic-American Adaptation)*, Minnesota: AGS

Grantham-McGregor, S., Y.B. Cheung, S. Cueto, P. Glewwe, L. Richter, B. Strupp and the International Child Development Steering Group (2007) 'Developmental Potential in the First 5 Years for Children in Developing Countries', *Lancet* 369.9555: 60-70

Gray, S., E. Plante, R. Vance, and M. Henrichsen (1999) 'The Diagnostic Accuracy of Four Vocabulary Tests Administered to Preschool-Age Children', *Language, Speech and Hearing Services in Schools* 30: 196-206

Gujarati, D. (2003) *Basic Econometrics* (4th ed.), New York: McGraw Hill

Hambleton, Ronald K., H. Swaminathan and H. Jane Rogers (1991) *Fundamentals of Item Response Theory*, Newbury Park, CA: Sage Publications

Koretz, D. (2008) *Measuring Up: What Educational Testing Really Tells Us*, Cambridge, MA: Harvard University Press

Linacre, Mike (2008) *A User's Guide to WINSTEPS MINISTEPS Rasch Model Computer Programs*, Chicago: MESA Press

Linacre, J.M. (2002) 'What do Infit and Outfit, Mean-Square and Standardized Mean?', *Rasch Measurement Transactions* 16.2: 878

Matlock-Hetzel, S. (1997) 'Basic Concepts in Item and Test Analysis', available at http://www.ericae.net/ft/tamu/Espy.htm (accessed 1 Dec 2008)

Oosterhof, A. (1990) *Classroom Applications of Educational Measurements*, Columbus, OH: Merrill

Restrepo, María Adelaida, Paula J. Schwanenflugel, Jamilia Blake, Stacey Neuharth-Pritchett, Stephen E. Cramer and Hilary P. Ruston (2006) 'Performance on the PPVT-III and the EVT: Applicability of the Measures with African American and European American Preschool Children', *Language, Speech, and Hearing Services in Schools* 37.1: 17-27

Stockman, I.J. (2000) 'The New Peabody Picture Vocabulary Test–III: An Illusion of Unbiased Assessment?', *Language, Speech, and Hearing Services in Schools* 31: 340-53

Ukrainetz, T.A. and C. Bloomquist (2002) 'The Criterion Validity of Four Vocabulary Tests Compared with a Language Sample', *Child Language Teaching and Therapy* 18: 59-79

Ukrainetz, T.A. and D.S. Duncan (2000) 'From Old to New: Examining Score Increases on the Peabody Picture Vocabulary Test–III', *Language, Speech, and Hearing Services in Schools* 31: 336-9

Washington, J.A. and H.K. Craig (1999) 'Performance of At-risk, African American Preschoolers on the Peabody Picture Vocabulary Test–III', *Language, Speech, and Hearing Services in Schools* 30: 75-82

Williams, K.T. and J. Wang (1997) *Technical References to the Peabody Picture Vocabulary Test - Third Edition,* Circle Pines, MN: American Guidance Service

THE AUTHORS

Santiago Cueto is Senior Researcher at GRADE, Lima.

Gabriela Guerrero is Adjunct Researcher at GRADE.

Juan Leon is a graduate student in Education at Penn State University.

Ismael Muñoz is a Research Assistant at GRADE

ACKNOWLEDGEMENTS

Special thanks to K. Mayuri (India), Teshome Nakatibeb (Ethiopia), Vu Thi Thu Thuy, Tran Thi Hien, Nguyen Viet Duc and Vu Hoang Dat (Vietnam), who participated in the pilot testing of the instruments reported here. Our thanks also to Dr. David Johnson from Oxford University, for his support in the Vietnam pilot test; and to Dr. Martin Woodhead from the Open University for his advice and support in the design of this component of the study. Finally, we would like to thank Dr. Patrice Engle and Dr. Richard Wolfe for their comments to a previous version of this report.

Department of International Development
University of Oxford
3 Mansfield Road, Oxford OX1 3TB, UK
Tel: +44 (0)1865 289966
Email: younglives@younglives.org.uk