Technical Note 32

Young Lives
An International Study of Childhood Poverty

Equating Children's PPVT Scores Across Survey Rounds and Age Cohorts in Peru

Juan León, Alejandra Miranda and Santiago Cueto

www.younglives.org.uk

JUNE 2015

Equating Children's PPVT Scores Across Survey Rounds and Age Cohorts in Peru

Juan León, Alejandra Miranda and Santiago Cueto

First published by Young Lives in June 2015

About Young Lives

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

Young Lives is funded from 2001 to 2017 by UK aid from the Department for International Development (DFID), It is co-funded by the Netherlands Ministry of Foreign Affairs from 2010 to 2014, and by Irish Aid from 2014 to 2015.

Funded by

UKaid
from the British people

Ministry of Foreign Affairs of the Netherlands

Irish Aid
An Roinn Gnóthaí Eachtracha agus Trádála
Department of Foreign Affairs and Trade

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

Tel: +44 (0)1865 281751 • E-mail: younglives@younglives.org.uk

Contents

Abstract

Young Lives gathers information from children and families through household and child questionnaires as well as children's cognitive and achievement tests. The Peabody Picture Vocabulary Test (PPVT) has been used in all survey rounds to date and with both age cohorts. This technical note presents the psychometric analysis performed (using Item Response Theory, IRT) in order to build cognitive measures comparable across Rounds 2 and 3 and the Younger and Older Cohorts for Peru. To achieve this, a one-parameter IRT model was used. We were able to identify a set of items across difficulty levels with good item fit as well as without item bias by gender and round. This set of items was used to equate the PPVT scores across rounds and age cohorts. Finally, we did an external validity analysis correlating the one-parameter PPVT scores with individual and family characteristics and the results showed that correlations were statistically significant with the expected signs.

The authors

Juan León is a PhD candidate in Educational Theory and Policy and Comparative and International Education at The Pennsylvania State University. He has a Bachelor's degree in Economics and a diploma in Liberal Arts from the Pontificia Universidad Catolica of Peru (PUCP). he is currently an Associate Researcher at the Group for the Analysis of Development (GRADE), Lima. Additionally, he is a Lecturer in the Department of Psychology at the Universidad Antonio Ruiz de Montoya in Lima and a Researcher at Young Lives.

Alejandra Miranda holds a Bachelor's degree in Economics from the PUCP. She is a Research Assistant at GRADE and for Young Lives.

Santiago Cueto holds a PhD in Educational Psychology from Indiana University. He has been a Visiting Researcher at the University of California at Davis and the University of Oxford. He is currently a Senior Researcher at GRADE, where he coordinates the Peru component of the Young Lives study. Additionally, he is a member of the National Education Council in Peru and Senior Lecturer in the Department of Psychology at the PUCP.

Acronyms and abbreviations

ANCOVA	analysis of covariance
ANOVA	analysis of variance
CDA	Cognitive Development Assessmen
CTT	classic test theory
DIF	differential item functioning
IRT	item response theory
MNSQ	mean square
PPVT	Peabody Picture Vocabulary Test
TVIP	Test de Vocabulario en Imágenes Peabody [Spanish version of PPVT]

1. Introduction

For longitudinal studies such as Young Lives, getting comparable measures of children's cognitive abilities over time is essential for identifying which child- family- and school-level variables affect children's development. Few longitudinal studies that follow birth/age cohorts have comparable cognitive measures across rounds; of those that do, the majority are from developed countries and almost none are from developing countries. For example, longitudinal studies such as the Early Childhood Longitudinal Study – Kindergarten (ECLS-K) or Education Longitudinal Study (ELS) in the USA have achievement measures (maths and reading comprehension) that are comparable across rounds (Najarian et al. 2009). Having comparable achievement measures facilitates the development of value-added or growth curve modelling analysis to identify variables at different levels (individual, family, school or community) associated with children's learning outcomes.

Our main goal in this technical note is to investigate how to build cognitive scores that are comparable across rounds and age cohorts for the Young Lives survey in Peru. Young Lives gathers information from children and families through household and child questionnaires as well as cognitive and achievement tests for children. The only common test across rounds and cohorts is the Peabody Picture Vocabulary Test (PPVT); therefore, we have used PPVT data to build cognitive measures that are comparable across Rounds 2 and 3 and across the two age cohorts.

To address our goal of comparable measures, we have used item response theory (IRT) to get standardised cognitive measures. As a first step, we estimated the scores using the one-parameter model (or Rasch model), which uses item difficulty as a parameter to estimate child ability. Our second step was to perform a differential item functioning (DIF) analysis by gender, cohort, and round to identify possible item bias that could be corrected. The last step consisted of equating the scores using common item equating (anchor items) as a means of obtaining comparable PPVT scores across rounds and cohorts.

In this technical note, we have five sections. After this brief introduction, we present a short literature review about the uses of the PPVT in other studies. The third section describes the methodology of analysis, and the fourth explains the main results of the analysis performed. Finally, the last section gives some final remarks about the use of the cognitive measures.

2. The use of PPVT scores in research studies

The PPVT was designed originally to measure children's English vocabulary. The examiner shows the child a set of four pictures simultaneously and asks him or her to select the picture that best represents the word spoken by the examiner. By 1986, a Spanish adaptation of the PPVT became available, named *Test de Vocabulario en Imágenes Peabody* (TVIP) and developed by L. Dunn, E. Padilla, D. Lugo and L. Dunn (Dunn et al. 1986). Like the English version of the PPVT, the Spanish equivalent has been used to measure children´s vocabulary skills in several research studies. However, one of the main limitations of the different studies reviewed is that none used the same PPVT or TVIP measure to compare children's vocabulary abilities over time; rather, the studies used the scores to compare the strengths of relationships.

Umbel et al. (1992) used the English and Spanish versions of the PPVT with Hispanic children to test the extent to which the language used at home affected their vocabulary skills in Spanish and English. The study was carried out with 105 Hispanic children from four public schools in Dade County, USA. The main findings of this study were that there were no differences in the PPVT scores in Spanish of children from Spanish-only and bilingual (English and Spanish) home environments; however, PPVT scores in English showed a difference of 1 SD in favour of children from bilingual homes.

Vogel et al. (2006) carried out a study to explore the relationship between a father´s presence at home and children's early development outcomes (cognitive, linguistic and socio-emotional measures) in low-income families in three ethnic groups: European American, African American and Latin American. The authors used data from the Head Start study in the USA. To measure linguistic aptitude, the study used the Peabody Picture Vocabulary Test III (PPVT-III)[1] for European and African American children and the TVIP for Latin American children. The main findings suggest that frequent contact with the biological father is beneficial for child outcomes but that the effects differ between ethnic groups.

Long (2012) used the Head Start study datasets to examine the effect of parental involvement (measured by home stimulation in language and cognition and emotional supportiveness) on language development among Hispanic children from low-income families at 3 years old. To test children's language development level, the PPVT-III was used for English vocabulary and the TVIP for Spanish vocabulary. The author divided the sample into two groups depending on the language used at home by the children: mainly Spanish or mainly English and used path analysis models. The main findings suggest that parental involvement predicts language development (vocabulary) in English and Spanish at the age of 3. However, the effects of the measures used were different for the two groups. For children whose home language environment was mainly English, parental emotional supportiveness had a positive effect on the PPVT scores in Spanish and English, while in

1 The third version was released with the same procedure as the others two, but with 204 items. It was developed by Dunn and Dunn in 1997.

families where Spanish was mainly used, home stimulation in language and cognition predicted language development at the age of 3 in Spanish and English.

Chien et al. (2010) explored the connection between children's classroom engagement and school readiness gains in pre-kindergarten (pre-school or early years settings). They classified children into four profiles of classroom engagement: free play, individual instruction, group instruction and scaffolded learning. Then they explored whether these profiles were linked to gains in school readiness during the pre-kindergarten year using the Emerging Academics Snapshot, a measure of children's classroom engagement that captures their moment-to-moment activities. The authors used two datasets: (i) The National Center for Early Development and Learning Multi-State Study of Pre-Kindergarten, and (ii) the follow-up study of the State-Wide Early Education Programs Study (SWEEP). Both studies tested children's vocabulary using the PPVT and TVIP as measures for school readiness and the sample included children from different ethnic backgrounds. The authors used analysis of variance (ANOVA) and analysis of covariance[2] (ANCOVA) to explore the effects of classroom engagement measures on school readiness. The results indicated that there were no statistically significant differences across the four child profiles as far as TVIP and PPVT scores were concerned.

Paxson and Schady (2005) explored the cognitive development of youth in Ecuador, using TVIP age-normed scores as cognitive development measures. They found that children's cognitive development was associated with socioeconomic status, child health and adequate parenting styles (measured by the degree to which parents are responsive or harsh toward children); the same results were found even using the raw TVIP scores. Given these results, the authors recommended setting up programmes to work on these three things, such as cash transfer programmes or early childhood development programmes like Head Start in the USA.

Finally, Lopez Boo (2013) explored socioeconomic disparities and cognitive gains before and after the early school years in different countries. She used data from the Younger Cohort of the Young Lives study and compared the results across countries. She used value-added analysis in order to compare the cognitive gains across countries. The main findings of her study showed Peru as the country with the largest socioeconomic disparities as well as the one with the highest persistence in cognitive development. However, one main difference between this study and previous ones is the use of raw scores instead of standardised scores. The use of raw scores provides comparable measures within a country although the standardised raw scores must be adjusted by child age.

In sum, different studies have used PPVT or TVIP measures for receptive vocabulary or language development but only one of them has (to our knowledge) tried to address the issue of having comparable tests scores across ethnic groups or time points in order to have accurate measures of children's cognitive development. Most of the studies used standardised scores to compare the results across ethnic groups. However, this is problematic because the TVIP standardised scores are outdated. The normalisation sample of the TVIP standardised scores was collected at the end of the 1980s in contrast with PPVT-III normalisation sample which was from the 2000s. Only Lopez Boo (2013) makes an effort

2 The variables used for the ANCOVA analysis were household size, poverty status, maternal education, gender, age and ethnicity.

to have comparable measures within each country; however, she used classic test theory (CTT) assumptions that do not necessarily give accurate comparable measures.

3. Methodology

For this technical note, we do not use TVIP age-normed scores,[3] for three main reasons. First, the normalisation sample for the normed scores dates from late 1980s. Second, the reference group is composed of Mexican and Puerto Rican children, who are not strictly comparable with children from Peruvian contexts. And, third, possible ceiling effects could occur since the TVIP age range goes from 2.5 to 17 years old and we have children who are 15 years old in our sample. Therefore, the use of normed scores could lead to biased interpretation and estimation of children's receptive vocabulary levels, as well as possible ceiling effects, since some items could be too easy for a group of children in our sample. However, it is necessary to point out that a Spanish version of the PPVT-III is available at TEA editions.[4] This Spanish translation has 192 items and the age range for vocabulary abilities goes from 2 to 90 years old but this test was not used for the Young Lives study.

3.1. The Young Lives study

Young Lives is a longitudinal study of childhood poverty that tracks the development of 12,000 children in Ethiopia, India (in the state of Andhra Pradesh), Peru and Vietnam over 15 years. Young Lives has been following two cohorts (one born in 1994 and other in 2001) since the beginning of the study, in 2002. In Peru, the original sample was chosen randomly from 20 sites across the country; however the 5 per cent richest districts were excluded from the sampling framework. Up to now, Young Lives has carried out four rounds of data collection (administered in 2002, 2006, 2009 and 2013); currently the data-cleaning process for Round 4 is still taking place. Young Lives gathers information on economic indicators, work patterns, the access to services of children and their families, children's nutritional status and children's educational progress.

In order to identify which variables (i.e. child, family and school characteristics) affect children's development, and to what extent, it is necessary to have comparable measures of children's cognitive abilities over time. Table 1 shows the measures of ability and achievement administered during Rounds 1 to 3 by round and cohort. As the table shows, the only common test administered across rounds was the PPVT; therefore, to build cognitive measures comparable across Rounds 2 and 3 and across the Younger and Older Cohorts, we are using the PPVT.

3 The age-normed scores are standardised raw scores based on samples from Mexican and Puerto Rican children.

4 More details about the PPVT-III Spanish version could be found at http://web.teaediciones.com/peabody-test-de-vocabulario-en-imagenes.aspx

Table 1. *Measures of ability and achievement administered in Young Lives, Rounds 1 to 3*

Round	Cohort	Cognitive	Reading	Mathematics
Round 1	Younger Cohort (age 1)	-	-	-
	Older Cohort (age 8)	Raven's Progressive Matrices for children	One item on reading and one on writing	One multiplication item
Round 2	Younger Cohort (age 5)	PPVT	-	Cognitive Development Assessment (CDA)
	Older Cohort (age 12)	PPVT	One item on reading and one on writing	One multiplication item and maths test
Round 3	Younger Cohort (age 8)	PPVT	One item on reading and one on writing Early Grade Reading Assessment (EGRA)	One multiplication item and maths test
	Older Cohort (age 15)	PPVT	Cloze test of reading comprehension	Maths test

3.2. Data

We used the data collected in Rounds 2 and 3 of the Young Lives survey in Peru, specifically the PPVT data. Our sample sizes for each round and cohort are presented in Table 2, which shows the full sample (all children tested in the TVIP in Rounds 2 and 3) and the sample analysed by us for this paper; the main difference between these is that the samples analysed exclude children who took the PPVT in Quechua, given the small sample size.[5]

Table 2. *Number of observations by type of sample and age cohort in Rounds 2 and 3*

	Round 2		Round 3	
	Full sample	Sample analysed	Full sample	Sample analysed
Younger Cohort	1,907	1,716	1,902	1,832
Older Cohort	673	656	667	666

Source: Young Lives, Rounds 2 and 3.

Table 3 shows the main background characteristics of children in the full sample and the panel sample.[6] We observe that there are no big differences between the full sample and the panel one. Both groups have the same child characteristics, such as percentage of girls, average age and parental background (i.e. parent's educational level).

[5] The rule of thumb according to Gorsuch (1983) and Bryant and Yarnold (1995) is to have a ratio of 5 between subjects and items. Ratios lower than 5 could give biased estimates.

[6] The full sample refers to those children who have TVIP scores in either round while the panel sample refers to those children who have TVIP scores for both rounds.

Table 3. *Main background characteristics of the Peruvian children in Round 3 by cohort and type of sample (full and panel sample)*

	Full sample		Panel sample	
	Younger Cohort	Older Cohort	Younger Cohort	Older Cohort
Average age (months)	94.9	178.7	94.9	178.6
Wealth Index, Round 3[a]	0.55	0.52	0.57	0.52
Female (%)	49.6	46.9	49.7	47.0
Mother completed secondary education (%)	57.4	54.1	61.9	55.5
Indigenous mother tongue (%)	10.3	10.3	4.1	8.3
Total[b]	1,809	659	1,646	634

[a] This is the simple average of three composite scores: Housing Quality Index, Consumer Durables Index and Services Index. Each composite score ranges from 0 to 1.

[b] Excludes children with missing values for mother's education, wealth index and mother tongue.

Source: Young Lives, Round 3.

3.3. TVIP administration and item coding for the analysis

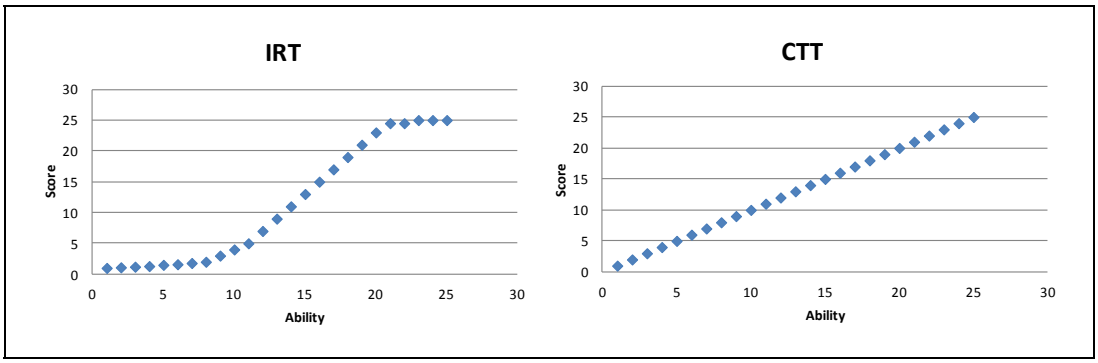
One main characteristic of the TVIP is that children do not have to answer all items since it has rules for when administrators should start and stop it. The test aims first to determine the basal vocabulary level for each examinee. This level is determined as the highest set of eight consecutive items answered correctly by the examinee. If a basal set of items is not determined, the first item answered correctly is taken as basal item. When a basal set of items is determined, children continue answering the items above this set until they have six errors in eight consecutive items, and the last item is taken as ceiling item. Thus, the logic of the TVIP is that any item below the basal set is assumed to be correct since the examinee has higher chances to get it right, while any item above the ceiling item is assumed to be incorrect since the chances of getting it wrong are high. Therefore, for this technical note, we coded all items below the basal set for each child as 1 (or correct) and any item above the ceiling item as 0 (or wrong).

3.4. Why IRT scores instead of CTT scores?

Unlike the CTT, the IRT is more focused on the item rather than the test. Additionally the standard error of measurement in the IRT is a function of the ability of individuals; thus it varies at each level of ability, and nonetheless, the interpretation is the same. The IRT estimates the probability of answering the item correctly through a logistic function based on the difference between the item difficulty and the individual's ability. The idea is that individuals with higher ability will have a greater probability of answering easier items correctly than difficult ones.

Figure 1 shows the relationship between an individual's ability and their score according to the CTT and the IRT. In the case of the CTT, we observe that the raw score increases in the same proportion as the ability; thus it follows a linear and monotonic trend. In contrast, in the IRT we observe that as the ability increases, the score does not increase in the same proportion; in other words the growth in scores is non-linear. This implies that, under the CTT, the rate of change is the same if ability changes from 10 to 15 as from 20 to 25; however, under the IRT it is not the same since it follows a functional form that relies on the characteristics of the items.

Figure 1. *Relationship between scores and ability according to the IRT and the CTT*



In order to build the children's composite scores for the PPVT in Peru, we used the IRT. The main advantages of using this statistical technique instead of CTT are: (i) the principle of invariance – the item parameters do not depend on an individual's ability, being invariant over different samples of examinees, and an individual's ability does not depend on the items presented, being invariant over different samples of items; (ii) allowance for comparing the ability of individuals from different populations if tested with instruments that have common items; and (iii) allocation of an individual's ability and item difficulty in the same scale or metric, which creates an interval scale in logits for both scores. Thus, using this statistical technique, we were able to build comparable scores by cohort and round.

3.5. The IRT model

The IRT models rely on two main assumptions. One is the local independence assumption, which asserts that the probability of an individual answering an item correctly depends on his/her ability only and not on his/her answer to other items. Second, the model assumes unidimensionality. In other words, it assumes that only one latent trait is measureable across all items or at least one dominant factor is observed behind the set of items tested. Of these two assumptions, the latter is the most difficult to demonstrate since different factors could be affecting the individual performance (e.g. test anxiety).[7]

IRT has three different models, which we detail below:

- **One-parameter model or Rasch model:** this is the most popular of the three. It assumes that an individual's ability depends on item difficulty and that all items have the same level of discrimination. Item discrimination refers to how well an item discriminates between high and low achievers in the test. The one-parameter model and the Rasch model are mathematically equivalent and the equation for this model is:

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \qquad i = 1,2,\dots,n$$

$P_i(\theta)$: the probability that an individual with ability θ get right the item i

b_i : the item difficulty

[7] For further information, see Cueto et. al. (2009)

n : the number of items in the test

θ : the individual's ability parameter

- **Two-parameter model:** this model assumes that an individual´s ability depends on two item parameters – item difficulty and item discrimination. This model allows for different levels if item discrimination. The equation for this model is:

$$P_i(\theta) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \qquad i = 1,2, \dots, n$$

$P_i(\theta)$: the probability that an individual with ability θ get right the item i

a_i : item discrimination

b_i : the item difficulty

n : the number of items in the test

θ : the individual´s ability parameter

- **Three-parameter model:** this model assumes that an individual's ability depends on three item parameters – item difficulty, item discrimination, and item guessing. This last parameter refers to the chances that an individual with low levels of ability has to get an item right. This model is mainly considered for multiple choice tests since these allow examinees to guess. The equation for this model is:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \qquad i = 1,2, \dots, n$$

$P_i(\theta)$: the probability that an individual with ability θ get right the item i

a_i : item discrimination

b_i : the item difficulty

c_i : guessing parameter

n : the number of items in the test

θ : the individual´s ability parameter

While any of the three IRT models may help estimate the probability of answering an item correctly, for this technical note we decided to employ the one-parameter model or Rasch model, given the findings by van de Vijver (1986) about the robustness of Rasch estimates. His study found: (i) different values of item discrimination do not affect Rasch accuracy estimates, (ii) the presence of item guessing could harm Rasch estimates but guessing could be present in any type of answer format (open-ended or multiple choice) and its detection does not depend on psychometric properties, (iii) two- and three-parameter models rely less on formal fit statistics than do Rasch or one-parameter models, and (vi) this analysis contributes to the normalisation of test curves.

3.6. Item fit analysis

The Rasch model uses 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 one-parameter model). A ratio of 1 indicates that the observed

variance equals expected variance; ergo, the error of measurement is almost zero or inexistent. When the MNSQ fit statistics value is greater than 1.0 (for example, 1.70) there is 70 per cent more variation in the observed variable than the one-parameter model predicted. When the fit statistics value is less than 1.0, there is less variation in the observed variable than the one-parameter model predicted.

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

- **Outfit mean square:** this 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 idea that there are individuals with higher ability who are answering incorrectly items that should have been easy for them, and vice versa. Problems with this indicator, although they should be considered, do not represent a serious threat to the reliability of the test (Linacre 2008).
- **Infit mean square:** this 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 how large the residuals in the estimated model are. Problems with this indicator indicate a threat to the reliability and also to the validity of the test (Linacre 2008).

These two statistics, when calculated, can be interpreted as follows: >2.00: off-variable noise is greater than useful information; 1.50–1.99: noticeable off-variable noise; 0.50–1.50: productive of measurement (and <0.50: overly predictable). In sum, infit and outfit values between 0.50 and 1.50 indicate that the item has a good fit.

Other fit statistics considered are the person and item reliability indexes. If the person reliability index is similar or equivalent to test reliability measures (e.g. Cronbach's alpha) then low values indicate a poor range of person abilities or short test length. Person reliability indexes above 0.80 indicate good person ability discrimination. The item reliability index is a new concept that has no equivalent in CTT. Poor item reliability indicates that we have a poor range of item measures (same level of difficulty) or small sample size, and item reliability measures above 0.80 indicate good item discrimination.

Finally, the last fit index used is the DIF analysis. An item is considered to have DIF if the probability of answering an it correctly differs across groups or memberships (e.g. gender), controlling for level of ability (Linacre 2008). DIF analysis, however, could be sensitive to sample size since the standard errors of the item difficulty depend on the size of the groups that are being compared. Thus, large sample sizes could lead to accept even small differences between item difficulties as DIF. Therefore, it is necessary to use normalised standard errors in order to have better estimates of DIF between groups. The Educational Testing Service in the USA as well as different scholars (Wright and Douglas 1976) suggest that for large sample sizes logit differences in item difficulty above 0.50 are signals of DIF between groups.

3.7. Equating

As mentioned before, one of the main advantages of using IRT modelling is that it helps to build comparable scores using common items. Hambleton (1989) indicates that if we have different tests (common items across them) and the items of those tests meet the IRT assumptions (good item fit indicators), then it is possible to estimate a score for each

individual that is independent of the group of items that he/she answered. Thus, it is possible to use those PPVT items with adequate fit index as anchors in order to have a score that could be comparable across rounds and cohorts.

The main types of test equating are as follows (Linacre 2008):

- **Common item equating:** there are different examinees but common items across all tests forms. Two different types of analysis could be performed. First, the common and non-common items could be analysed simultaneously (e.g. equating tests for different school grades). Second, common items across all test forms are analysed and calibrated in order to use them to adjust the mean and standard deviation of each est form.
- **Common person equating:** there are different tests in the same subject (e.g. maths) but common examinees across tests. The average ability of the common examinees is used to adjust examinees' mean and standard deviations.
- **Virtual item equating:** there are different examinees and two different tests but both tests cover the same subject (e.g. maths). This type of equating involves identifying test pairs of items that cover the same subject and using them as pseudo-anchor items for the equating analysis.

For our analysis, we have used the common item equating approach since we have the same test across cohorts and rounds. It is not possible to use common person equating since having the scores of the same examinee at two different time points is similar to having different examinees.

Finally, the subsequent procedures for the equating analysis are to: (i) run separate Rasch analysis (Younger Cohort, Older Cohort and pooled sample), checking for the item fit, (ii) identify those items with poor item fit, deleting them from the analysis, (iii) identify those items with presence of DIF and consider them as different items, and (iv) run the Rasch analysis again, using as anchor items those with the absence of DIF by gender, round and age cohort for each age cohort and pooled sample. The Rasch analysis was carried out using WINSTEP version 3.68.2.

4. Results

This section presents the results of the analysis performed using the raw and Rasch scores for the TVIP.

4.1. Raw scores

The first measures that are comparable across rounds and cohorts are the TVIP raw scores. Given the logic of the tests, the raw scores estimate the level of vocabulary of each examinee. Thus, comparing raw scores across rounds gives us an idea of learning over time, but this measure has limitations, such as possible item bias.

Table 4 shows the mean raw scores by round and age cohort. Across rounds Younger Cohort children increase their vocabulary by about 29 points and the Older Cohort by 11 points . These results indicate that children's learning rate is not necessarily linear or monotonic; instead, we have non-linear learning growth, as different longitudinal research studies have found (LoGerfo et al. 2005; Cheadle 2008; León and Cueto 2013). This effect could be related to possible ceiling effects in the children's receptive vocabulary since Older Cohort children were 15 years old in Round 3 and the TVIP measures receptive vocabulary up to 17 years old.

Table 4. *Raw scores in the TVIP by round and age cohort*

	Round 2	Round 3
Younger Cohort	31.15	60.13
	(0.44)	(0.39)
	[1,716]	[1,832]
Older Cohort	85.97	96.78
	(0.74)	(0.67)
	[656]	[666]

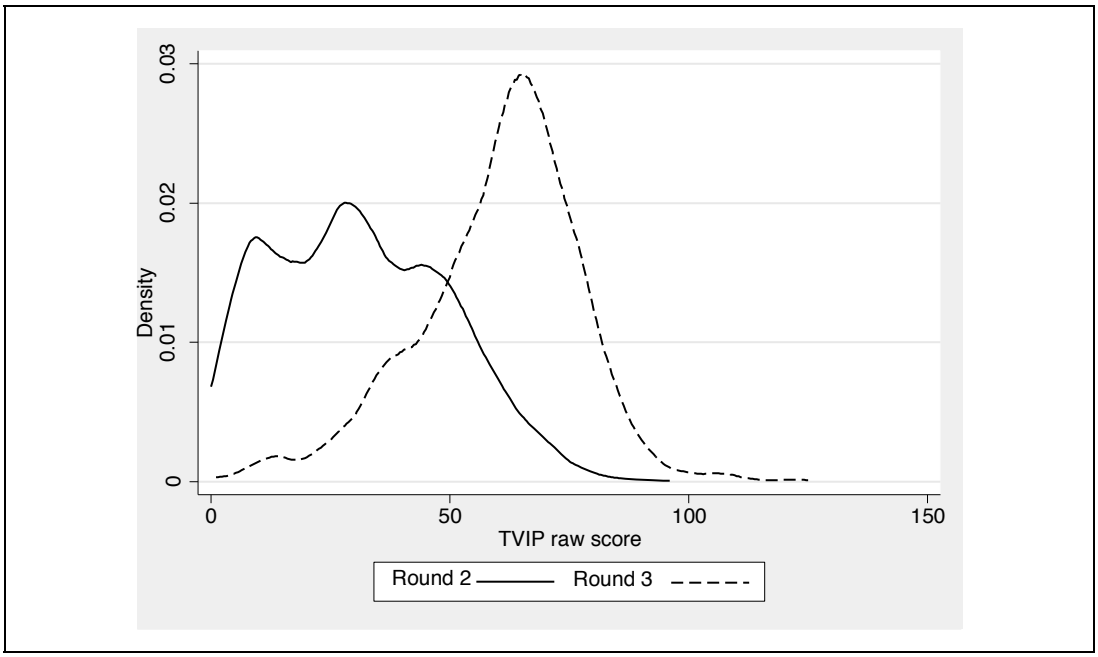
Source: Young Lives, Rounds 2 and 3.

Note: Standard error in round brackets and number of children reported in square brackets.

Figures 2 and 3 shows the raw score distribution for the Younger and Older Cohorts respectively. Both figures show the shift in the distribution of the scores between Rounds 2 and 3, with the shift for the Younger Cohort being larger. In terms of the normality of the raw scores, the four measures do not follow a normal distribution; even for some scores a multimodal[8] distribution can be observed.

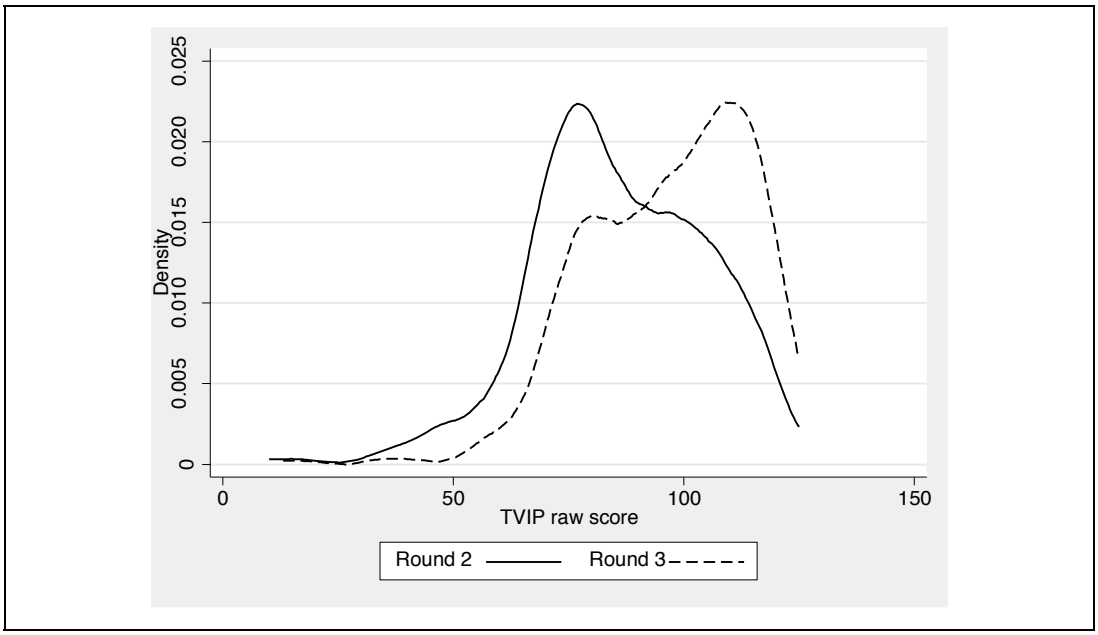
[8] More than one peak in the variable distribution.

Figure 2. *Distribution of raw TVIP scores for the Younger Cohort*



Source: Young Lives, Rounds 2 and 3.

Figure 3. *Distribution of raw TVIP scores for the Older Cohort*



Source: Young Lives, Rounds 2 and 3.

Finally, we tested whether the raw scores followed a normal distribution. As we can see in Table 5, all the scores are non-normally distributed.

Table 5. *Characteristics of the raw score distribution*

	Younger Cohort	Older Cohort
Round 2	NS/LK	NS/LK
Round 3	PS/PK	NS/LK

ND = Normal distribution, NS = Negatively skewed, PS = Positively skewed, LK = Leptokurtic, PK = Platikurtic.

4.2. Rasch scores

As mentioned above, using raw scores may result in biased estimates; therefore we estimated the Rasch scores for the TVIP. We estimated three different Rasch models. The first two models equate the TVIP scores across rounds for the Younger and Older Cohorts, and the last Rasch model equates the TVIP scores by cohort and rounds (pooled). Table 6 shows the reliability indexes for the three Rasch models estimated. We can see that item and child reliability indexes are close to 1.00, ensuring an adequate internal consistency at item and child level.

Table 6. *Reliability indexes for the Rasch models (separation index)*[a]

	Child	Item
Younger Cohort	0.99	1.00
Older Cohort	0.98	0.99
Pooled sample	0.99	1.00

[a] The child reliability or separation index for Rasch models is similar to the Cronbach's alpha or Kuder-Richardson 20 index in CTT.

Source: Young Lives, Rounds 2 and 3.

Table 7 shows the correlation matrix of the item difficulty among the three models. All correlations are statistically significant and their effect sizes above 0.98. These results indicate that it does not matter what model specification we choose, the item order or ranking remain the same across them.

Table 7. *Correlation matrix of item difficulty across models (p-value)*

	Younger Cohort	Older Cohort	Pooled sample
Younger Cohort	1.00		
	(0.00)		
Older Cohort	0.98	1.00	
	(0.00)	(0.0)	
Pooled simple	0.99	0.99	1.00
	(0.00)	(0.00)	(0.00)

Source: Young Lives, Rounds 2 and 3.

Also, we checked the item fit for each of the three models. We checked each item's infit mean square (see Tables A1 and A2 in the Appendix) and present the results in Table 8.[9] None of the items of the TVIP for the Younger Cohort and the pooled model showed poor fit, but five items in the Older Cohort Rasch model showed poor item fit.

Table 8. *TVIP items flagged with poor item fit for each Rasch model*

	Items with infit mean square above 1.5 or below 0.5
Younger Cohort	None
Older Cohort	Item 2, Item 5, Item 7, Item 32 and Item 80
Pooled sample	None

Source: Young Lives, Rounds 2 and 3.

Finally, we checked for item bias by gender, round and cohort. For models 1 and 2, we checked for item bias by gender and round, while for model 3 we checked for item bias by gender, round and cohort. Table 9 shows the number of items flagged with bias for each type or combination of types.[10] For models 1 and 2 around half or more of the items do not show item bias by gender or round. However for model 3, the number of items without bias was around one-third of the total, but these items cover a wide range of item difficulty, which facilitates the item equating since we have items at different levels of difficulty.

Table 9. *Number (and percentage) of TVIP items flagged with bias for gender, round and cohort*

	Younger Cohort	Older Cohort	Pooled sample
Gender	12 (10%)	21 (17%)	14 (11%)
Round	34 (27%)	11 (9%)	16 (13%)
Cohort			23 (18%)
Gender and round	12 (10%)	2 (2%)	6 (5%)
Gender and cohort			14 (11%)
Round and cohort			11 (9%)
Gender, round and cohort			8 (6%)
Without bias	67 (54%)	91 (73%)	33 (26%)

Source: Young Lives, Rounds 2 and 3.

Once biased items were identified, we proceeded to use the items without bias as anchor items across rounds for the Younger and Older Cohort models, and we used non-biased items as anchors across rounds and cohorts for the pooled model. The items flagged with bias were split and considered as new items for each group. Therefore, the final equated models, instead of having 125 items, count with more items, as shown in Table 10.

9 We did not use the outfit mean square fit indicator since it does not represent a serious threat to the reliability of the test (Linacre 2008).

10 See Tables A3, A4 and A5 in the Appendix for details of the items flagged with bias for each of the characteristics presented in Table 9.

Table 10. *Number of TVIP items by model before and after equating*

	Before equating	After equating[a]
Younger Cohort	125	207
Older Cohort	125	163
Pooled sample	125	327

[a] There is an increase in the number of the overall items, given the splitting method, but all children only have 125 items

Source: Young Lives, Rounds 2 and 3.

Then, we run the new Rasch models with the adjustment for the new number of items, using as anchor items those items that did not show any type of bias by gender, round or cohort. Table 8 shows a summary of the number of items flagged with poor item fit using this approach. As shown, none of the items in the three models showed poor item fit.

Table 11 shows child and item reliability indexes after equating. Reliability indexes are high, ranging from 0.98 to 0.99 for child reliability and from 0.99 to 1.00 for item reliability. These results, as with those shown before, indicate the good internal consistency of the item difficulty and child ability measures obtained with the Rasch analysis.

Table 11. *Reliability indexes for the Rasch models (separation index) with final equated models[a]*

	Child	Item
Younger Cohort	0.99	1.00
Older Cohort	0.98	0.99
Pooled	0.99	1.00

[a] The child reliability or separation index for Rasch models is similar to the Cronbach's alpha or KR20 in CTT.

Source: Young Lives, Rounds 2 and 3.

Table 12 shows the mean Rasch equated scores. The scores were centred on the mean of the Younger Cohort in Round 2. As we saw with the raw scores, the equated Rasch scores show that the Younger Cohort learnt more vocabulary between Rounds 2 and 3 than the older one. The Younger Cohort increased their scores by 1.53 SD while the Older Cohort increased theirs by 0.55 SD.

Table 12. *Rasch scores equated by round and cohort*

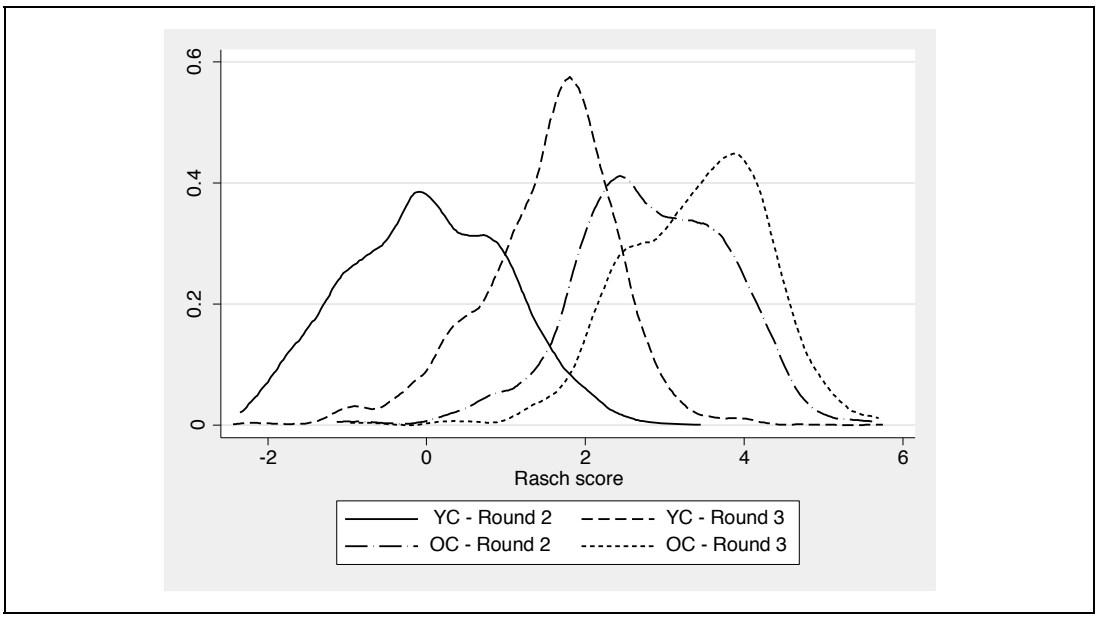
	Sample analysed		Panel sample	
	Round 2	Round 3	Round 2	Round 3
Younger Cohort	0.00	1.53	0.00	1.62
	(0.024)	(0.020)	(0.024)	(0.020)
	[1,715]	[1,832]	[1,645]	[1,645]
Older Cohort	2.83	3.38	2.84	3.40
	(0.037)	(0.034)	(0.038)	(0.035)
	[656]	[666]	[634]	[634]

Note: Standard error in round brackets and number of children reported in square brackets. The scores were centred on the mean for the Younger Cohort in Round 2.

Source: Young Lives, Rounds 2 and 3.

Also, Figure 4 shows the score distribution at the same scale since all the scores are now comparable. The graph shows the same results as Table 12: children from the Younger Cohort show a higher displacement of the score distribution than children from the Older Cohort,. This could be accounted for by possible ceiling effects, as we mentioned before, since the TVIP measures receptive vocabulary for children up to 17 years old, and some of the children from the Older Cohort (aged 15 years old) could be finding some items too easy.

Figure 4. *Distribution of the Rasch equated scores by round and cohort using the pooled sample*



Source: Young Lives, Rounds 2 and 3.

We then tested whether the Rasch equated scores followed a normal distribution pattern. As Table 13 shows, the Rasch equated scores for the Older Cohort are non-normally distributed.

Table 13. *Characteristics of the Rasch equated score distribution*

	Younger Cohort	Older Cohort
Round 2	PK	NS/LK
Round 3	NS/LK	NS/LK

ND = Normal distribution, NS = Negatively skewed, PS = Positively skewed, LK = Leptokurtic, PK = Platikurtic.

Finally, we calculated the correlation between the equated scores and the previous scores calculated for Rounds 2 and 3 (Table 14).[11] The correlation between the equated scores and previous estimated scores is almost perfect and statistically significant.

11 We used the calculated scores for Rounds 2 and 3 that are available in the international dataset at the UK Data Service (http://discover.ukdataservice.ac.uk/). Also, for details about the procedures followed to build previous scores, see Cueto et al. (2009) and Cueto and León (2013).

Table 14. *Correlation between Rasch equated scores and previous Rasch scores estimated in Rounds 2 and 3*

Younger Cohort – R2	0.99 (0.00)
Younger Cohort – R3	0.99 (0.00)
Older Cohort – R2	0.99 (0.00)
Older Cohort – R3	0.95 (0.00)

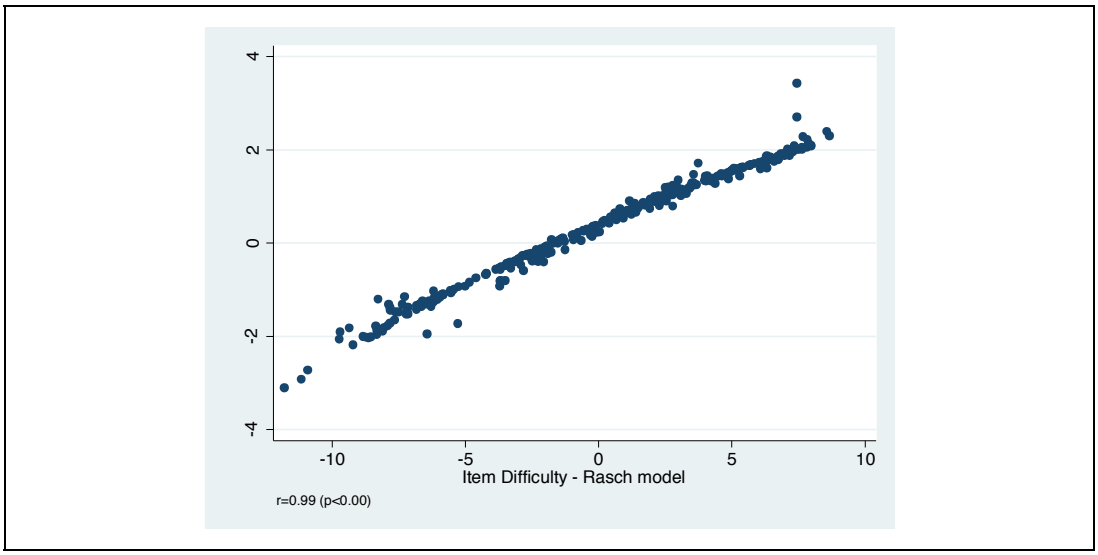
Source: Young Lives, Rounds 2 and 3.

4.3. Comparing Rasch scores with three-parameter IRT scores (pooled model)

As a final exercise, we calculated the TVIP scores using a three-parameter IRT model to validate the results obtained with the Rasch model used in the previous section.[12] We also wanted to take into consideration possible differences in discrimination (different slopes) across items as well as the probability of the child guessing (different intercept), given that the TVIP is a multiple choice test. As we mentioned in Section 3.5, the Rasch model considers item discrimination to be equal across items and assigns a value of zero for item guessing.

First, we correlated the item difficulty estimated by both models. As we could observe, in Figure 5, the results are pretty similar and the correlation between these two scales is 0.99 ($p<0.01$). It indicates that taking into consideration possible differences in item discrimination and the probability of children guessing does not affect the item difficulty.

Figure 5. *Scatterplot of the estimated item difficulty using Rasch model and three-parameter IRT model (pooled sample)*

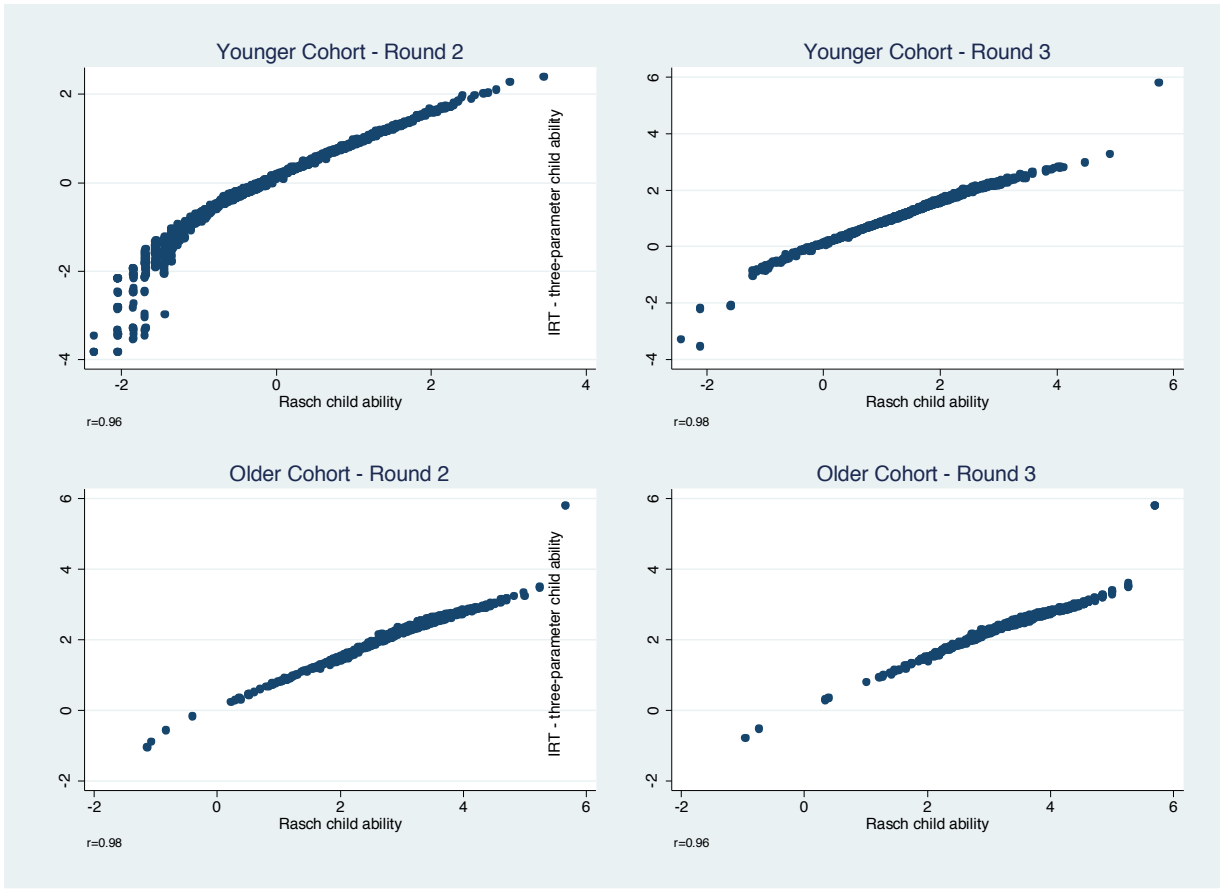


Source: Young Lives, Rounds 2 and 3.

12 To estimate the three-parameter model, we used the openirt ado file developed for the STATA software.

Finally, we correlated the Rasch and three-parameter equated scores in order to check for possible differences. In Figure 6, we could observe that the correlation between these two sets of scores by round and age cohort is above 0.95 ($p<0.01$) in all cases. This result indicates that both models give us the same information about the children's ability in the TVIP.

Figure 6. *Comparing child's ability Rasch equated scores with three-parameter IRT equated scores for the pooled sample, by round and age cohort*



Source: Young Lives, Rounds 2 and 3.

5. Final considerations

In this technical note we have presented the psychometric analysis performed on the TVIP scores gathered in Rounds 2 and 3 of the Young Lives survey for the Older and Younger Cohorts. The psychometric analysis had as its main objective to estimate comparable TVIP scores by round and cohort. For this purpose, we used Rasch models since this statistical technique allows us not only to estimate standardised scores but also to equate test scores.

In terms of item fit, the reliability indexes at child and at item level, before and after equating, are pretty good. Child reliability indexes range from 0.98 to 0.99 while item reliability indexes range from 0.99 to 1.00, showing a good internal consistency for the item difficulty and child ability measures estimated among all models.

Regarding DIF, we were able to identify a set of items across different item difficulty levels that did not have DIF by gender and round in the case of the Younger and Older Cohort analysis, and without gender, round and cohort bias for the pooled analysis, and we used them as anchor items for the equating analysis. These items could be used in subsequent rounds as anchors to equate TVIP scores.

In terms of test equating, we did not drop the items flagged with bias; instead we used them as new items, splitting them by DIF. This strategy gives us ability measures free of DIF as well as increasing the accuracy of the item estimates. Also, avoiding dropping items flagged with DIF reduces the chances of a ceiling effect, given the reduction in test length.

The equated Rasch scores show the same pattern as the raw scores measures. The change over time of TVIP scores for the Younger Cohort is higher (1.5 SD) than for Older Cohort children (0.6 SD). This finding is similar to those reported by other longitudinal studies, carried out both internationally and locally (LoGerfo et al. 2005; Cheadle 2008; and León and Cueto 2013), where higher learning rates are evident during the first years of schooling than in the later grades.

Also, equated Rasch scores are highly correlated with the measures estimated in previous rounds; correlation measures were above 0.90 and all of them were statistically significant. This finding shows that these new scores, as well as the previous ones estimated, are valid to measure child ability, the main difference being that the equated scores allow researchers to compare children's ability over time.

Finally, the Rasch model and the three-parameter model give us the same results in terms of item difficulty and child ability. Both scales have a high degree of correlation (above 0.90), indicating that both of them contain the same information.

References

Bryant, F.B. and P.R. Yarnold (1995) 'Principal Components Analysis and Exploratory and Confirmatory Factor Analysis' in L.G. Grimm and P.R. Yarnold (eds) *Reading and Understanding Multivariale Statistics*, Washington, DC: American Psychological Association.

Cheadle, Jacob E. (2008) 'Educational Investment, Family Context, and Children's Math and Reading Growth from Kindergarten Through Third Grade', *Sociology of Education* 81: 1–31.

Chien, N.C., C. Howes, M. Burchinal, R.C. Pianta, S. Ritchie, D.M. Bryant and O.A. Barbarin (2010) 'Children's Classroom Engagement and School Readiness Gains in Pre-kindergarten', *Child Development* 81.5: 1534–49.

Dunn, L.M., D.E. Lugo, E.R. Padilla and L.M. Dunn (1986) *Test de Vocabulario en Imagenes Peabody*, Circle Pines, MN: American Guidance Service.

Cueto, S. and J. León (2013) *Psychometric Characteristics of Cognitive Development and Achievement Instruments in Round 3 of Young Lives*, Technical Note 25, Oxford: Young Lives.

Cueto, S., J. León, G. Guerrero and I. Munoz (2009) *Psychometric Characteristics of Cognitive Development and Achievement Instruments in Round 2 of Young Lives*, Technical Note 15, Oxford: Young Lives.

Gorsuch, R.L. (1983) *Factor Analysis* (2nd edn), Hillsdale, NJ: Erlbaum.

Hambleton, R.K. (1989) 'Principles and Selected Applications of Item Response Theory' in R.L. Linn (ed.) *Educational Measurement* (3rd edn), New York: Macmillan.

León, Juan and S. Cueto (2013) 'Igualdad y Equidad en los Aprendizajes: La Magnitud de los Factores Socio-demográficos en el Rendimiento de los Estudiantes Peruanos', unpublished article for the Inter-American Development Bank, Education Unit.

Linacre, J.M. (2008) *Winsteps: A Rasch Analysis Computer Program* [Version 3.68], Chicago, IL (http://www.winsteps.com).
Logerfo, Laura, Austin Nichols and Sean Reardon (2006) *Achievement Gains in Elementary and High School*, Washington, DC: The Urban Institute.

Long, Yanjie (2012) 'The Impact of Parental Involvement on Preschool Children's Later Language Development in Low-income Hispanic English Language Learners', Open Access Theses and Dissertations from the College of Education and Human Sciences, University of Nebraska, Lincoln, Paper 144, http://digitalcommons.unl.edu/cehsdiss/144 (accessed 26 September 2013).

Lopez Boo, Florencia (2013) *Intercontinental Evidence on Socioeconomic Status and Early Childhood Cognitive Skills: Is Latin America Different?*, Inter-American Development Bank, Working Paper Series No. 435, Washington, DC: IADB.

Narajian, Michele, Judith Pollack, Alberto Sorongon and Elvira Germino Hausken (2009) *Early Childhood Longitudinal Study-Kindergarten Class of 1998-99 (ECLS-K), Psychometric Report for the Eighth Grade*, Washington, DC: US Department of Education, National Center for Education Statistics.

Paxson C. and N. Schady (2005) *Cognitive Development among Young Children in Ecuador: The Roles of Wealth, Health and Parenting*, World Bank Policy Research Working Paper 3,605, Washington, DC: World Bank.

Umbel, V.M., B.Z. Pearson, M.C. Fernandez and D.K. Oller (1992) Measuring Bilingual Children's Receptive Vocabularies, *Child Development* 4: 1012–20.

van de Vijver, Fons (1986) 'The Robustness of Rasch Estimates', *Applied Psychological Measurment* 10.1: 45–57.

Vogel, Ch., R. Bradley, H. Raikes, K. Boller and J. Shears (2006) 'Relation Between Father Connectedness and Child Outcomes', *Parenting: Science and Practice* 6: 2–3.

Wright B.D. and G.A. Douglas (1976) *Rasch Item Analysis by Hand*, MESA Research Memorandum Number 21, Chicago, IL: Statistical Laboratory, Department of Education, University of Chicago.

Appendix: Item fit statistics

Table A1. *Item infit by model estimated before equating*

	Younger Cohort	Older Cohort	Pooled
it 1	1.2	0.7	1.2
it 2	1.4	1.6	1.4
it 3	1.2	1.0	1.2
it 4	1.1	1.0	1.1
it 5	1.4	2.7	1.4
it 6	1.4	1.0	1.4
it 7	1.3	1.8	1.3
it 8	1.2	0.7	1.2
it 9	1.1	0.9	1.1
it 10	1.1	1.1	1.1
it 11	0.8	1.0	0.8
it 12	1.2	0.6	1.2
it 13	1.1	0.8	1.1
it 14	1.0	0.8	1.0
it 15	0.9	0.6	0.9
it 16	1.1	1.3	1.1
it 17	1.0	1.4	1.0
it 18	1.0	2.5	1.0
it 19	0.9	0.5	0.9
it 20	1.2	1.0	1.2
it 21	1.2	1.1	1.2
it 22	1.1	0.6	1.1
it 23	1.0	1.1	1.0
it 24	1.0	0.6	0.9
it 25	0.9	0.7	0.9
it 26	0.9	1.3	0.9
it 27	0.8	0.2	0.8
it 28	0.9	1.5	0.9
it 29	1.0	0.9	1.0
it 30	1.1	0.9	1.1
it 31	1.0	0.2	1.0
it 32	1.1	1.7	1.1
it 33	1.0	0.8	1.0
it 34	0.5	0.2	0.5
it 35	1.2	0.9	1.2
it 36	1.2	1.2	1.2
it 37	1.2	0.8	1.2
it 38	1.3	1.1	1.3
it 39	1.0	1.2	1.0
it 40	1.1	1.2	1.1
it 41	1.0	1.0	1.0
it 42	1.1	1.4	1.1
it 43	0.9	0.8	0.9
it 44	1.0	0.9	1.0
it 45	1.1	1.1	1.1
it 46	0.9	1.3	0.9
it 47	0.8	0.9	0.8
it 48	0.9	1.1	0.9
it 49	1.0	1.3	1.0
it 50	1.0	0.9	1.0
it 51	1.4	1.1	1.3
it 52	1.0	1.0	1.0
it 53	0.9	0.9	0.9
it 54	1.2	1.1	1.2
it 55	1.0	1.0	1.0
it 56	1.1	0.9	1.1
it 57	1.3	1.0	1.3
it 58	1.2	1.0	1.2
it 59	0.9	0.8	0.9
it 60	0.8	0.8	0.8
it 61	1.2	0.8	1.0
it 62	1.4	1.2	1.3
it 63	1.0	1.0	1.0
it 64	0.8	0.8	0.8
it 65	0.9	0.7	0.8
it 66	0.9	1.1	0.9
it 67	1.0	1.0	1.0
it 68	1.3	1.2	1.2
it 69	0.9	0.9	1.0
it 70	1.1	1.0	1.0
it 71	0.9	0.9	0.8
it 72	1.1	1.3	1.2
it 73	1.0	1.1	0.9
it 74	0.9	1.0	0.9
it 75	1.0	1.1	1.3
it 76	0.9	1.0	1.0
it 77	1.1	1.4	1.2
it 78	1.0	1.4	1.3
it 79	0.8	1.2	0.9
it 80	0.9	1.6	1.3
it 81	1.2	1.5	1.6
it 82	0.8	0.9	0.7
it 83	1.1	1.1	1.0
it 84	1.0	1.3	1.2
it 85	1.0	1.3	1.2
it 86	0.8	0.9	0.8
it 87	0.9	1.1	1.1
it 88	0.8	0.9	0.8

	Younger Cohort	Older Cohort	Pooled
it 89	1.0	1.1	1.1
it 90	1.1	1.2	1.2
it 91	1.0	1.2	1.3
it 92	1.0	1.1	1.1
it 93	1.1	1.1	1.1
it 94	1.0	0.8	0.8
it 95	0.8	0.7	0.6
it 96	1.1	1.1	1.0
it 97	0.9	1.1	1.1
it 98	0.9	1.2	1.1
it 99	0.8	0.5	0.5
it 100	0.7	0.7	0.6
it 101	1.1	1.1	1.1
it 102	1.1	1.0	1.0
it 103	0.7	0.8	0.8
it 104	0.8	0.8	0.7
it 105	0.9	1.0	1.0
it 106	0.8	1.2	1.1
it 107	0.7	0.9	0.9

	Younger Cohort	Older Cohort	Pooled
it 108	0.7	0.8	0.7
it 109	0.8	0.9	0.9
it 110	0.8	1.1	1.0
it 111	0.5	0.6	0.6
it 112	0.8	0.9	0.9
it 113	0.7	0.8	0.8
it 114	0.7	1.2	1.1
it 115	0.7	0.9	0.9
it 116	1.0	1.0	1.0
it 117	1.0	0.6	0.6
it 118	0.7	1.0	1.0
it 119	0.6	0.8	0.8
it 120	1.3	1.1	1.1
it 121	0.7	0.8	0.8
it 122	0.9	1.2	1.2
it 123	0.7	0.9	0.8
it 124	1.1	1.0	1.0
it 125	0.7	1.2	1.2

Table A2. *Item Infit by model estimated after equating*

Item	Younger Cohort	Older Cohort	Pooled
It 1	1.15	0.69	1.37
It 2	1.38	1.60	1.06
It 3	1.07	1.00	1.31
It 4	1.31	1.00	0.81
It 5	0.81	2.62	0.96
It 6	0.96	1.00	0.90
It 7	0.90	1.78	1.09
It 8	1.10	0.70	0.94
It 9	0.91	0.90	0.91
It 10	1.08	1.13	0.81
It 11	0.91	1.00	0.94
It 12	0.81	0.63	1.07
It 13	0.94	0.80	0.48
It 14	1.09	0.75	1.21
It 15	0.49	0.58	1.00
It 16	1.23	1.30	1.11
It 17	1.22	1.34	0.96
It 18	1.01	2.51	1.23
It 19	1.13	0.54	0.77
It 20	0.97	0.97	1.29
It 21	0.96	1.11	1.04
It 22	0.87	0.58	0.96
It 23	1.26	1.02	0.83
It 24	0.79	0.60	1.06
It 25	1.04	0.69	1.00
It 26	0.91	1.32	1.02
It 27	1.05	0.26	0.92
It 28	0.94	0.93	0.98
It 29	1.10	0.92	0.98
It 30	1.03	0.26	1.10
It 31	1.05	0.79	0.83
It 32	0.96	0.26	1.00
It 33	0.88	0.92	1.15
It 34	1.04	1.24	1.25
It 35	1.01	0.79	1.06
It 36	1.06	1.12	1.20
It 37	0.96	1.17	0.92
It 38	1.11	1.23	1.03
It 39	0.86	1.03	1.19
It 40	0.90	1.41	1.24
It 41	0.81	0.83	0.81
It 42	0.73	1.32	0.79
It 43	1.10	1.07	1.12
It 44	1.12	1.32	1.15
It 45	0.70	0.94	0.95
It 46	0.78	1.06	0.77
It 47	0.89	1.03	0.71
It 48	0.79	0.89	1.22

Item	Younger Cohort	Older Cohort	Pooled
It 49	0.67	1.09	0.97
It 50	0.74	0.99	1.11
It 51	0.80	1.00	0.88
It 52	0.77	1.00	1.00
It 53	0.55	0.85	1.20
It 54	0.82	0.75	1.32
It 55	0.70	0.84	1.02
It 56	0.66	1.18	0.84
It 57	0.72	0.78	1.07
It 58	0.97	0.66	1.06
It 59	1.01	1.17	1.06
It 60	0.72	0.92	0.94
It 61	0.58	1.05	0.88
It 62	1.29	0.94	1.20
It 63	0.68	1.07	1.35
It 64	0.88	1.13	1.11
It 65	0.65	1.04	1.13
It 66	1.05	1.36	1.23
It 67	0.68	0.89	1.12
It 68	1.26	1.05	1.01
It 69	1.05	1.34	1.05
It 70	1.21	0.86	1.22
It 71	1.25	1.12	0.96
It 72	0.93	1.07	1.04
It 73	1.04	0.70	1.15
It 74	1.26	1.11	0.95
It 75	0.88	1.12	1.12
It 76	0.81	0.53	0.85
It 77	1.06	0.68	1.21
It 78	0.91	0.83	1.18
It 79	1.24	0.82	1.40
It 80	1.22	0.78	0.93
It 81	0.98	1.06	1.02
It 82	1.13	0.64	1.14
It 83	1.06	0.95	1.04
It 84	0.89	1.00	0.92
It 85	1.02	0.63	0.83
It 86	1.35	1.00	1.18
It 87	0.88	1.11	1.05
It 88	1.01	0.78	1.29
It 89	0.98	0.86	1.11
It 90	1.00	1.02	0.96
It 91	1.13	1.17	1.08
It 92	1.21	1.52	0.85
It 93	1.36	1.54	1.06
It 94	1.11	1.11	1.15
It 95	1.14	1.19	0.91
It 96	1.24	1.04	1.07

Item	Younger Cohort	Older Cohort	Pooled
It 97	1.12	0.88	0.95
It 98	1.02	0.91	0.87
It 99	0.99	1.18	0.83
It 100	1.06	0.93	1.05
It 101	1.14	1.26	0.91
It 102	1.27	1.69	0.95
It 103	1.07	1.54	0.93
It 104	0.96	0.86	1.09
It 105	1.27	1.09	1.04
It 106	0.98	0.72	1.02
It 107	1.04	1.22	0.77
It 108	1.11	1.13	1.07
It 109	1.01	0.97	0.80
It 110	1.09	0.81	0.72
It 111	1.19	0.87	0.69
It 112	0.83	0.74	0.76
It 113	0.88	1.59	0.73
It 114	0.82	1.71	0.55
It 115	0.88	0.83	0.98
It 116	1.00	1.17	0.67
It 117	0.92	0.89	0.63
It 118	0.81	0.91	0.43
It 119	0.77	1.02	0.63
It 120	0.94	1.04	0.97
It 121	0.77	0.94	0.85
It 122	1.00	1.45	0.76
It 123	0.88	1.42	1.01
It 124	0.95	1.47	0.93
It 125	0.99	0.84	1.04
It 126	1.18	1.17	0.96
It 127	1.41	0.88	1.34
It 128	0.94	1.14	1.04
It 129	1.02	1.06	1.23
It 130	1.14	1.06	0.70
It 131	1.04	0.96	1.11
It 132	0.92	0.78	0.53
It 133	0.89	0.92	0.68
It 134	0.84	1.40	0.83
It 135	0.95	1.24	0.82
It 136	1.19	1.23	0.77
It 137	1.11	1.23	0.64
It 138	1.06	1.31	0.63
It 139	1.32	1.14	0.78
It 140	1.03	1.08	1.43
It 141	1.33	1.28	1.01
It 142	1.15	0.87	0.92
It 143	0.88	1.23	0.86
It 144	1.16	1.10	1.05
It 145	1.36	1.14	1.26
It 146	0.86	1.06	1.40

Item	Younger Cohort	Older Cohort	Pooled
It 147	0.89	1.26	0.76
It 148	0.96	1.12	1.02
It 149	1.30	1.15	0.93
It 150	1.09	1.13	0.82
It 151	0.99	0.91	1.12
It 152	1.02	1.17	1.42
It 153	0.98	0.91	1.15
It 154	0.85	1.10	1.17
It 155	0.92	1.26	0.92
It 156	1.02	0.89	1.07
It 157	0.78	1.01	1.12
It 158	0.75	0.99	1.31
It 159	0.77	1.03	0.91
It 160	1.43	0.92	1.05
It 161	1.02	0.98	0.97
It 162	0.93	0.69	0.90
It 163	0.94	0.86	0.99
It 164	0.64		1.24
It 165	0.86		0.87
It 166	0.86		0.69
It 167	0.87		1.00
It 168	0.84		1.06
It 169	1.02		0.91
It 170	0.96		0.75
It 171	1.00		0.97
It 172	1.39		0.86
It 173	0.76		1.07
It 174	1.14		1.12
It 175	1.03		0.93
It 176	0.72		0.92
It 177	0.93		0.66
It 178	1.08		1.05
It 179	0.97		0.75
It 180	1.04		0.95
It 181	0.76		1.11
It 182	1.10		1.17
It 183	1.27		0.92
It 184	1.43		0.77
It 185	1.16		1.35
It 186	1.13		1.08
It 187	1.17		1.18
It 188	0.75		1.04
It 189	0.89		0.72
It 190	0.85		0.94
It 191	0.89		0.87
It 192	1.05		1.06
It 193	0.84		0.87
It 194	0.74		0.88
It 195	0.54		0.98
It 196	1.32		0.97

Item	Younger Cohort	Older Cohort	Pooled
It 197	0.92		1.00
It 198	1.31		0.76
It 199	1.04		0.93
It 200	0.91		0.89
It 201	0.96		0.95
It 202	1.05		0.88
It 203	1.05		0.94
It 204	0.89		0.61
It 205	0.75		0.78
It 206	0.97		1.00
It 207	1.00		0.85
It 208			0.83
It 209			1.32
It 210			0.68
It 211			0.93
It 212			1.00
It 213			1.17
It 214			1.16
It 215			1.03
It 216			1.07
It 217			0.87
It 218			0.77
It 219			0.78
It 220			1.13
It 221			1.06
It 222			1.25
It 223			1.00
It 224			1.06
It 225			0.87
It 226			0.86
It 227			0.90
It 228			0.80
It 229			1.00
It 230			1.00
It 231			0.94
It 232			1.11
It 233			1.32
It 234			0.88
It 235			1.15
It 236			0.89
It 237			1.30
It 238			1.07
It 239			1.01
It 240			0.77
It 241			0.86
It 242			0.27
It 243			1.18
It 244			0.95
It 245			0.87
It 246			0.85

Item	Younger Cohort	Older Cohort	Pooled
It 247			0.64
It 248			1.10
It 249			1.39
It 250			1.15
It 251			1.27
It 252			1.09
It 253			0.17
It 254			1.31
It 255			1.12
It 256			0.62
It 257			0.81
It 258			0.67
It 259			1.17
It 260			1.12
It 261			1.08
It 262			1.17
It 263			1.26
It 264			0.92
It 265			0.86
It 266			0.94
It 267			0.83
It 268			0.75
It 269			0.81
It 270			0.97
It 271			0.87
It 272			1.13
It 273			0.96
It 274			1.00
It 275			1.04
It 276			1.02
It 277			1.01
It 278			1.24
It 279			0.97
It 280			1.48
It 281			0.76
It 282			0.91
It 283			1.01
It 284			1.22
It 285			1.57
It 286			1.57
It 287			1.14
It 288			0.93
It 289			1.08
It 290			0.79
It 291			0.90
It 292			1.26
It 293			1.84
It 294			1.46
It 295			1.39
It 296			1.13

Item	Younger Cohort	Older Cohort	Pooled
It 297			0.89
It 298			0.98
It 299			1.02
It 300			0.98
It 301			0.71
It 302			0.76
It 303			1.06
It 304			1.30
It 305			1.05
It 306			1.05
It 307			0.89
It 308			1.06
It 309			0.82
It 310			1.17
It 311			1.07
It 312			1.46

Item	Younger Cohort	Older Cohort	Pooled
It 313			1.01
It 314			0.92
It 315			0.81
It 316			1.54
It 317			1.64
It 318			1.42
It 319			1.40
It 320			1.00
It 321			0.74
It 322			0.78
It 323			0.80
It 324			1.31
It 325			1.14
It 326			1.48
It 327			1.42

Table A3. *Item bias for Younger Cohort*

	Bias by	
	Gender	**Round of survey administration**
It 1		
It 2		
It 3		T
It 4		
It 5	G	T
It 6		T
It 7		
It 8	G	
It 9		T
It 10		T
It 11		
It 12		T
It 13		T
It 14		
It 15		
It 16		
It 17		T
It 18	G	
It 19		
It 20	G	
It 21	G	
It 22		
It 23	G	T
It 24		T
It 25	G	
It 26		
It 27		
It 28		
It 29		T
It 30		
It 31		T
It 32	G	T
It 33	G	
It 34		
It 35		T
It 36		
It 37		
It 38	G	
It 39		
It 40		
It 41		
It 42		T
It 43	G	
It 44		T
It 45	G	T

	Bias by	
	Gender	**Round of survey administration**
It 46	G	
It 47	G	T
It 48	G	T
It 49	G	T
It 50	G	T
It 51		T
It 52		
It 53		
It 54		
It 55	G	
It 56		T
It 57		T
It 58		T
It 59		T
It 60		
It 61		T
It 62		T
It 63		
It 64		T
It 65		T
It 66		T
It 67	G	
It 68		T
It 69	G	T
It 70		T
It 71		
It 72		
It 73		T
It 74		
It 75		T
It 76		T
It 77		
It 78		
It 79		T
It 80		T
It 81	G	
It 82	G	T
It 83		
It 84		
It 85		T
It 86		T
It 87		
It 88		T
It 89	G	T
It 90	G	T

	Bias by	
	Gender	**Round of survey administration**
It 91		
It 92		
It 93		
It 94		
It 95		T
It 96		
It 97		
It 98		
It 99		
It 100		
It 101		
It 102		
It 103		
It 104		
It 105		
It 106		
It 107		
It 108		

	Bias by	
	Gender	**Round of survey administration**
It 109		
It 110		
It 111		
It 112		
It 113		
It 114		
It 115		
It 116		
It 117		
It 118		
It 119		
It 120		
It 121		
It 122		
It 123		
It 124		
It 125		

Note: G = Gender bias, T = Round bias.
For all the item bias analysis, the Rasch-Welch method was used (Linacre 2008).

Table A4. *Item bias for Older Cohort*

	Gender	Round
It 1		
It 2		
It 3		
It 4		
It 5		
It 6		
It 7		
It 8		
It 9		
It 10		
It 11		
It 12		
It 13		
It 14		
It 15		
It 16		
It 17		
It 18		
It 19		
It 20		
It 21		
It 22		
It 23		
It 24		
It 25		
It 26		
It 27		
It 28	G	
It 29		
It 30		
It 31		
It 32	G	
It 33		
It 34		
It 35		
It 36		
It 37		
It 38		
It 39		
It 40		
It 41		
It 42		
It 43		
It 44	G	
It 45	G	
It 46		
It 47	G	
It 48		

	Gender	Round
It 49		
It 50		
It 51		
It 52		
It 53		
It 54		
It 55		
It 56	G	
It 57		
It 58		
It 59		
It 60		
It 61		
It 62		
It 63	G	
It 64		
It 65		
It 66	G	
It 67	G	
It 68		
It 69		
It 70		
It 71		
It 72		T
It 73		
It 74	G	T
It 75		
It 76		
It 77		
It 78	G	
It 79		T
It 80	G	
It 81	G	
It 82		
It 83		
It 84		T
It 85		
It 86	G	
It 87	G	
It 88		
It 89		
It 90		T
It 91		T
It 92		
It 93		T
It 94	G	
It 95		
It 96		

	Gender	Round
It 97		
It 98	G	
It 99		
It 100		
It 101	G	
It 102		T
It 103		
It 104		
It 105	G	
It 106		T
It 107		T
It 108		
It 109	G	
It 110		
It 111		

	Gender	Round
It 112	G	T
It 113	G	
It 114		T
It 115		
It 116		
It 117		
It 118		
It 119	G	
It 120		
It 121		
It 122		T
It 123		
It 124		
It 125		

Note: G = Gender bias, T = Round bias
For all the item bias analysis, the Rasch-Welch method was used (Linacre 2008).

Table A5. *Item bias for pooled sample*

	Gender	Round	Cohort
It 1			C
It 2			
It 3		T	
It 4			
It 5	G	T	
It 6		T	
It 7			
It 8	G		
It 9		T	
It 10		T	
It 11			
It 12		T	
It 13		T	
It 14			
It 15			
It 16			
It 17		T	
It 18	G		
It 19			C
It 20	G		
It 21	G		C
It 22			C
It 23	G	T	
It 24			
It 25	G		
It 26			
It 27			
It 28			
It 29		T	
It 30			
It 31		T	C
It 32	G	T	C
It 33	G		
It 34			
It 35		T	
It 36			
It 37	G		
It 38	G		
It 39			
It 40			
It 41			
It 42		T	C
It 43	G		C
It 44		T	
It 45	G	T	
It 46	G		
It 47	G		C
It 48	G	T	

	Gender	Round	Cohort
It 49	G		C
It 50	G	T	C
It 51		T	C
It 52			C
It 53			C
It 54			
It 55	G		C
It 56	G	T	C
It 57		T	
It 58		T	
It 59		T	C
It 60			
It 61		T	C
It 62			
It 63			
It 64			C
It 65		T	C
It 66		T	
It 67	G		C
It 68		T	C
It 69	G	T	C
It 70			C
It 71			C
It 72		T	C
It 73			C
It 74			C
It 75		T	C
It 76			
It 77			C
It 78	G	T	C
It 79	G	T	
It 80	G	T	C
It 81	G	T	C
It 82	G		C
It 83			C
It 84	G		C
It 85	G	T	C
It 86	G		
It 87	G		C
It 88			
It 89	G		C
It 90	G	T	
It 91			C
It 92	G		C
It 93		T	
It 94	G		C
It 95			C
It 96			C

	Gender	Round	Cohort
It 97			
It 98	G		
It 99			C
It 100			C
It 101	G		
It 102			
It 103			C
It 104			C
It 105	G		
It 106		T	C
It 107		T	
It 108			C
It 109	G		
It 110			
It 111			C

	Gender	Round	Cohort
It 112	G		C
It 113	G		C
It 114		T	
It 115			
It 116			
It 117			C
It 118			
It 119	G		
It 120			
It 121			C
It 122		T	C
It 123			
It 124			
It 125			

Note: G = Gender bias, T = Round bias, C = Cohort bias
For all the item bias analysis, the Rasch-Welch method was used (Linacre 2008).

Equating Children's PPVT Scores Across Survey Rounds and Age Cohorts in Peru

Young Lives gathers information from children and families through household and child questionnaires as well as children's cognitive and achievement tests. The Peabody Picture Vocabulary Test (PPVT) has been used in all survey rounds to date and with both age cohorts. This technical note presents the psychometric analysis performed (using Item Response Theory, IRT) in order to build cognitive measures comparable across Rounds 2 and 3 and the Younger and Older Cohorts for Peru. To achieve this, a one-parameter IRT model was used. We were able to identify a set of items across difficulty levels with good item fit as well as without item bias by gender and round. This set of items was used to equate the PPVT scores across rounds and age cohorts. Finally, we did an external validity analysis correlating the one-parameter PPVT scores with individual and family characteristics and the results showed that correlations were statistically significant with the expected signs.

Young Lives
An International Study of Childhood Poverty

About Young Lives

Young Lives is an international study of childhood poverty, involving 12,000 children in 4 countries over 15 years. It is led by a team in the Department of International Development at the University of Oxford in association with research and policy partners in the 4 study countries: Ethiopia, India, Peru and Vietnam.

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

Young Lives Partners

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

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

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

Young Lives
An International Study of Childhood Poverty