Young Lives

INSIGHTS REPORTS

October 2020

Measuring the Development of Cognitive Skills Across Time and Context: Reflections from Young Lives

Andrew Dawes

Introduction

Young Lives has traced the development of a range of cognitive skills and psychosocial traits from early childhood through to young adulthood across cohorts in four countries – Ethiopia, India (the states of Telangana and Andhra Pradesh), Peru and Vietnam. These include components of cognitive development (language and mathematical skills), as well psychosocial attributes such as educational and occupational aspirations, self-esteem, self-efficacy, emotional well-being and risky behaviour (Boyden et al. 2019; Boyden, Dawes and Tredoux 2018; Dercon and Krishnan 2009; Favara and Sanchez 2017).

Young Lives has collected data on children's cognitive skills in both its household and school surveys. This paper looks at the longitudinal measurement of cognitive skills, drawing primarily on the household survey experience and a longitudinal analysis conducted using those data (Boyden, Dawes, and Tredoux 2018; Tredoux and Dawes 2018). Yorke and Ogando Portela (2018) cover the properties of Young Lives measures of children's psychosocial functioning. School survey measures are discussed in Iyer and Moore (2017).

Two central points for consideration in longitudinal studies are their temporal design, and the adaptation of tests of cognitive skills for use in populations for which they were not originally designed.

Reflections on

- the timing, frequency, and spacing of observations in a longitudinal study of children's development
- the choice of measures of cognitive functions and skills in a cross-cultural longitudinal study
- points to consider when measuring cognitive skills at successive ages
- adapting, translating, and ensuring the equivalence and fairness of measures for use across cultural communities.

Temporal design

Temporal design refers to the timing, frequency, and spacing of observations in a longitudinal study. Timing of measurement has important consequences for understanding developmental trajectories. Baltes and Nesselroade (1979) advanced reasons for conducting longitudinal research in the field of child development which remain valid today (Grimm, Davoudzadeh, and Ram 2017).

First, longitudinal studies permit examination of *intraindividual* stability and change in a skill such as receptive vocabulary over time (Tredoux and Dawes 2018). Second, they enable us to observe *interindividual* differences and similarities in *intraindividual* change in the trait over time. Here we observe variation between individuals in the amount of development in a trait between say 5 and 8 years old; some children of the same age will develop faster, while others will show less growth. Figure 1 illustrates both of these, using findings from Young Lives in Ethiopia. Each line represents a child's receptive vocabulary growth trajectory. Interindividual differences are also evident.

Third, and provided the measures are equivalent in all groups, the method allows us to observe *intergroup* differences in change and stability; that is, the effects of differences in the child's social group characteristics (e.g. wealth, caste or ethnicity) on their development. For example, Georgiadis et al. (2017) have shown that children from better off Young Lives households are more likely to recover from growth stunting post-infancy.

Fourth, changes in one area of development are likely to be associated with changes in other domains at the same age and across time. For example, reading skills at age 8 are likely to be associated with mathematical ability. Growth in language abilities from 8 to 12 years of age may be associated with improvement in mathematics skills, as children have to read well enough to understand problems and their solutions (Watts et al. 2015). Self-esteem and confidence may grow over time as the child's problem-solving skills improve.

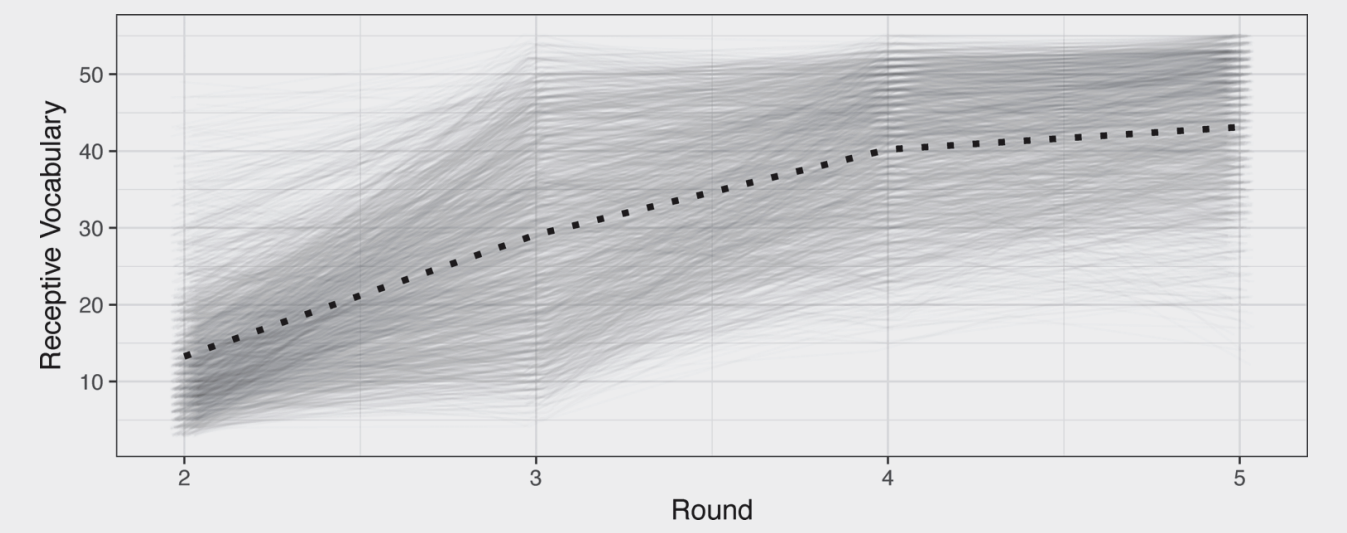
Finally, and if the design permits this, we are able to observe influences that are associated with both intraindividual and interindividual change (as in Figure 1), as well as the effect of variations in societal conditions during the same developmental period. The Young Lives design makes this possible in each country. For example, Dornan and Georgiadis (2015) report that children whose households benefited from social protection measures and improved sanitation and health services, were more likely to show recovery in growth stunting post-infancy.

Study goals obviously play a key role in deciding what to measure and when. As Collins (2006: 508) notes: 'the most appropriate temporal design is one chosen not primarily on the basis of logistics, but instead on the basis of correspondence with the theoretical model of change. Thus, for example, if the theoretical model suggests that change is rapid, or characterized by many ups and downs, then more frequent observation may be called for.'

In studies of children's development, the choice of temporal design should therefore be justified scientifically and not simply be a decision to measure every five years (or some other period). That might be appropriate for tracking household economic well-being linked to government five-year plans, but such an interval may not be helpful to an understanding of language development.

For example, if we wished to understand the role of the child care environment during the first four years of life on the development of vocabulary and other outcomes associated with readiness to learn in the kindergarten year, we would need to measure the relevant indicators at several points from birth to 5 years old, as was done in the NICHD Study of Early Child Care (NICHD Early Child Care Research Network 2002) and the Cebu Longitudinal Health and Nutrition Survey in the Philippines (Adair et al. 2011). In Cebu, mothers were interviewed every two months for the first two years. Five subsequent rounds of data were

Figure 1. *Intraindividual and interindividual differences in receptive vocabulary growth from 5 to 15 years of age*



Source: Tredoux and Dawes (2018).

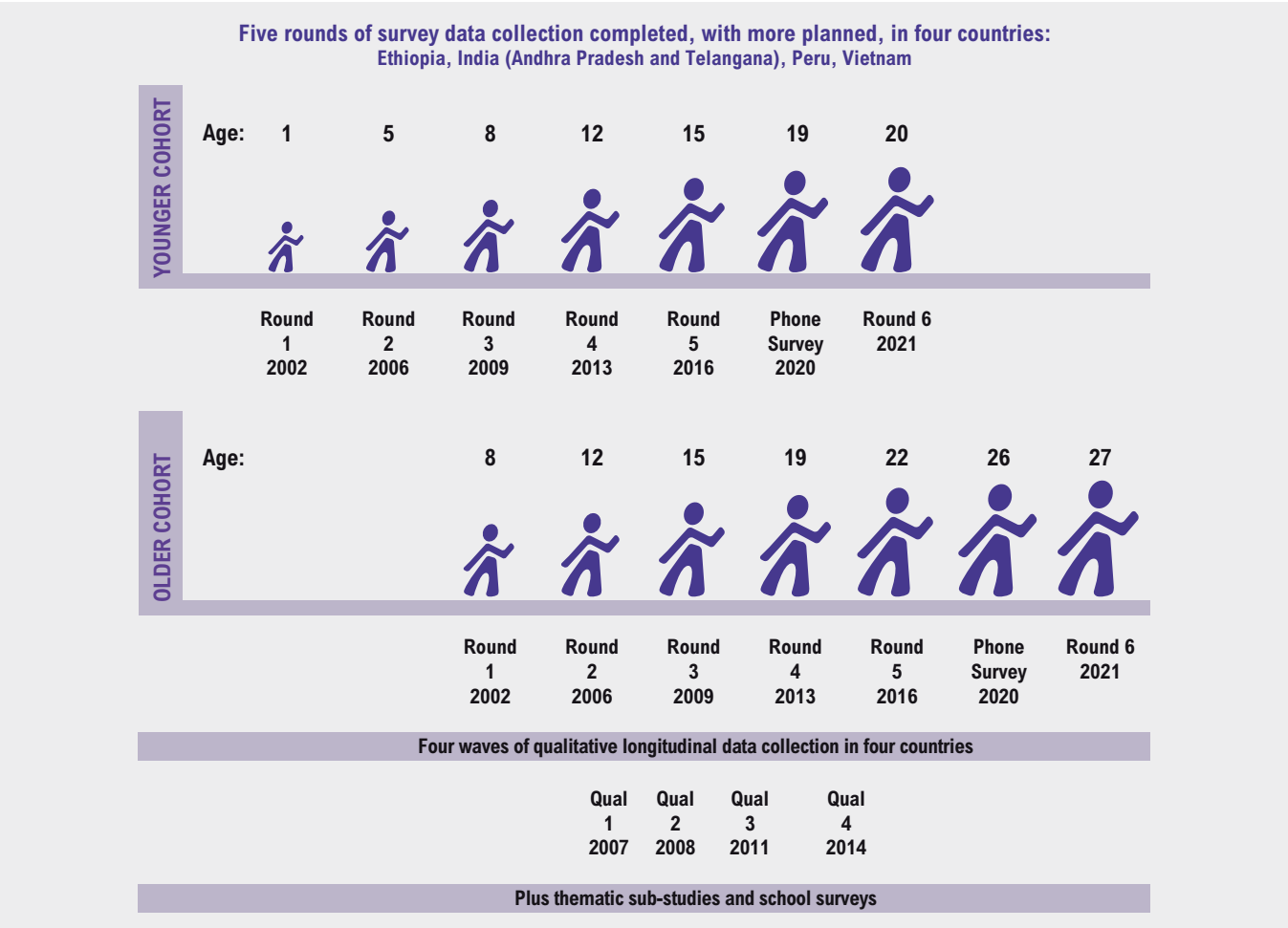
collected at four-year intervals on both mothers and their offspring (including health and nutrition status, school progress, and maths and language ability). Cebu is a multi-disciplinary collaboration informed by a health determinants model that includes individual, household and community level influences on child health and development outcomes. The study required frequent collection of detailed early data on mother and child and their circumstances. Later data points were linked to significant life-course transitions such as beginning high school.

When planning longitudinal study waves, it is important to recognise that life-course transitions are socially constructed. As Elder (1998: 3), doyen of the life-course approach to the study of human development, said: 'the life course of individuals is embedded in and shaped by the historical times and places they experience over their life-time'. While there are many commonalities in the transitions experienced by the young in our increasingly globalised world, we should not lose sight of cultural particularities in life-course transitions that are likely to influence the course and pace of development. These should be taken into account in temporal designs and in the interpretation of study findings.

Young Lives was not designed to provide a fine-grained examination of any particular feature of child development, as was the case with Cebu. Rather, it is a multi-disciplinary multi-purpose investigation in which a range of interests had to be satisfied. The study was originally conceived as a tool for tracking progress towards the Millennium Development Goals. A central objective was to understand the impact of poverty on children and its impact on their later life chances so that appropriate policy recommendations could be made. Understanding the development of cognitive and psychosocial skills under varied socio-economic and policy conditions, as well as their influence on school completion and transitions to the labour market, has been central to the study since inception.

The Young Lives design is unique in being both cross-national and cohort cross-sequential. The latter feature seeks to control for the effects of social change by enrolling more than one cohort in the same year, but at different ages (Figure 2). A unique feature of the design is that longitudinal qualitative data ('Qual' in Figure 1) has been gathered on 100 children in each cohort to explore changes in their everyday experiences and perspectives in greater depth. In addition to the collection of regular data from the Older and Younger Cohorts, the study design also incorporates complementary school effectiveness surveys to evaluate the quality and effectiveness of schooling within the Young Lives sites (Boyden and James 2014).

Figure 2. *The Young Lives cohort cross-sequential design*



Young Lives enrolled two cohorts in 2002, when the Younger Cohort were around 1 year old, and the Older Cohort around 8 years old. Provided the data are suitable, a two-cohort cross-sequential design such as this permits observation of historical change on child development when social policy interventions benefit one and not the other cohort. For example, the introduction of a primary school feeding scheme when the Younger Cohort were enrolled might benefit their growth and cognitive skills, but not those of the Older Cohort who had already moved on to secondary school (Boyden et al. 2019).

Young Lives has been able to take advantage of each of the opportunities for studying developmental change outlined by Baltes and Nesselroade (1979). This is evident in many papers issuing from the study. One example is latent growth modelling of predictors of intraindividual, interindividual, and intergroup changes in mathematics abilities and receptive vocabulary from 5 to 15 years of age in the Younger Cohort (Tredoux and Dawes 2018).[1] As we prepared the data, we pondered the suitability of these age points. Did they represent the best times in the life course to observe stability, transition and growth? How well aligned were they to known biological and social life-course changes likely to affect these abilities? Did they take sufficient account of socially determined points of transition that would be likely to mark transitions? And did they have the necessary predictor variables (informed by theory and evidence) in the datasets to enable a meaningful analysis of relevance to policy? These are all key questions for longitudinal study designs that seek to produce policy-relevant findings.

Young Lives rounds were carried out at more or less at four-year intervals for each cohort, permitting an examination of child development (intra-individual change) against the background of possible societal and household influences and changes. For the most part, given the broad objectives of the multidisciplinary study, the data points seemed appropriate. Measurement of Younger Cohort cognitive skills at 5 years old (Round 2 in 2005) prior to enrolling in Grade 1 at school has enabled Young Lives to investigate a range of influences on the development of language and early numeracy (see Table 1) during early childhood and effects on school performance (e.g. Crookston et al. 2013). These include household wealth, ethnicity, access to services (water, electricity and sanitation), maternal psychological well-being, preschool attendance, and growth stunting at 12 months old.

Arguably, however, our findings and policy recommendations for the early years are likely to have been enhanced had finer grained assessments of cognitive functions between 12 months old and 5 years been undertaken alongside measurement of the nature and quality of the children's early care and learning environments. Both have been shown, respectively, to predict psychosocial well-being and readiness for school (NICHD Early Child Care Research Network 2002; Melhuish et al. 2008). Similarly, gaps in the Young Lives data on pregnancy, birth, early feeding practices and other pertinent variables, meant that the study was unable to determine the reasons why some children recovered from early growth stunting and others did not (Boyden and Walnicki, forthcoming).

In addition, while we measured the effects of preschool enrolment on cognitive outcomes, Young Lives was not able to measure preschool quality or dose (sessions attended). Both are key contributors to the development of cognitive skills among those who are enrolled (Melhuish et al. 2008). While analyses of Young Lives data have shown that preschool participation contributes to cognitive skill development (e.g. Cueto et al. 2016), had preschool quality data been available, they would have enriched our analyses and recommendations. Young Lives has subsequently collected data on primary and secondary schooling quality, with the aim of supporting these types of analyses in later phases of children's development.

At Round 3 of the study in 2008, the Younger Cohort were around 8 years old and the Older Cohort around 12. This enabled the former to be assessed at the same age as their Older Cohort counterparts for cross-sequential cohort comparison. However, there is no clear cognitive developmental or scholastic progress rationale for choosing 8 years old, when most children would be in the middle of the primary education phase (unless they enrolled late). While the timing of later rounds can be debated, if one is interested in factors that affect growth in cognitive skills, ideally one would want rounds to correspond to key schooling phases and established phases of neurocognitive development (and the onset of puberty). These would be the end of primary school and middle childhood (11-12 years) which is also when concrete operational thinking (in Piaget's terms) is likely to be well established in all children, and middle (14-15 years) and late adolescence (18-19 years), points of significance for both phases of schooling and neurocognitive development. These points are appropriately covered in the Young Lives survey. Care must be taken to align rounds to the particular education phases in the study country, which may be a challenge in a multi-country study. For example, in Ethiopia children are expected to enrol at 7 years of age while in the other three countries this is 6 years old. Conducting the surveys when the children in both cohorts were 8 years old could be seen as a practical compromise.

In sum, longitudinal study rounds should as far as possible be informed by the study goals, knowledge of the area of development that is the focus of the research, and the factors likely to influence intraindividual change and stability over time. This can be a particular challenge in a study in which members have questions arising from different interests and disciplines. The cost of data collection is a major consideration, as is the measurement burden on participants, which is also an important ethical consideration. This can become quite acute as the child ages and investigators from different disciplines argue about the addition of new modules to answer questions

1 Latent growth modelling (LGM) is a statistical technique that permits description of both individual developmental trajectories over time, and differences between individuals as a function of variations in their backgrounds and other influences (predictors of the trajectories measured at each study round).

that become relevant (for example, in adolescence). The orientations of study team members also play a role in these decisions. For example, those with an interest in human capital development rather than in the process of human development, are likely to have different ideas as to what should be measured and when.

Regardless of the temporal design, questions around the suitability, validity and reliability of the tests of language and mathematics abilities used in Young Lives in four countries with children from different socio-economic backgrounds and many ethnic and language groups had to be considered.

Adaptation of measures in cross-cultural studies

This section draws on Young Lives child survey experience and other studies to highlight the main issues that need to be addressed when tracking children's psychological development across time and cultures. Iyer and Moore (2017) provide a discussion of challenges in using cognitive tests to measure learning quality in the diverse social and cultural contexts in which the Young Lives school surveys were conducted.

There are three main questions to consider when measuring cognitive skills and psychosocial functioning at successive ages in a longitudinal study.

1. What sort of test[2] is most appropriate for measuring the construct of interest in the socio-economic, language and cultural groups of the study sample? This is no trivial matter, particularly when the study is to be conducted among samples characterised by rural-urban and socio-economic variation, as well as cultural and linguistic differences.
2. If the study includes children from different language and cultural groups, what adaptations might be necessary to ensure that the test is a valid, reliable and fair and equivalent measure of the trait in each group?
3. How would we determine whether the chosen test has measurement invariance – that is, does it measure children on the same construct at each age point?

Then, when choosing tests, we need to consider three further questions.

- What are the skills and functions (our dependent variables) that we seek to measure over time and at which specific points in development?
- What is our theoretical model of change in these traits?
- How does our model guide our choice of predictor measures (independent variables) at particular times of measurement in the temporal design? What time-invariant (e.g. ethnic identity), and time-varying covariates (e.g. wealth status) may predict changes and stability in the development of the trait measured by the test?

If possible, a review of tests used in other studies conducted in the same or similar study context is the first step. As the goal is to contribute to knowledge, consideration should be given to using these tests if they have been shown to be psychometrically sound (Economic and Social Research Council 2017). Alternatively, other tests, particularly those of local origin (see below) should be considered if they can be shown to reliably measure the same construct (concurrent validity).

The next key consideration is whether the tests are likely to be fit for purpose in the cultural context of the proposed study. Are the measures likely to reliably, validly and fairly assess the traits of interest in the new study population, and if not, what work is necessary to ensure that they do?

Historically, virtually all valid and reliable instruments measuring children's cognitive skills have been developed in the Global North and then used in countries of the Global South. Referring to East Africa, for example, Mwaura and Marfo (2011: 138) comment critically that, 'imported instruments [are] often adopted with little or no adaptations'. When conducting studies in low and middle-income (LMIC) countries of the Global South, a key question is therefore whether to use a test developed and standardised in Europe or USA and which is well regarded as a sound measure, or to select an instrument that has been developed or adapted in the study country and which has been shown to be a valid, reliable and fair assessment of the abilities of the children regardless of their language or cultural background? Perhaps this question has not been sufficiently addressed. But the situation is changing (Serpell and Simatende 2016), and an increasing number of psychometrically sound instruments have been developed or adapted for LMIC children in early and middle childhood (Baddeley, Gardner, and Grantham-McGregor 1995; Dowd et al. 2016; Pisani, Borisova, and Dowd 2015; Matafwali and Serpell 2014; Snelling et al. 2019). A further point is that if the intention is for study findings to inform educational policy, then results from culturally appropriate locally validated tests are more likely to be accepted by policymakers.

Test adaptation

The International Test Commission guidelines make a distinction between adaptation and translation:

'Test adaptation refers to all of the activities including: deciding whether or not a test in a second language and culture could measure the same construct in the first language; selecting translators; choosing a design for evaluating the work of test translators (e.g., forward and backward translations); choosing any necessary accommodations; modifying the test format; conducting the translation; checking the equivalence of the test in the second language and culture and conducting other necessary validity studies. Test translation … [is] restricted to the actual choosing of language to move the test from one language and culture to another to preserve the linguistic meaning.' (Gregoire 2018: 103).

2 'Test' is used to refer to instruments used to measure children's skills and psychosocial functioning.

Adaptation of a test to novel cultural groups requires investigation of the tests' cross-group equivalence. The psychometric properties of the adapted instrument need to be established, including their reliability, validity, measurement invariance, factor structure, item difficulty, and fairness (Carter et al. 2005; Hambleton 2002; Milfont and Fischer 2015; Peña 2007; Peña and Quinn 1997).

Translation

The words and phrases used in the translation (target language) should represent the sense of the source language as far as possible. Further, instructions should be rendered in the *standard* rather than 'professional' use of the language commonly provided by professional translators. Double-translation is increasingly recommended because translator idiosyncrasies can arise when using single forward and back translation (Gregoire 2018; Hambleton 2002; Hambleton and Zenisky 2010). Recommended practice is for at least two native speakers of the study population's language who are familiar with local language usage to translate from the source language and arrive at a common position. This version is then back translated by a third person and the three work on any outstanding discrepancies. Further issues should be addressed when piloting in each language. In the final product, the language used in test instructions must feel natural and accessible to the child rather than be a literally equivalent version of the original. In my experience this often occurs when professional translators unfamiliar with local usage are used.

While this takes time, it is important. Careful attention to the translation process and maintenance of a standard approach in the field is important to prevent the test administrators using their own approach, which will introduce variations that undermine the reliability of the test, the procedures and the integrity of the results.

The Young Lives approach to translation

A number of major and minor languages were spoken by Young Lives children. The Vietnamese sample included speakers of seven languages, with 11 in Ethiopia, eight in India, and six in Peru. These numbers exclude variations in dialect and usage. Such a range requires a careful processes of test adaptation to ensure that the children's performance in each language can be regarded as a reliable and fair indication of their abilities.

Young Lives used double-translation, whereby two mother tongue speakers of the indigenous language who were proficient in English (or Spanish in Peru), undertook independent translations of all the tests into the main national languages of the study sample. They then conferred to finalise the translated version prior to piloting with small samples. Following this, further adaptations might be required. This did not apply to the smaller languages which were translated during test administration. Cueto and León (2012) report that translation was not always standardised for these smaller groups, raising questions regarding the reliability of scores and validity of the tests in Rounds 2 and 3.

When instruments are adapted for use in new populations, as in the case of Young Lives, it is essential to conduct pilots and psychometric analyses to establish test reliability and validity. In Young Lives, these were undertaken prior to Round 2 (Cueto et al. 2009) and at later points when tests were introduced for older groups. Classical test theory methods are used to establish the reliability and validity of the tests. Item response theory methods were used to assess whether items vary in difficulty across test languages, and whether they discriminated unfairly between children from different language or cultural backgrounds who had the same level of ability (Cueto and León 2012; León and Singh 2017). Where appropriate for analysis, scores should be converted to the same scale (z-score transformations). Care should be taken with language groups that are too small for rigorous psychometrics and for which reliability and validity has not been established. Because of this concern, in modelling predictors of growth in receptive vocabulary, Tredoux and Dawes (2018) only included children who had completed the test in a familiar language for which psychometric properties had been established.

Investigation of the properties of tests used in new settings is time consuming but essential. It is likely to be necessary at more than one point in a longitudinal analysis. For a full discussion, see the citations above and others in the Young Lives technical notes,[3] and other key papers that describe the procedures used to address these matters (e.g. McArdle et al. 2009; Milfont and Fischer 2015).

In Young Lives, cognitive skills were mainly measured using existing instruments or adaptations thereof (Table 1). School survey psychometry is discussed by Azubuike, Moore and Iyer (2017).

While the Young Lives mathematics and other tests were standard across languages and countries, this was not the case with the adapted receptive language vocabulary test, which should be regarded as specific to each language. Cueto and León (2012) therefore advise against comparing children's receptive vocabulary across languages and countries. Growth in children's ability over time and predictors thereof should be examined within languages (e.g. Tredoux and Dawes 2018). This caution should be born in mind by other studies using adapted language assessments for each of several language groups.

The following sections discuss some of the guidance on test adaptation and translation that has emerged in recent years, using examples from Young Lives and other experience.

3 Available at: https://www.younglives.org.uk/publications-search/%2A?f%5B0%5D=im_field_document_type%3A196

Table 1. *Cognitive skills assessed in the Young Lives household survey*[4]

Peabody Picture Vocabulary Test [PPVT III and R (Peru)]: Administered at age 5 to the Younger Cohort, and to both cohorts at ages 8, 12, 15, and 19. PPVT 111 was developed and standardised in the US to measure receptive vocabulary. The PPVT-R Castellano version was developed for Latin America and translated into Quechua for indigenous children in Peru. The test was translated into the main languages of the other study countries. Psychometric analyses of Round 3 data resulted in shortening the test and adjusting the item order (to reflect item difficulty) in the Ethiopian, Indian and Vietnamese versions. In these countries, the test was no longer the PPVT but a receptive vocabulary test (León and Singh, 2017).
Early Grade Reading Assessment (EGRA): Administered to the Younger Cohort only at age 8. The EGRA was developed by USAID and measures basic skills for literacy acquisition in the early grades: recognising letters of the alphabet, reading simple words, understanding sentences and paragraphs, and listening with comprehension (Cueto and León 2012).
CLOZE Reading Comprehension Test: Administered to the Older Cohort at age 15. Developed by GRADE for Young Lives to measure reading comprehension (Cueto and León 2012).
Reading comprehension: Administered to both cohorts at ages 15 and 19. Items were drawn from the International Student Assessment (PISA), the UNESCO Literacy Assessment and Monitoring Programme (LAMP), and the Young Lives Peru school survey.
Reading and writing achievement tests in local languages: Administered to both cohorts at ages 8 and 12 (Cueto and León 2012).
Cognitive Development Assessment (CDA): Administered to the Younger Cohort only at age 5. The CDA was developed by the International Association for the Evaluation of Educational Achievement. Only the quantities subscale items were used due to the unreliability of the other scales (Cueto and León 2012).
Mathematics: Administered to both cohorts. Items were drawn from PISA and TIMMS mathematics achievement tests used internationally to track children's mathematical ability (Cueto and León 2012). Test items increase in difficulty at each age point and so are not equivalent across rounds. Transformation of scores to the same scale (z-scores) at each point is necessary for comparing achievement over time (Tredoux and Dawes 2018). Three items were consistent across study rounds and were sometimes used to compare progress across rounds. Mathematics skills were measured at ages 5, 8, 12, 15, and 19.

Assessment of equivalence of tests in cross-cultural studies

Cultural fairness

If children from different cultural and language groups are not acquainted with tests it is probable that the tests are not measuring the underlying construct accurately across groups. It is important to ensure as far as possible that children from different language and cultural backgrounds are likely to be equally familiar with the tasks demanded in the items. This can be both a familiarity and a translation issue. Cognitive tests for children developed in the UK or the US may require LMIC children to undertake tasks that are unfamiliar to them. Even an exercise such as drawing a human figure may present challenges to a poor rural African preschool child with very limited experience of holding a pencil and drawing.

The PPVT test of receptive vocabulary was chosen for Young Lives as the same instrument could be used at each study round to track stability and change in an important indicator of language development. Using the same test is one way of reducing measurement invariance. In the PPVT test used in Young Lives, children are shown pages containing four pictures (three distractors and a target), and asked to point to the image that depicts the word said by the assessor: for example, 'put your finger on the tractor'.

Language tests produce particular challenges in translation (Carter et al. 2005). The PPVT translation process conducted by Young Lives illustrates some that can arise. The test was first used in the second round of the study when the Younger Cohort were 5 years old and the Older Cohort were 12. Translators found that in some languages, the English target word had several equivalents, reflecting variations in local usage. And these target words may have been more or less familiar to children even though the order of word presentation followed that in the source PPVT test. This could result in a more difficult and less familiar word in the local language than in the original being presented to a child earlier in the word order than would be the case in the English.

To ensure greater fairness, in all study rounds, children could respond to the PPVT in languages other than the main languages of adaptation, if these were not sufficiently familiar to them. In these cases, fieldworkers translated the target words. However, this is likely to have introduced variations in administration and reduced reliability in the case of some of the small languages where translation in the field was applied. It may therefore be best to only use results from children speaking the main languages, where one can be more certain of the reliability of the data, for the purposes of analyses (Tredoux and Dawes 2018).

While not strictly a matter of cultural fairness, one could argue that many tests of children's abilities are 'school like' and that the actual testing situation is not dissimilar to testing in schools – particularly when it comes to

4 Details of Young Lives psychometric analyses of cognitive skills can be found in León (2020); León and Singh (2017); Cueto and León (2012); Cueto et al. (2009); Tredoux and Dawes (2018).

'paper and pencil' based tests used in the assessment of mathematical abilities. These methods are a cultural practice of a kind that is more familiar to children who are schooled than those who are not. Nunes and colleagues showed that while Brazilian children working the streets could do currency exchanges in their heads, they failed in similar operations in school tests (Nunes, Schliemann and Carraher 1993; Greenfield 1997). The implication is that while a test may have sound reliability in schooled populations, it may not be a reliable reflection of the true ability of those less exposed to school.

Functional equivalence

Psychological variables such as receptive vocabulary or executive functioning are latent constructs that cannot be measured directly (Milfont and Fischer 2015). The child's test performance is an indicator of the construct and must function in the same way across cultural groups. Functional equivalence is therefore established when the test instructions elicit the same behaviour in children from different backgrounds.

In Young Lives, the coloured version of the Raven's Progressive Matrices (developed in the UK during the 1930s) was administered to the Older Cohort when the children were 8 years old. Raven's is a non-verbal test of reasoning and fluid intelligence in which the child is presented with a series of designs (drawn on paper), each of which has a missing element. The task is to select the missing element from a set of options. While it is supposedly free of cultural bias, children from rural backgrounds (particularly in Ethiopia) had difficulty in understanding the test instructions. Also, the task was very unfamiliar. The functional equivalence of the test within and across the countries was not demonstrated and it was not used in later rounds.

Qualitative data from pilot studies is necessary to compare the performance of children across the various language groups on each test prior to test finalisation. Where there are clear differences between groups (one group performs much worse than others), the reasons must be explored.

The author and colleagues recently developed tests of language and mathematics proficiency for children entering Grade 1 in South Africa. These were double-translated into the 11 official languages – a considerable exercise. In one of the number sense and operations items, children are asked to count in ones, up to 20. The child is awarded one point for each number that is counted in the correct order. Pilot results showed a bimodal distribution with a significant proportion not being able to count beyond 10. On further analysis and discussion with teachers, we found that many of these children spoke African languages and were given the instructions in their language. We assumed that this would ensure fairness so they would count in their home language. We established that many had only been taught to count to ten in their African language, so they stopped at that point as they did not know the names of higher numbers in that tongue. Despite our best efforts to produce a linguistically equivalent and fair test, we failed because we did not take sufficient account of classroom practice. In our revised instructions, like in Young Lives, children were told they could choose the language in which they had learnt to count and could even switch languages if needs be.

The lesson was that when undertaking adaptations and translations prior to piloting, it is good practice to ask local informants such as teachers whether difficulties like this might arise. It is also advisable to conduct a pre-pilot phase to investigate whether children from the same backgrounds and age as those that will be participating in the study understand the provisional test instructions (Biersteker and Dawes 2019). Questions such as: did they follow the instruction easily, is there a need to adjust the instructions, and if so, what change needs to be made, are important to ask.

Linguistic equivalence

According to Peña (2007: 1257), 'The main goal for linguistic equivalence is to make certain that the words and linguistic meaning used in the instruments and instructions are the same for both versions'. Peña goes on to say that while direct translation 'usually satisfies the standard for ensuring linguistic equivalence', this is not necessarily the case. Translation can lead to an item being understood differently by those who have the same home language but are from different socio-economic backgrounds. For example, when developing a South African measure of preschool children's ability to count in classes, one of the tasks was for children to select five marbles from a larger assortment of objects and place them in a box. We translated the English word into Afrikaans as *albaster*, which is technically correct. But when we asked Afrikaans-speaking 5 year olds in our pilot study to put five *albasters* in the box, many had no idea what to do. When asked, they said they called marbles *gatjies*, a slang term used in their working-class community. Once we used that term, they understood the task (Snelling et al. 2019). This and other experiences led us to find out what children called the objects and to use the vernacular if necessary, provided this did not undermine the construct we were measuring. In such cases, all test administrators need to be trained to use the same term and, should the need arise, to ask the child whether she understands the instruction or uses another term for the object in question.

Cultural equivalence

Here the task is to ensure that items are not likely to prejudice the performance of particular groups of children due to lack of familiarity. In Young Lives, the Raven's test failed on both functional and cultural equivalence: many children did not understand the instructions and were not familiar with the tasks demanded in the test. It is very unlikely that children will be familiar with the actual task required for a test such as Raven's, but some exposure to similar tasks in the child's everyday world is likely to enable her to complete a new and unfamiliar task. For example, completing Raven's patterns would be enabled by experience with other forms of pattern completion, such as using blocks to complete or copy designs as commonly used in modern preschools. If a child has never undertaken the type of task expected in the test, this is likely to prevent them from performing as well as their latent trait might otherwise predict.

As noted in Table 1, Young Lives used the Peabody Picture Vocabulary Test (PPVT III) to measure receptive vocabulary. In that test, study participants are shown a card containing four pictures. They are instructed to identify the picture that matches the word spoken by the test administrator, for example, 'put your finger on the duck'. The total number of correct word–object matches is the child's score. When used in a new population, key issues for a test such as this are familiarity with the objects and actions depicted in the pictures, and whether there are single words of equivalent difficulty in a language of translation (see metric equivalence below). In Young Lives translations this was often not the case. For example, in one PPVT item, the child is asked to indicate a picture of 'hurdling' (the target picture shows a boy jumping over a hurdle in a race). The Indian language Telugu has no word for 'hurdling' and the literal translation equates to 'the boys jump over the fence', which would not be acceptable as equivalent to the English.

A South African study that used PPVT-IV to assess receptive vocabulary in rural preschool children faced similar issues. Several words and images had to be adapted as they were found during piloting to be unfamiliar to the children (Biersteker, Dawes, and Hendricks 2012; Biersteker and Dawes 2019). For example, the target word 'cookie' and the associated picture were unfamiliar. Both the word and picture were changed to 'bread'. In another case, to correctly identify the picture for the word 'lamp', the child had to put her finger on the picture of a bedside lamp. This was unfamiliar to many rural children who did not have electricity. They did, however, know what a paraffin lamp was, and a picture of a hurricane lamp was therefore substituted for the bedside lamp.

Metric equivalence

Measurement invariance is the central property of metric equivalence. It is assessed using statistical tests, including factor analysis and item response modelling, to determine whether the instrument (e.g. the PPVT) measures the same construct in the same scale in each of the study groups and at each study round (McArdle et al. 2009; Grimm, Davoudzadeh, and Ram 2017; Millsap 2010; Milfont and Fischer 2015). Measurement invariance is necessary if we are to compare children's scores on the same measure within and across time. Furthermore, if a test has been designed, for example, to measure executive functioning in preschool children, it will not be appropriate for adolescents. Cognisance will have to be taken of the fact that while some components of the test may remain the same (potentially acting as 'anchor items' to allow tests to be put onto the same scale), others more appropriate to adolescent development will be introduced, and the equivalence of the measure across groups will need to be revisited at that point.

Item difficulty may vary from the source language when a test is translated. Peña (2007) notes, for example, that a word may be used with high frequency in the original language but less so in the language of translation. In a test of receptive vocabulary, words used with high frequency and which are more familiar should come earlier in the item set. Simply translating the source word into the other language, without taking account of familiarity and difficulty in the order in which items are presented in the language of translation, will not result in a test of equivalent difficulty to the source. In the Young Lives translation of PPVT into local languages, the order in which items were presented in Rounds 2 and 3 was left the same as the source test. However, analysis of item difficulty in each language of translation using Round 3 data established that the word difficulty in the translations for Ethiopia, India and Vietnam was not equivalent to the English, and that the progressive difficulty of the words was not the same as that in the source (León and Singh, 2017). This meant that the tests in the different languages were not metrically equivalent. To address this, in later rounds the list of target words for Ethiopia, India and Vietnam were reduced in number and re-ordered to correspond to their difficulty so as to produce a more reliable set. For example, in the English source test, the first three words and images are: *bus*, *drinking*, *hand*. In the Vietnamese translation the first three became *key* (word 4 in the source), *fly* (word 12 in the source), and *feather* (word 16 in the source). A further complication, especially in India, was the language in which children chose to respond. They could answer in whichever language they felt most comfortable and sometimes opted to use different languages than those into which the test had been translated. For these reasons, and following psychometric analyses (León and Singh, 2017), significant adaptations to the PPVT were made from Round 4 (Table 1).

Concluding points

This reflection aimed to extract the essentials of a set of complex issues that bear on both temporal design and the quality and integrity of latent variables measured in longitudinal studies. As will have hopefully been evident, considerable effort needs to be put into the selection and adaptation of measures of cognitive skills and psychosocial functioning for application in novel cultural settings. Some of the work of adaptation, including gaining an understanding of local practices and language usage through piloting and stakeholder engagements, needs to be undertaken before the study goes into the field. This will assist in ensuring that when that time comes, the measures are as fit for purpose as possible. These points need to be taken into account in the preparation of funding proposals.

I close with five key guidelines on adapting and translating tests for use in cross-cultural settings drawn from the International Test Commission (International Test Commission 2016; Gregoire 2018), my own experience, and other helpful sources (Baddeley, Gardner, and Grantham-McGregor 1995; Hambleton and Kanjee 1995; Biersteker and Dawes 2019; Carter et al. 2005; Peña 2007; Van Widenfelt et al. 2005).

1. When choosing tests to measure cognitive skills in LMICs, rather than simply importing instruments, consider tests that have been developed in those contexts if they are psychometrically robust. A test developed for an African Kiswahili sample is as likely to require psychometric scrutiny when adapted for use in another language in Africa or elsewhere. It is inappropriate to use norms of tests that have not been standardised on the study population. Regardless of their source, all psychological variables to be used in longitudinal studies should be finalised in consultation with local informants and colleagues who are thoroughly familiar with the cultural and linguistic conditions of the study culture. The chosen measures must then be subjected to psychometric analyses. Both aspects of this work must be budgeted for.

2. Ensure that the adaptation process takes full account of the linguistic and cultural differences of the populations for whom adapted versions of the test are intended. Pre-pilot work is necessary to understand the ethnolinguistic context of the study participants, with sufficient time following the pre-pilot to allow changes to be made prior to a full pilot.

   In my experience, as far as possible adaptations should be developed in conjunction with mother tongue speakers of the assessment language, who grew up in the local area (or have resided there for some time), are familiar with local practice and the culture, and can make a judgement as to the likely suitability of the test materials and approach to test administration.

   It is also essential to be aware of local taboos such as avoidance of direct gaze from child to test administrator, and conventions for engagements with strangers, elders and members of the opposite gender. For young children in particular, it is advisable to position the administrator next to the child and not opposite. A face-to-face testing arrangement can induce anxiety, particularly in cultures where gaze avoidance is normative between children to adults. When designing timed tasks, consider the cultural view of speed and performance and whether the task will measure the desired construct if the salience of speed is different to that anticipated in the test. This is an issue for piloting.

3. When adapting tests, provide evidence that the language used in the instructions to the child and items themselves are appropriate for the cultural and language populations for whom the test is intended. Descriptions of how this was done should be available.

4. Materials familiar to children in the study area can enhance familiarity and engagement with the test (e.g. bottle tops or sticks for counting). On the other hand, novel presentations such as touch screens or presenting items on mobile devices can be very engaging even to children not that familiar with them (Pitchford and Outhwaite 2015). Use practice items and prompts, where appropriate, to reduce the chance of errors due to misunderstanding of task requirements.

5. Test instructions to the child for each item should be in both source and target languages to minimise the influence of unwanted sources of variation in administration. This procedure provides the administrator with both the source and translated instructions. Optimally, mother tongue speakers of the assessment language should be trained to carry out or assist in assessment procedures. This removes the need for translators to accompany an assessor who is unfamiliar with the child's language. If this is not possible and translation is required, this must be standardised across all translators to reduce the risk of poor reliability when more than one translator is used.

References

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

Adair, L.S., B.M. Popkin, J.S. Akin, D.K. Guilkey, S. Gultiano, J. Borja, L. Perez, C.W. Kuzawa, T. McDade, and M.J. Hindin (2011) 'Cohort Profile: the Cebu Longitudinal Health and Nutrition Survey', *International Journal of Epidemiology* 40.3: 619-625.

Baddeley, A., J.M. Gardner, and S. Grantham-McGregor (1995) 'Cross-cultural Cognition: Developing Tests for Developing Countries', *Applied Cognitive Psychology* 9.7: S173-S195.

Baltes, P.B., and J.R. Nesselroade (1979) 'History and Rationale of Longitudinal Research' in J.R. Nesselroade and P.B. Baltes (eds) *Longitudinal Research in the Study of Behavior and Development*, 1–39, New York: Academic Press.

Biersteker, L., and A. Dawes (2019) 'South Africa: Measuring up – The Sobambisana Evaluation' in H. Penn and A.T. Kjorholt (eds) *Early Childhood and Development Work: Theories, Policies, and Practices*, 91-112, Cham: Springer.

Boyden J., and D. Walnicki (forthcoming) 'Opportunities, Challenges and Strategies in Generating and Governing Longitudinal Data; Learning from two Decades of Research at Young Lives', Oxford: Young Lives.

Boyden, J., A. Dawes, P. Dornan, and C. Tredoux (2019) *Tracing the Consequences of Child Poverty: Evidence from the Young Lives Study in Ethiopia, India, Peru and Vietnam*. Bristol: University of Bristol Policy Press.

Boyden, J., A. Dawes, and C. Tredoux (2018) 'Improving Children's Chances: Using Evidence from Four Low- and Middle-income Countries to set Priorities for the Sustainable Development Goals' in S. Verma and A. Petersen (eds) *Developmental Science and Sustainable Development Goals for Children and Youth*, 257-275, Cham: Springer.

Boyden, J., and Z. James (2014) 'Schooling, Childhood Poverty and International Development: Choices and Challenges in a Longitudinal Study', *Oxford Review of Education 40*.1: 10-29.

Carter, J.A., J.A. Lees, G.M. Murira, J. Gona, B.G. Neville, and C.R. Newton (2005) 'Issues in the Development of Cross-cultural Assessments of Speech and Language for Children', *International Journal of Language and Communication Disorders* 40.4: 385-401.

Collins, L.M. (2006) 'Analysis of Longitudinal Data: The Integration of Theoretical Model, Temporal Design, and Statistical Model', *Annual Review of Psychology* 57: 505-528.

Crookston, B.T., W. Schott, S. Cueto, K.A. Dearden, P. Engle, A. Georgiadis, E.A. Lundeen, M.E. Penny, A.D. Stein, and J.R. Behrman (2013) 'Postinfancy Growth, Schooling, and Cognitive Achievement: Young Lives', *The American Journal of Clinical Nutrition* 98.6: 1555-1563.

Cueto, S., A. Miranda, J. León, and M.C. Vásquez (2016) 'Education Trajectories: from Early Childhood to Early Adulthood in Peru', Country Report, Oxford, Young Lives.

Cueto, S., and J. León (2012) *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. Muñoz (2009) *Psychometric Characteristics of Cognitive Development and Achievement Instruments in Round 2 of Young Lives,* Technical Note 15, Oxford: Young Lives.

Dawes, A., L. Biersteker, and L. Hendricks (2012) 'Towards Integrated Early Childhood Development: An Evaluation of the Sobambisana Initiative', Cape Town: Ilifa Labantwana. http://ilifalabantwana.co.za/wp-content/uploads/2016/04/An-Evaluation-of-the-Sobambisana-Initiative.pdf (accessed 14 August 2020).

Dercon, S., and P. Krishnan (2009) 'Poverty and the Psychosocial Competencies of Children: Evidence from the Young Lives Sample in Four Developing Countries', *Children, Youth and Environments* 19.2: 138–163.

Dornan, P., and A. Georgiadis (2015) *Nutrition, Stunting and Catch-up Growth,* Policy Brief 27, Oxford: Young Lives.

Dowd, A.J., I. Borisova, A. Amente, and A. Yenew (2016) 'Realizing Capabilities in Ethiopia: Maximizing Early Childhood Investment for Impact and Equity', *Journal of Human Development and Capabilities* 17.4: 477–493.

Economic and Social Research Council (2017) 'Longitudinal Studies Strategic Review: 2017 Report to the Economic and Social Research Council', https://esrc.ukri.org/files/news-events-and-publications/publications/longitudinal-studies-strategic-review-2017 (accessed 14 August 2020).

Elder Jr, G.H. (1998) 'The Life Course as Developmental Theory', *Child Development* 69.1: 1-12.

Favara, M. and A. Sanchez (2017) 'Psychosocial Competencies and Risky Behaviours in Peru', *IZA Journal of Labor and Development* 6.3: 1-40.

Georgiadis, A., L. Benny, L.T. Duc, S. Galab, P. Reddy, and T. Woldehanna (2017) 'Growth Recovery and Faltering Through Early Adolescence in Low- and Middle-income Countries: Determinants and Implications for Cognitive Development', *Social Science & Medicine*, 179. 81-90.

Greenfield, P.M. (1997) 'You Can't Take it with You: Why Ability Assessments Don't Cross Cultures', *American Psychologist* 52.10: 1115.

Gregoire, J. (2018) 'ITC Guidelines for Translating and Adapting Tests', *International Journal of Testing* 18.2: 101-134.

Grimm, K.J., P. Davoudzadeh, and N. Ram (2017) 'Developments in the Analysis of Longitudinal Data', *Monographs of the Society for Research in Child Development* 82.2: 46-66.

Hambleton, R.K., and A. Kanjee (1995) 'Increasing the Validity of Cross-cultural Assessments: Use of Improved Methods for Test Adaptations', *European Journal of Psychological Assessment* 11.3: 147-157.

Hambleton, R.K., and A. Zenisky (2010) 'Translating and Adapting Tests for Cross-cultural Assessment', in D. Matsumoto and F. van de Vijver (eds) *Cross-cultural Research Methods*, 46–74, New York: Cambridge University Press.

Hambleton, R.K. (2002) 'Adapting Achievement Tests into Multiple Languages for International Assessments', in *Methodological Advances in Cross-National Surveys of Educational Achievement*, 58-79, Washington, DC: The National Academies Press.

International Test Commission (2016) 'The ITC Guidelines for Translating and Adapting Tests (Second edition),' www.InTestCom.org, (access 14 August 2020).

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

León, J. (2020) *Equating Cognitive Scores across Rounds and Cohorts for Young Lives in Ethiopia, India, Peru and Vietnam,* Technical Note 51, Oxford: Young Lives.

León, J., and A. Singh (2017) *Equating Test Scores for Receptive Vocabulary Across Rounds,* Technical Note 40, Oxford: Young Lives.

Matafwali, B., and R. Serpell (2014) 'Design and Validation of Assessment Tests for Young Children in Zambia', *New Directions for Child and Adolescent Development* 146: 77–96.

Mwaura, P.A., and K. Marfo (2011) 'Bridging Culture, Research, and Practice in Early Childhood Development: The Madrasa Resource Centers in East Africa', *Child Development Perspectives* 5.2: 134–139.

McArdle, J.J., K.J. Grimm, F. Hamagami, R.P. Bowles, and W. Meredith (2009) 'Modeling Life-span Growth Curves of Cognition Using Longitudinal Data with Multiple Samples and Changing Scales of Measurement', *Psychological Methods* 14.2: 126-149.

Melhuish, E., M. Phan, K. Sylva, P. Sammons, I. Siraj-Blatchford, and B. Taggart (2008) 'Effects of the Home Learning Environment and Preschool Center Experience Upon Literacy and Numeracy Development in Early Primary School', *Journal of Social Issues* 64.1: 95–114.

Milfont, T.L., and R. Fischer (2015) 'Testing Measurement Invariance Across Groups: Applications in Cross-cultural Research', *International Journal of Psychological Research* 3.1: 111-130.

Millsap, R.E. (2010) 'Testing Measurement Invariance Using Item Response Theory in Longitudinal Data: An Introduction', *Child Development Perspectives 4*.1: 5-9.

NICHD Early Child Care Research Network (2002) 'Early Child Care and Children's Development Prior to School Entry: Results from the NICHD Study of Early Child Care', *American Educational Research Journal* 39.1: 133-164.

Nunes, T.N., A.D. Schliemann, and D.W. Carraher (1993) *Street Mathematics and School Mathematics,* New York: Cambridge University Press

Peña, E.D. (2007) 'Lost in Translation: Methodological Considerations in Cross-cultural Research', *Child Development* 78: 1255-1264.

Peña, E.D., and R. Quinn (1997) 'Task Familiarity: Effects on the Test Performance of Puerto Rican and African American Children', *Language, Speech, and Hearing Services in Schools* 28.4: 323-332.

Pisani, L., I. Borisova, and A.J. Dowd (2015) 'International Development and Early Learning Assessment Technical Working Paper', http://resourcecentre.savethechildren.se/sites/default/files/documents/idela_technical_working_paper_v3_nodraft.pdf (accessed 14 August 2020).

Pitchford, N.J., and L.A. Outhwaite (2016) 'Can Touch Screen Tablets be Used to Assess Cognitive and Motor Skills in Early Years Primary School Children? A Cross-cultural Study', *Frontiers in Psychology* 7: Article 1666.

Serpell, R., and B. Simatende (2016) 'Contextual Responsiveness: An Enduring Challenge for Educational Assessment in Africa', *Journal of Intelligence* 4.1: 3.

Snelling, M., A. Dawes, L. Biersteker, E. Girdwood, and C.G. Tredoux (2019) 'The Development of a South African Early Learning Outcomes Measure: A South African Instrument for Measuring Early Learning Program Outcomes,' *Child Care Health and Development* 45: 257–270.

Tredoux, C., and A. Dawes (2018) *Predictors of Mathematics and Literacy Skills at 15 years in Ethiopia, India, Peru and Vietnam,* Working Paper 179, Oxford: Young Lives.

Van Widenfelt, B.M., P.D. Treffers, E. De Beurs, B.M. Siebelink, and E. Koudijs (2005) 'Translation and Cross-cultural Adaptation of Assessment Instruments Used in Psychological Research with Children and Families', *Clinical Child and Family Psychology Review* 8.2: 135-147.

Watts, T.W., G.J. Duncan, M. Chen, A. Claessens, P.E. Davis-Kean, K. Duckworth, and M.I. Susperreguy (2015) 'The Role of Mediators in the Development of Longitudinal Mathematics Achievement Associations', *Child Development* 86.6: 1892-1907.

Yorke, L., and M.J. Ogando Portela (2018) *Psychosocial Scales in the Young Lives Round 4 Survey: Selection, Adaptation and Validation,* Technical Note 45, Oxford: Young Lives.

The project

The ‘Methodological Learning and Lessons from Young Lives’ project, funded by the ESRC, aims to strengthen capacity and effectiveness in the conduct of longitudinal research in low-and-middle-income countries, while also contributing to a growing community of practice. Through this project, Young Lives is reflecting on its experience and practices over fifteen years of research to create a dialogue with others involved in large-scale longitudinal studies in international development and related fields.

The author

Andrew Dawes is Associate Professor Emeritus in Psychology at the University of Cape Town and a Research Associate with Young Lives. Over the past 40 years he has undertaken research of relevance to child policy and programme implementation in South Africa, including on the development of children’s rights and well-being indicators for South Africa, the prevention of maltreatment and violence to young children, and evaluations of early childhood development programmes in low-income African settings. This work has included the development of standardised instruments suitable for the cross-cultural assessment of language, numeracy, and cognitive functioning in preschool children. Together with Colin Tredoux, he has used the Young Lives household survey data to undertake longitudinal analyses of growth in mathematics and receptive vocabulary in the Younger Cohort. He is one of the authors of Tracing the Consequences, which summarises the findings and policy implications of the first five Young Lives study rounds.

Acknowledgements

I am most grateful for the helpful comments provided by Gina Crivello, Jo Boyden and Rhiannon Moore on an earlier draft of this paper. Thanks to Garth Stewart for the design of this report, Adam Houlbrook for copy-editing and Emily Cracknell for shepherding production.

This Insights Report was funded by a grant from the Global Challenges Research Fund (GCRF), as part of the two-year project – Methodological Lessons and Learning in Young Lives funded by the Economic and Social Research Council (ESRC). The support of the ESRC is gratefully acknowledged.

GCRF
Global Challenges Research Fund

UKRI Economic and Social Research Council

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

www.younglives.org.uk
Tel: +44 (0)1865 281751 • Email: younglives@qeh.ox.ac.uk • Twitter: @yloxford

Young Lives