What mediates the effects of wealth on children's vocabulary developmental trajectories in Ethiopia? A longitudinal analysis.

Nardos Tesfay
Doctor of Philosophy in Education
University of Oxford, Wolfson College
Email: nardos.tesfay@education.ox.ac.uk

Supervisors: Professor Lars-Erik Malmberg
Department of Education, University of Oxford
Email: lars-erik.malmberg@education.ox.ac.uk

Dr. Maria Evangelou
Honorary Research Fellow
Department of Education, University of Oxford
Email: maria.evangelou@education.ox.ac.uk

Contents

Thesis abstract

This D.Phil study is motivated by gaps in educational research in low-income countries. First, it uses a child developmental, rather than economic, model to investigate how the timing of poverty shapes children's vocabulary trajectories and the mediational pathways through which poverty influences growth at different stages of childhood. Second, it investigates the reliability and validity of the measure used to assess vocabulary, addressing inter-cultural issues that have not been dealt with in the literature and challenging some of the assumptions underlying the use of psychometric approaches in low-income countries. Data from a sample of 1,321 children in Ethiopia, who took part in the Young Lives longitudinal study at ages five, eight, twelve and 15 years, were analysed. The Peabody Picture Vocabulary Test (PPVT) was used to measure vocabulary ability. A composite wealth index was used to measure living standards.

The psychometric properties of the PPVT were analysed within a Rasch measurement model. The test was screened for unidimensionality, item targeting, differential item functioning and scale stability. To compare the scores obtained at different time points, scores from the repeated test were calibrated onto a single metric using a test equating procedure. The Rasch estimates were obtained by anchoring the scores on the most stable items within each of respective language groups: Amharic, Oromifa and Tigrinya.

A multilevel model for change approach was used to estimate individual growth trajectories. On average, children in the highest wealth quintile outperformed children in the lowest quintile. The difference in the mean growth trajectories between top and bottom quintiles followed a constant path, except in Amharic in which the average gap narrowed by 0.47 SD. Initial wealth, measured at one year of age, predicted onset ability at five years for all language groups, and the growth rate between 5-15 years in Amharic. Changes in wealth did not further help to explain the variance in the intercept or the slope.

The analysis was extended to account for the factors that mediate the poverty-growth relationship. The results showed that 0-10% of the wealth effect on onset ability was mediated by the height-for age z-score, a measure of nutrition. Mothers' education mediated 1-8% of the wealth effect. Preschool participation mediated 6-14%. The findings reinforce existing literature about the importance of these factors in mitigating the effects of poverty on children's vocabulary development, particularly in early childhood. The study further highlights that expanding this knowledge base requires an in-depth understanding of the contexts in which children develop and learn, drawing on insights from substantive theories and local understandings of development and learning.

Acknowledgements

I wish to express my sincere appreciation to my supervisors, Professor Lars-Erik Malmberg and Dr. Maria Evangelou, who provided me with guidance, wisdom and relentless encouragement at every stage of the D.Phil, especially during the toughest times.

My deepest gratitude to my family, who have shown unwavering patience. Thank you for your kind interest in my work. “Mamma’s book” would not have been possible without you.

List of tables

List of figures

Map of Ethiopia

Map No. 4188 Rev. 5 UNITED NATIONS
March 2012

Department of Field Support
Cartographic Section

Source: United Nations, 2012.

Chapter 1 Introduction

This first chapter introduces the reader to the D.Phil study. The researcher's rationale is explored in the first three sections. Following this, the research questions are identified, and the subsequent chapters are outlined.

1.1 General rationale

The impetus for the D.Phil study is to understand how poverty influences children's vocabulary trajectories in Ethiopia, the extent to which poverty continues to shape trajectories beyond early childhood and what can be done to mitigate its effects throughout the course of childhood. There is broad consensus within the academic literature that children who experience poverty in early childhood have poorer cognitive outcomes than children who do not (Grantham-McGregor et al., 2007). Disparities between the rich and poor have been documented in both high and low-income countries, including Ethiopia (see Cueto et al., 2016). The negative consequences of developmental disadvantage in the first years of life for future learning and for other areas of human development, including health, employability and earnings are also well-established (Walker et al., 2007; Nelson, 2000 and Naudeau et al., 2011).

Despite the abundance of research on the effects of early childhood poverty on children's development and learning in low-income countries, little is known about how disadvantages evolve over the course of childhood in relation to children's socioeconomic contexts. The few studies that have documented such dynamics offer mixed results (Fernald et al., 2011; Reynolds et al., 2017; Hamadani et al., 2014), and conflicting answers to this question. The sociological hypothesis of cumulative advantage, referred to as the Matthew Effect in educational research (Stanovich, 1986), suggests that because later learning depends on earlier learning, "learning deficits are cumulative" and thus disparities widen with time (Jensen, 1966 and 1974a; Merton, 1968; DiPrete and Eirich, 2006). However, a widening gap is not necessarily inevitable. Disparities in cognitive outcomes may be mitigated or narrowed with time if, for example, disadvantaged children are provided with appropriate support for their development and learning (Hall et al., 2009; Barnett and Hustedt, 2005; Puma et al., 2010). Progress over time also depends on the cognitive outcome being studied. Skills, such as working memory and executive function, which are abilities that develop rapidly in early childhood (Nelson, 2000), may follow a more constant growth path in later childhood than abilities that are acquired slowly over time (Guo, 1998).

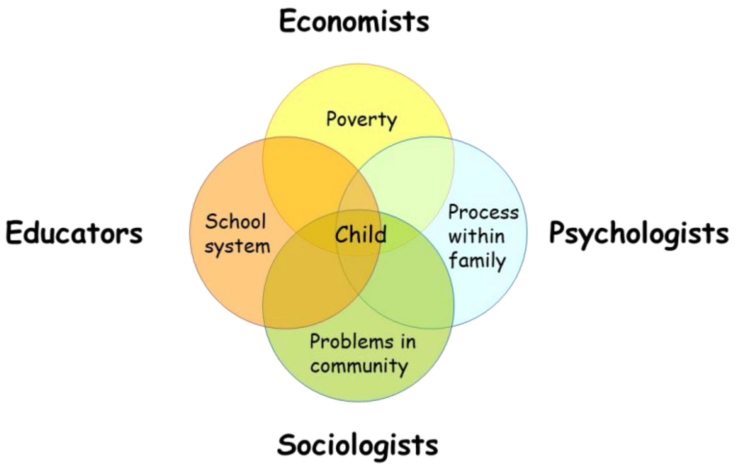
The focus of the current research is on understanding how poverty, experienced at different stages of childhood, shapes children's vocabulary developmental paths, and the mechanisms underlying this relationship. To what extent do family living standards in early childhood influence developmental trajectories? Does an increase in family living standards at a later stage of childhood improve a child's developmental trajectory? Does a decline in living standards contribute to developmental stagnation? These questions are particularly important in Ethiopia, a country that is characterised by persistent poverty, with many families moving in and out of poverty from one period to the next.

The observed gap in developing world educational research is largely due to the absence of quality longitudinal data that assess children's development and learning, while simultaneously tracking the wider environments that shape childhood outcomes. The measurement of children's development and learning is expensive, technical skill intensive and substantial investment in training is required for successful measurement (Fernald et al., 2009). Assessors require expertise in the developmental domain under study as well as experience in working with children. In developing country contexts, relevant human resources and competencies are often lacking. Longitudinal surveys are further challenged by high rates of attrition and the lack of funding for participant follow-up (Campbell and Rudan, 2011). Moreover, while children's socioeconomic contexts are widely researched in low-income countries, less is known about the processes, relationships and experiences that shape children's development. However, such longitudinal information is necessary for identifying opportunities to improve children's outcomes at all stages of childhood.

1.2 Theoretical rationale

Empirical research in low-income countries has not probed deeply enough to foster culturally meaningful and context informed models of children's development and learning. The recent expansion of interdisciplinary research on children in low-income countries has contributed to a growing awareness of different scientific perspectives about what shapes individual differences in childhood outcomes (as illustrated in Figure 1.1 below). However, much of the empirical work is still underpinned by economic theories of human development, due to the primacy of basic needs in the research contexts; the need to inform international development assistance programmes; and the need to monitor progress towards the achievement of national development and poverty reduction goals. However, knowledge from other disciplines is also needed to successfully address developmental and learning disadvantage: theories of environment and human development, individual differences in development, learning theory, as well as educational and psychological measurement theory.

Figure 1.1 Perspectives of how the environment shapes individual differences



Source: Sameroff and Fiese, cited by Shonkoff & Meisels, 2000

Currently, substantive insights about child development and learning from education, psychology and sociology may be integrated in research in low-income countries but are less extensively explored. The human development approach within economics evolved on the basis that economic growth does not necessarily lead to greater opportunities for all, and thus there is a need to enhance human capabilities directly through access to health, education and a decent standard of living (United Nations, 1995). Theoretical advancements in this field have further emphasised the social, empowerment and distributional role of investments in human development (Dreze and Sen, 1995). The concept has been extended to better understand individual differences in children's developmental outcomes, and to mobilise financial and technical support to improve children's outcomes in low-income countries. As articulated by the World Bank's Early Childhood Development Centre,

> Smart investments in the physical, cognitive, linguistic, and socio-emotional development of young children —from before birth until they transition to primary school— are critical to put them on the path to greater prosperity, and to help countries be more productive and compete more successfully in a rapidly changing global economy.
>
> (The World Bank Group, 2020)

The current study seeks to move this literature forward by drawing on theories of child development which posit that an understanding of children's environments is central to improving developmental outcomes. *An understanding of context* is regarded as a key principle of the D. Phil study, and guides each stage of the research with theoretical, methodological, and analytical implications. The notion of context is the common thread linking the overlapping strands of literature that are reviewed and the subsequent empirical chapters that follow. The researcher's perspective is that any investigation of childhood outcomes must be rooted in an understanding of the environments in which children's development takes place.

At a theoretical level, emphasis is placed on articulating clearly defined constructs of inquiry in terms of the outcome of interest, i.e. vocabulary development, as well the contextual influence, i.e. living standards. The theoretical literature is reviewed to document divergent views of how both living standards and children's wider contexts should be conceptualised and modelled. Moreover, the empirical literature is examined not only to contemplate the impact of poverty on growth, but also to better understand how the measures used to assess both poverty and cognitive outcomes have been contextualised in the various research settings.

From a methodological and analytical standpoint, the research takes a more cautious approach to measurement in comparison with the conventional literature. The literature review seeks to understand what constitutes meaningful operational measures of both living standards and cognitive ability, and the implications for the present study. Measures of poverty and child development are not adopted uncritically. Rather, the research asks, given the nature and the extent of poverty in Ethiopia, what measure of living standards is most appropriate for addressing the research questions? Moreover, what constitutes a valid and reliable measure of cognitive ability? Thus, the review explores the concepts underlying the measures, the operational measures themselves, and their fit-for-purpose. The importance of these questions is heightened when instruments are used in a multicultural or multilingual setting, or when transferred from one context to another.

1.3 Contextual rationale

Ethiopia has made significant progress towards reducing poverty in recent decades. Between 2000 and 2011, the share of the population living in poverty reduced by over one-third (World Bank, 2015). Moreover, between 1995 and 2015, the proportion of people living in absolute poverty, i.e. those who are severely deprived of their basic human needs (United Nations, 1995), declined from approximately one-half to one-quarter (Araya and Woldehanna, 2019). Advancements in household living standards have been attributed to steady economic expansion in both urban and rural areas, with a focus on the service sector and agricultural development (World Bank, 2015). The Productive Safety Net Programme (PNSP), established in 2005, has been cited a one of the main mechanisms for lifting people out of poverty in rural areas and smoothing negative economic shocks overall (Araya and Woldehanna, 2019).

Despite steady improvements in well-being, the persistence and cyclical nature of poverty in Ethiopia renders it a structural problem. Households often move in and out of poverty over periods of time (Dercon and Krishnan, 2007). Rural households are particularly vulnerable to this precariousness as many are unable to protect their consumption against shocks due unpredictable rainfall, drought and rapid food price inflation (Hill and Porter, 2017). Moreover, according to the Multidimensional Poverty Index (MPI), which conceptualises poverty as multiple deprivations that are simultaneously experienced, poverty in Ethiopia only reduced by 10% during 2000 and 2011 (Carranza and Gallegos, 2013). The MPI characterises an individual living in poverty as someone who is below the threshold in one-third or more of the weighted indicators (Alkire et al., 2011). In 2011, it was estimated that 87% of the Ethiopian population was MPI poor (Carranza and Gallegos, 2013).

According to the Multiple Overlapping Deprivation Analysis (MODA), which is a non-monetary measure of poverty conceptualised based on the Multidimensional Poverty Index (MPI), 88% of children in Ethiopia are deprived in at least three key dimensions and thus multidimensionally poor (CSA and UNICEF, 2018). The MODA studies nine dimensions of well-being overall. Stunting, nutrition, health, water, sanitation and housing are considered core dimensions. Other dimensions include education, health-related knowledge and information and participation. Virtually all Ethiopian children are deprived in at least one of the six key dimensions (CSA and UNICEF, 2018). For children under five, deprivations were largest for housing, sanitation and nutrition. In terms of children's learning, low educational achievement and school retention are critical challenges for Ethiopia's educational system. Since the introduction of the school-fee abolition programme, Ethiopia has witnessed a surge in enrolment which has put pressure on the educational system. Huge wealth, regional and urban-rural disparities persist across schooling participation indicators (DHS data tables, 2011). Composite scores from the National Learning Assessments showed that mean student achievement among Grade 4 and Grade 8 students was below the minimum requirement of 50% for grade promotion (NOE, 2011: data tables). The results further showed that student performance declined during the period of educational expansion, from 48% in 2004 to 40% in 2013 for grade 4s, and from 40% to 35% for grade 8s (NOE, 2004a, b, 2011: data tables).

While poverty levels are among the highest in the world, income inequality is moderate compared with other sub-Saharan African countries (WIDER, 2007). However, inequalities are growing with the expansion of economic growth. The Gini coefficient, which is an index that measures the distribution of wealth across the population, is modest relative to neighbouring countries but rising. Inequality has risen from 0.29 in 1995/6 to 0.33 in 2015/16 (Araya and Woldehanna, 2019). A coefficient of zero indicates a perfectly equal distribution of income in a population, and one indicates perfect inequality. Moreover, horizontal inequalities, which refers to inequalities between cultural groups rather than individuals, are significant.[1] In terms of children's access to basic services, such as education and health care, large disparities have been documented between rural and urban areas (Woldehanna et al, 2008), by language/ethnicity (Tesfay and Malmberg, 2014), and across geographic regions (CSA and UNICEF, 2018).

1.4 Research aims and questions

The current study seeks to address the following *overarching research question*: how does poverty shape children's vocabulary development in Ethiopia? The overall aim of

[1] Stewart (2000) defines cultural groups as groups encompassing common cultural identities such as ethnicity, language, race, religion, or class (Stewart, 2000: 253).

the research is to better understand how poverty, experienced at different stages of childhood, shapes children's developmental paths and the different pathways through which children's development and learning can be promoted in contexts of poverty and adversity. To carry out the investigation, the study draws on the Young Lives prospective longitudinal survey of children who were followed up over a fifteen-year period from their first year of life through to adolescence. The Young Lives survey uses the Peabody Picture Vocabulary Test (PPVT-III) as a proxy measure of cognitive ability, since the test has been shown to correlate strongly with varied measures of intelligence in other samples (Dunn and Dunn, 1997; Campbell, 1998; and Campbell et al., 2001; Gray et al.,1999). The test was administered to the sample of children when they were five, eight, 12 and 15 years old. The wealth index is the primary instrument used in the Young Lives survey to measure living standards.

The overall research aim and question is broken down into more *specific research objectives and questions*. The first question addresses the effectiveness of the assessment instruments that were used to measure children's cognitive ability among the multicultural and multilingual sample of children. The second question asks to what extent household living standards, as measured by household wealth, influence cognitive trajectories? Finally, what are the mechanisms through which wealth influences developmental outcomes? These broader questions correspond to three sets of descriptive and explanatory subsidiary research questions:

1.1 To what extent is the Peabody Picture Vocabulary Test (PPVT-III) a valid and reliable proxy measure of vocabulary ability among the sample of children?

1.2 How consistently does the PPVT perform across children from different linguistic backgrounds and across time?

1.3 To what extent can growth trajectories be reliably estimated, considering the cultural differences across the sample vis-à-vis language?

2.1 Based on the constructed metric of children's abilities, to what extent do individual developmental trajectories differ in relation to the overall mean trend, in terms of initial status, the growth rate and the shape of the trajectory?

2.2 What is the effect of wealth on vocabulary developmental paths, i.e. how does wealth status in early childhood predict initial ability and the growth rate? How do patterns of development differ between the top and bottom wealth quintiles?

2.3 To what extent is change over time in wealth associated with patterns of development (i.e. change in ability)?

3.1 How is the relationship between wealth and children's vocabulary development explained by the children's nutritional status, their home environment and preschool exposure?

1.5 Structure of the thesis

First, the reviewed academic literature is discussed in Chapter 2. The purpose of the review is to establish the foundations for a quantitative study of the relationship between poverty and children's cognitive growth over the course of childhood. The overall aim is to identify a context informed conceptual framework for the study. The chapter explores theories of environment and human development to better understand the role of context and culture in children's development in general, and specifically in relation to children's cognitive development and vocabulary development in particular. The review further examines the empirical evidence on the effects of childhood poverty on cognitive outcomes, with a focus on evidence from low-income countries and the measures used in the research. Different concepts of poverty are assessed to identify an appropriate construct and operational measure of poverty for the current enquiry. Finally, the main measurement issues in cross-cultural educational and psychological assessment, that pertain to the present study, are discussed. The aim of this final component of the review is to better understand the properties of good measures with a view to establishing a valid and reliable metric of children's vocabulary abilities over time.

Chapter 3 discusses the methodology for the D.Phil study, drawing on the findings of the literature review. It introduces the Young Lives longitudinal survey with a focus on the survey design aspects. This section presents the data collection process, the instruments used in the analysis, as well as sampling, attrition, and missing data. The chapter provides the background to the Peabody Picture Vocabulary Test (PPVT), and the process of contextualising the PPVT instruments for the Ethiopian context. The research designs for each of the three empirical studies are then presented and justified. Each stage of the research design describes how the data are used. The analytical strategy for the psychometric analysis is discussed, including a justification for using the Rasch measurement model and an explanation of how the model is used. The statistical models that are used to estimate children's growth trajectories are then discussed. The first model tests the magnitude and significance of the effect of wealth in early childhood on vocabulary development. The second model investigates whether changes in wealth over time further help to explain patterns of development. The chapter then discusses how the statistical models are adapted to investigate the mechanisms that mediate the poverty-vocabulary growth relationship. The ethical considerations associated with secondary data analysis are presented in the final section.

Chapter 4 addresses the first set of research questions on the performance of the PPVT when used across the multilingual sample of children and across time. The chapter seeks to assess the reliability and validity of the measure, as well as the consistency of the measure over time and across linguistic subgroups. The analysis contributes to the emergent literature on the robustness of educational and psychological measures when used in cross-cultural research. It further seeks to illustrate the importance of test localisation for construct validity. The chapter concludes by discussing the extent to which growth trajectories can be reliably estimated based on the PPVT measure.

Chapter 5 addresses the second set of research questions about the effects of poverty on children's vocabulary developmental trajectories. The aim of this component of the research is to better understand patterns of change in the relationship between poverty and children's vocabulary ability over time. The PPVT data that were scaled according to the Rasch methodology are modelled to establish the mean trend in vocabulary development for each language subgroup, to investigate the extent to which living standards in early childhood explain individual differences in development, and whether changes in living standards over the course of childhood further help to explain changes in development. Thus, two statistical models are discussed. The results of the "launch model" of developmental change are presented to discuss how wealth in early childhood influences vocabulary trajectories. The results of the "change-to-change" model are then discussed to determine if changes in wealth predict developmental change.

Chapter 6 presents the findings of the final empirical study. It investigates how the relationship between wealth and children's vocabulary development is explained by the children's nutritional status, their home environment and preschool exposure. The purpose of this final stage of analysis is to identify the pathways through which poverty affects children's vocabulary development. Such pathways can be viewed as "resilience promoting" mechanisms such that they promote "positive patterns of adaptation in the context of risk and adversity" (Masten and Gewirtz, 2006: 23). Understanding these developmental and resilience processes are necessary for the design of policies and interventions that address developmental disadvantage at all stages of childhood.

Chapter 7 discusses the findings of the research in relation to each of the three sets of research questions. This chapter is presented as an analytical discussion chapter, and thus, the results are interpreted with reference to the literature review and the broader aims of the research. This chapter also concludes the thesis. The lessons emerging from the D.Phil study are discussed, the limitations of the study are identified, and future research directions are highlighted.

Chapter 2 Literature review

2.1 Introduction

In reiteration, the present study seeks to address the following overarching research question: how does poverty shape children's vocabulary development in Ethiopia? The aim of the research is to better understand poverty effects on developmental trajectories and the different pathways through which children's development and learning can be promoted in contexts of poverty. To carry out the investigation, the study draws on the Young Lives prospective longitudinal survey of children who were followed up over a fifteen-year period from infancy through to adolescence. The broad research aim of the D.Phil study is broken down into more specific research objectives and questions. The first question addresses the effectiveness of the assessment instruments in measuring children's vocabulary ability. The second question asks to what extent household living standards, as measured by household wealth, shape vocabulary trajectories? Finally, what are the mechanisms through which wealth influences developmental outcomes?

The literature review is divided into three parts. The findings from the literature review will inform the subsequent empirical components of the research. The first section establishes the theoretical origins of the study. It explores the role of context and culture in children's development in general, in relation to children's *cognitive* development, and vocabulary development in particular. The purpose of the section of the review is to establish a foundation for a quantitative study of the relationship between poverty and vocabulary growth and to identify an appropriate model of development, one that is culturally meaningful, and context informed.

The second part of the review examines evidence of the effects of poverty on children's cognitive development, with a focus on research in low-income countries. The section begins with a review of the empirical evidence on the effects of childhood poverty on cognitive development. The primary aim is to document the effects of poverty in early childhood on development and to establish whether changes in socioeconomic conditions over the course of childhood further help to explain developmental paths. As a secondary objective, consideration is given to the quality of the measures used in the reviewed studies. Thirdly, the pathways through which poverty shapes trajectories are examined to identify the contextual influences that vary with living standards and shape cognitive outcomes. This is followed by a review of poverty concepts to identify an appropriate construct and operational measure of poverty for the current enquiry.

The final part of the review identifies the main measurement issues in cross-cultural educational and psychological assessment that pertain to the present study. The purpose of this component of the review is to better understand the properties of good measures and to set up a framework for assessing the reliability and validity of the PPVT instrument used in the current research. The issues that are highlighted in the review are as follows: (i) cultural sources of test bias and the validity of the claims made using biased test data; (ii) test translation vs. test adaptation as a process of decomposing broad concepts into culturally meaningful constructs and operational measures; (iii) transparency and documentation of test development, adaptation and administration processes; (iv) the focus on solving methodological problems rather than substantive considerations about the appropriateness of measures; and related to this point (v) the reliance of psychometric research on statistical theory ahead of educational and psychological theory.

2.2 Theoretical origins for the study

This section of the literature review establishes the theoretical origins of the study with a view to develop a framework for a quantitative study of the relationship between poverty and vocabulary growth, one that is culturally meaningful, and context informed. The review seeks to better understand how context and culture have been studied in relation to children's development and learning. The first part, section 2.2.1, discusses the main theoretical perspectives within the social sciences on the role played by children's wider contexts in shaping their development. Section 2.2.2 looks at the role of context and culture in relation to children cognitive development, and vocabulary development specifically. Section 2.2.3 discusses the issue of stability in cognitive outcomes. The final section 2.2.4 explores influences on children's vocabulary development.

2.2.1 Context and culture in children's development

The influence of context on child development has been widely researched across various disciplines in the social sciences. This section of the review explores more recent debates, launched in the 1970s, which sought to formalise the child-context relationship. Much of this earlier research was guided by criticism of prevailing theories within cognitive psychology, most notably the ideas of Jean Piaget. Piaget asserted that children's psychological processes develop internally according to universal stages of development (e.g. Piaget, 1952; 1983), with no substantial role played by social experience or cultural exposure. Early post-Piagetian research questioned both the notion of universality in development and the lack of attention given to external sources of influence. Regarding the latter, critics argued that individual processes could not be understood without considering the social reality within which the person develops. This shift in thinking is summarised by Vygotsky (1978), "human learning presupposes a specific social nature and a process by which children grow into the intellectual life of those around them" (Vygotsky, 1978: 88). In other words, by interacting with others, children's learning and development is a social and cultural process.

Vygotsky's publication of *Mind in Society* (1978) revealed among the earliest efforts to formalize the social interactional origins of development. Vygotsky's concept of the zone of proximal development (ZPD), which relates the process of learning to social settings, exemplifies this transformation. Vygotsky defines ZPD as,

> … the distance between the actual developmental level as determined by independent problem solving and the level of potential development as

> determined through problem solving under adult guidance or in collaboration with more capable peers.
>
> (Vygotsky, 1978: 86)

For Vygotsky, studies of children's development need to account for both the capabilities that have already matured, "the fruits of development" (Vygotsky, 1978: 86), as well as those that are in the process of being formed, "the buds or flowers of development" (Vygotsky, 1978: 86). Capabilities that have not yet reached maturation require assistance from external actors: parents, teachers, and peers, and through this social interaction developmental paths are shaped. His reasoning is articulated as follows,

> … learning awakens a variety of internal developmental processes that are able to operate only when the child is interacting with people in his environment and in cooperation with his peers. Once these processes are internalized, they become part of the child's independent developmental achievement.
>
> (Vygotsky, 1978: 90)

Early efforts to model the relationship between context and individual development is further demonstrated by Bruner who conceptualised the notion of *scaffolding*, defined as a "process that enables a child or novice to solve a problem, carry out a task or achieve a goal which would be beyond his unassisted efforts…" (Wood et al., 1976: 90). This work aligns with Vygotsky's vision, asserting that the developing child requires a supportive framework, with emphasis on guidance from adults, and does not occur in isolation from others.

Building on this, Bronfenbrenner endeavoured to explicitly model the layers of social interaction involved in child development in bio-ecological systems theory, initially published in *The Ecology of Human Development* (1979). According to Bronfenbrenner, children's development is a product of both biological and environmental factors and takes place within various social settings, both proximal and distal. Proximal influences involve children directly. These are the "engines of development" and relate to the interactions in the child's immediate environment (Bornstein, 2012: 19). Distal processes are less direct, relating to the “nonfamilial social structures, settings, or other related circumstances that help shape the course of development” (Bornstein, 2012: 24), such as school, peers, and culture. Four levels of environmental influence were initially hypothesised by Bronfenbrenner, hierarchically according to the degree of proximity to the child (Bronfenbrenner, 1979). The *microsystem* refers to the direct interactions between the child and the social settings, such as interpersonal relationships with parents. The *mesosystem* involves the interactions between two or more microsystems, e.g. between the child’s home and school.   The *exosystem* refers to settings that have an indirect influence on the developing child, e.g. parent’s workplaces. The *macrosystem* consists of the overarching institutional structures within which the micro-, meso-, and exosystem are embedded, such as political, cultural, and religious contexts. These descriptions were extended to include the *chronosystem*, which refers to “change [.] over time not only in the characteristics of the person but also of the environment in which the person lives…” (Bronfenbrenner, 1994: 40). The bio-ecological framework has been hugely influential in human developmental research. Bronfenbrenner’s later writings reveal a broadening of his understanding of how the individual engages with the systems, highlighting the role played by individuals themselves in shaping their own development (Bronfenbrenner, 1989), a notion that was absent from his earlier theoretical work.

Sameroff’s Transactional Model of Development sought to address the limitation in the literature which downplayed the role of the child in shaping their own development. The transactional model conceptualises a bidirectional relationship between the individual and the environmental (Sameroff and Chandler, 1975; Sameroff, 1983; Sameroff and Fiese, 2000; Sameroff, 2009). Children are not viewed as passive recipients of external sources of influence. Rather, it is argued that children develop through engagement with their wider contexts, which in turn contributes to their own personal outcomes. The transactional model rationalises a more simplified notion of context in comparison with Bronfenbrenner’s approach. Three systems are proposed, which take into account both the characteristics of the child and the various social settings. The *genotype* refers to innate biological sources of influence. The *phenotype* relates to the child's observable characteristics, such as the child's physical appearance. The *environtype* includes external social influences, such as the family, peers, and the community. A child's behaviour at a given point in time is considered the product of transactions between the three categories of influence (Sameroff, 2009: 14). Sameroff's typologies of influence are closely linked with Bronfenbrenner's nested systems. However, Sameroff draws closer attention to the significance of evolving non-biological individual characteristics such as personality, attitudes, and behaviour. Therefore, the main contribution of the transactional model lies in recognising that the relationship between individual development and environmental events or conditions are an endogenous two-way process, and in a constant state of transition. It demonstrates an important transformation in thinking, from perspectives that view children's development as reactions to external influences, a one-way cause and effect relationship, towards a theory of the child that is underpinned by the active participation of the child in structuring their own world (Sameroff, 1975, 2009).

Up to this point, the review has documented the pioneering research that spearheaded the transformation towards context informed understandings of children's development and learning. However, despite this broadening of perspective and the expansion of research, the notion of culture has been ambiguously applied in quantitative educational research, often limited to socio-demographic indicators of cultural difference such as ethnicity or religion. The remaining part of this review considers cultural understandings of child development in the academic literature: Whiting and Whiting's psychocultural model, Super and Harkness's Developmental Niche framework, and Rogoff's view of "cultural apprenticeships".

Whiting and Whiting's psychocultural model (1975) approached the cultural influences on development in a more tangible way than convention at the time. The anthropologists disagreed with the expression of culture as a broad level of external influence on the child, insisting on “unpackaging” the concept. Cultural practices and beliefs were described as processes derived from environmental conditions, historical experiences, and everyday life. They argued that children develop within a *cultural learning environment* characterised as “… an activity in progress, a physically defined space, a characteristic group of people, and norms of behaviour” (Whiting, 1980: 97). By moving towards an operational definition of culture as a process, consistent with contemporary theories of development (e.g. Bronfenbrenner, Sameroff), the Whitings’ theory helped forge a bridge between anthropology and developmental psychology and paved the way for empirical studies in this area.

Super and Harkness’s highly influential Developmental Niche framework sought to refine the notion of a cultural learning environment further, specifying how culture manifests in children’s lives and drives developmental paths. They hypothesise that a child’s progress is the product of the relationship between his/her “inherited dispositions” (Harkness et al., 2007: 34S) and three interacting subsystems of influence: (i) the child’s physical and social settings; (ii) the customs of child care; and (iii) the psychology of caretakers, e.g. ethnotheories about child behaviour and child rearing (Super and Harkness, 1986: 551-558). However, Super and Harkenss disagree with the Whitings’ implied direction of causality, arguing that belief systems shape the environmental conditions around the developing child as well as the everyday practices of parenting, and not the other way around.

As the theory of environment, culture and human development continues to evolve, concerns about the validity of causal assumptions and of the treatment of development and context as separate entities have been echoed across disciplines. As argued by Rogoff,

> … it is incomplete to focus only on the relationship of individual development and social interaction without concern for the cultural activity in which personal and interpersonal actions take place. And it is incomplete to assume that development occurs in one plane and not in others (e.g., that children develop

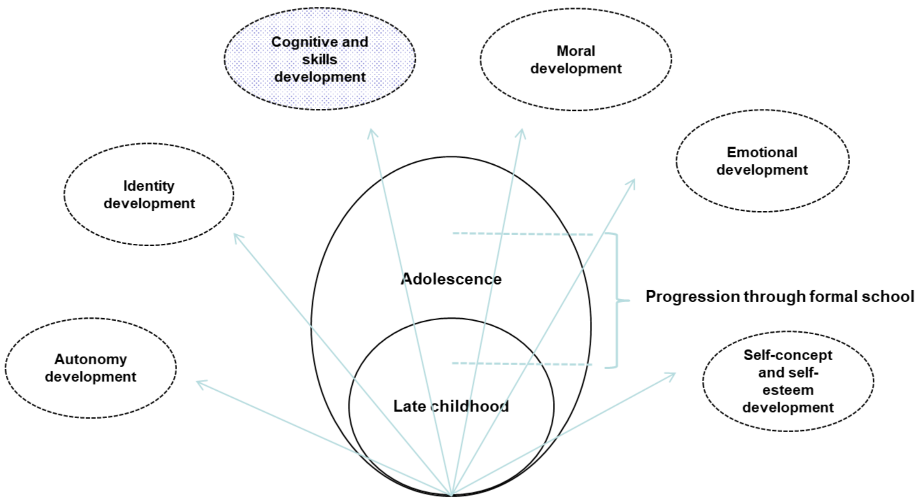
> but that their partners or their cultural communities do not) or that influence can be ascribed in one direction or another or that relative contributions can be counted (e.g., parent to child, child to parent, culture to individual).
> (Rogoff, 1995, p. 141)

Rogoff's theory of cultural apprenticeships stems from the premise that the child develops as he/she "participate[s] in and contribute to cultural activities that themselves develop with the involvement of people in successive generations" (Rogoff, 2003: 52). Thus, more recent theoretical trends emphasise greater "cultural competence" in research, and appreciation for the inherent variability in processes of human development. This transformation has important implications for the conceptualisation, measurement, and analysis of children's development in different cultural contexts, and particularly in multicultural settings. From this more nuanced perspective, development should be understood as a complex process, "linked to a child's everyday experiences, which are embedded in the cultural practices or "scripts" of his or her family and society" (Shonkoff and Phillips, 2000: 68).

2.2.2 Context and culture in children's cognitive development

Moving from broad contextual explanations of children's development, this section of the review focuses on children's cognitive development specifically (Figure 2.1). It introduces the notion of *cognitive systems* from the natural sciences, advancing the prevailing view that the environment and experiences in early life are central to children's development. Within the natural sciences, there is an emphasis on the biological processes driving development and on the critical periods of life that are most important for cognitive development. Two different strands of literature are discussed: developmental neuroscience and behavioural genetics. The aim is to better understand how the role of environment is conceptualised and approached methodologically within the two fields.

Figure 2.1 Different domains of children's development



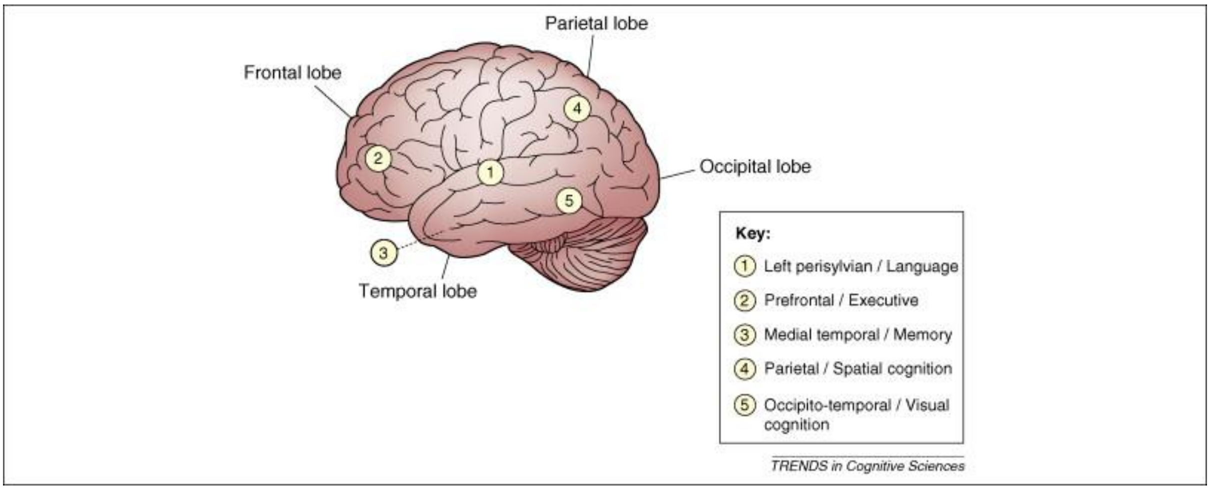
Source: author's own illustration

In recent decades, there has been tremendous progress towards understanding how individual experiences, environmental conditions and biological factors interact to shape cognitive outcomes, as evidenced by rapid theoretical and methodological advancements in the field of developmental neuroscience. The research emerging from these debates has produced important insights into how cognitive systems evolve throughout the course of childhood. We now know that the earliest years of life constitute the most critical period for brain development (Shonkoff and Phillips, 2000), that cognitive abilities are largely formed during this early phase of life (Nelson, 2000), and that higher level cognitive abilities such as the prefrontal and executive functions are strongly associated with intelligence (Grey and Thompson, 2004; Hackman and Farah, 2009; Farah et al., 2006).

The prefrontal cortex develops at a rapid speed in the early years (Figure 2.2). It predicts the formation of other cognitive systems located in different regions of the brain, including vocabulary and language, memory, and spatial and visual processing (Nelson, 2000; Noble et al., 2007; Malkova et al., 2000; Farah et al., 2006). It is now well-established that abnormalities in the development of the frontal region and delays in the formation of core cognitive functions may have long lasting consequences for children's development and can cause cumulative damage over time (Shonkoff et al., 2009).

Within developmental neuroscience, a wide range of direct observation-based techniques are used to assess cognitive activity using tools such as electrophysiological measures and neuroimaging. Individual differences in development and behaviour are explained mainly by biological processes, which are viewed as the driving force behind growth in cognitive systems. Associations between brain development and environmental influences such as nutritional factors, abusive or neglectful treatment, exposure to toxins and disease are increasingly being researched in relation to neural development (Shonkoff et al, 2012; Turecki et al., 2016; McLaughlin et al., 2017). However, investigating such effects on brain activity is challenged by the need for “intrusive procedures” to facilitate direct observation (Shonkoff and Philips, 2000: 41).

Figure 2.2 The five basic neurocognitive systems



Source: Hackman and Farah, 2009: 66.

Whereas in cognitive neuroscience emphasis is placed on understanding the biological mechanisms driving brain development, behavioural genetics stresses the isolation of biological processes to pinpoint the contextual factors that moderate their effects on childhood outcomes. The focus on contextual processes implies less intrusive techniques. Methodological advancements have made it possible to explain the strength of heritable influences on development, and to isolate the role played by children’s wider environments and experiences in shaping development. The two main research designs, longitudinal twin and adoption studies, partition the variance in individual developmental outcomes by genetic and environmental factor. The Twins' Early Development Study (TEDS) provides one of the most prominent examples (Oliver and Plomin, 2007; Trouton et al., 2002). TEDS is a nationally representative longitudinal survey of twins in the UK, which tracks over 15,000 families who are followed from birth for over 20 years. The study has produced important insights into how genes and environment interact to shape cognitive and behavioural outcomes (Asbury et al., 2005; Plomin and Davis, 2006). The research has further documented how this relationship can be moderated by appropriate responses to genetic sources of risk (Asbury and Plomin, 2005), or put differently, "the environmental contexts in which heritable influences are expressed" (Shonkoff and Phillips, 2000: 41).

The advantage of research designs in behavioural genetics lies in the possibility to explain how contextual factors modify the influence of genes on cognitive outcomes. As an illustrative example, researchers have shown that among children in families with high socioeconomic status, individual differences in cognitive ability are largely explained by genes as "environmental influences are in effect 'at ceiling' in this group", whereas among poor families, variation in ability is attributed to environment, which is "the limiting factor" in contexts of poverty (Hackman and Farah, 2006: 69; also see Turkheimer, 2003). A second advantage relates to the longitudinal aspect, which offers the possibility to investigate the causal processes driving development and the dynamic nature of that process. In other words, the moderating effect of a given contextual factor can be studied at different stages of childhood to identify when its effect is most significant.

It is important to recognise the differences between the two strands of research. Developmental neuroscience focuses on the first few years of life based on the reasoning that the brain's capacity for change decreases with age (National Scientific Council on the Developing Child, 2007). This period of life is viewed as critical not only for the development of core functions but also for future learning, health, and behaviour (Nelson, 2000). Individual differences in cognitive abilities are assumed to form throughout the early years and then to stabilise. Early developmental gaps are thought to persist and potentially widen throughout childhood, leading to disparities in other areas of human development (see Nelson, 2000).

While behavioural genetics also stress the importance of early cognitive outcomes for later development and learning, the focus is on explaining heritable patterns in cognitive outcomes (Asbury and Plomin, 2013; also see Plomin et al., 2013 and Plomin, 1986). The factors that underlie the development of a cognitive outcome are examined at different stages of childhood to isolate both the genetic and the environmental effects on its level/trajectory. Consistent with findings from developmental neuroscience, gene studies have shown that genes are the driving force behind the stability of cognitive outcomes, that environmental influences help to explain individual differences, and that these effects are time specific and decline with age (Hoekstra et al., 2007). More recent research argues that this does not mean that environment is less important than genetics, rather "genetic influences on cognition are the result of accumulating environmental experiences and depend on exposure to high-quality environmental contexts over time" (Tucker-Drob et al., 2013: 349).

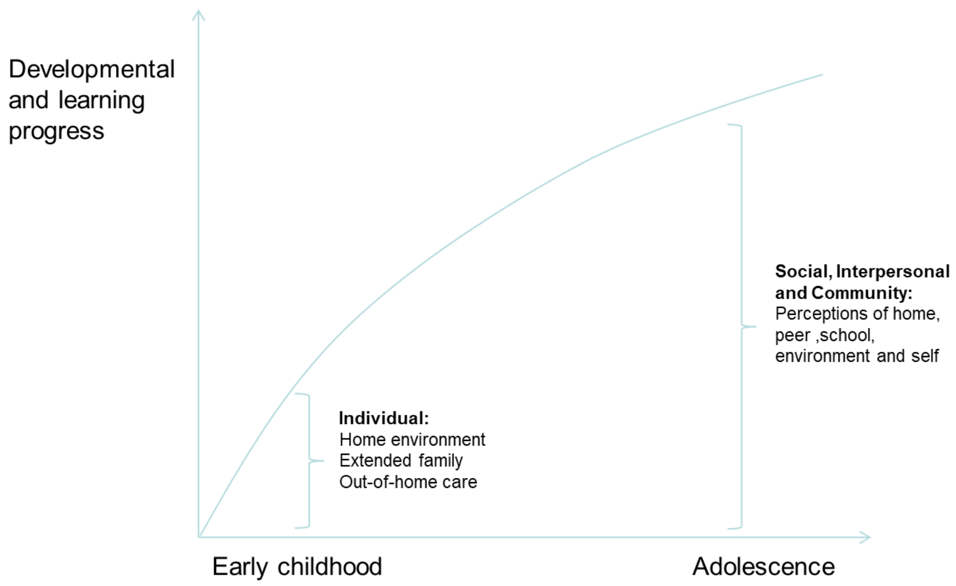
Reflecting on both strands of literature draws attention to the importance of defining each phase of childhood explicitly when conducting longitudinal research on children's development. Early childhood generally refers to the age group spanning birth to five years (Grantham-McGregor et al., 2007; Walker et al., 2007; Engle et al., 2007). Graber and Brooks-Gunn define two separate cohorts in early childhood – the prenatal period up to 2 years and 3-6 years (Graber and Brooks-Gunn, 1996). This phase of life is characterised by rapid changes in a child's physical, cognitive, language and socioemotional development. Both definitions underscore the significance of early childhood as a critical period of life (Shonkoff and Phillips, 2000). Late or middle childhood corresponds to the first years of primary school, age 7-10 years, which is followed by adolescence, age 11-19 years (Graber and Brooks-Gunn, 1996).

2.2.3 Stability in cognitive outcomes

The empirical research has shown that the degree to which context shapes cognitive development depends on the environmental influence, the specific cognitive construct under study and the stage of childhood. The literature reviewed so far has documented the importance of biological endowments and family processes for the development of children's cognitive skills and abilities in early childhood. Both nature and nurture play a vital role in establishing foundational skills

In middle childhood and adolescence, as interactions with peers, the school and community expand (Figure 2.3), biological characteristics and family experiences become more distal influences (Mayes and Lewis, 2013; Chase-Lansdale et al., 1997). The start of middle childhood coincides with the beginning of formal schooling. Developmental tasks broaden beyond the physical, cognitive, and socioemotional aspects to include the development of the individual (i.e. self-concept, personality and motivation) and the cultivation of social relationships (Collins, 1984). Adolescence, which is the final stage of childhood, begins with the onset of puberty. During this phase of life, identity formation and self-direction become primary developmental tasks (Backes and Bonnie, 2019).

Figure 2.3 Influences on children's learning and development throughout childhood



Source: author's illustration drawing on Aber et al., 1997.

Thus, a distinction has been made in the literature between cognitive ability and educational achievement, two different but connected concepts. *Academic achievement* refers to the extent to which a student has accomplished the educational goals set out in different domains of learning (Steinmayr et al., 2014). The term *cognitive ability* is used to describe the primary cognitive functions largely formed in early childhood, which set the stage for future development and learning. Thus, whereas achievement is acquired through formal schooling and settings outside the home, ability is conceptualised as a process that "tends to be determined by both environmental and genetic factors in early life" (Guo, 1998: 257). Cognitive transformations continue throughout childhood but are not as dramatic as in early childhood: (i) behaviour changes gradually not abruptly; (ii) children develop at different rates in different domains; and (iii) different children develop in different ways (Fischer and Bullock, 1984).

The contextual factors that are important for academic achievement may not explain individual differences in cognitive ability. Due to its acquired nature, achievement is shaped by a range of factors that have varying degrees of influence depending on the stage of childhood (Draper, 1974; Carroll, 1974; Scriven, 1974; Guo, 1998, Plomin, 1994). As an illustrative example, consider the concept of "motivation to learn" and its effects on academic achievement. In the early primary school years, basic literacy and numeracy is just beginning to develop and academic self-concept is still in its earliest stages. Children are generally eager to learn such that motivation is not a limiting factor (Entwisle and Alexander, 1989). However, as children progress through the grades, their level of motivation changes in accordance with their opportunities to learn and an increasing awareness of the socioeconomic contexts in which their learning is taking place (Henderson and Dweck, 1990; Zimmerman, 1990; Schunk et al., 2014). As argued by Mickelson (1990), "the student's perceptions of what her or his efforts and accomplishments in school ultimately will bring from the larger society" is a critical predictor of success in school in later childhood (Mickelson, 1990: 59).

2.2.4 Influences on children's vocabulary development

The importance of vocabulary knowledge for reading achievement and later academic success is well-established in the literature (Thurstone, 1946; Davis, 1968; Anderson and Freebody, 1981; Beck et al., 1982; Stahl and Fairbanks, 1986; Stanovich, 1986). However, what shapes vocabulary learning has traditionally been less well understood, due to the nature of vocabulary processes. Language skills comprise both receptive vocabulary, which is the ability to understand words that are spoken, signed, or written by others, and expressive vocabulary, which is the ability to produce words. Being able to progress from understanding a word receptively to expressing it productively evolves incrementally over time (Nagy and Herman, 2005). Related to this is the distinction between breadth and depth of vocabulary knowledge, where breadth of vocabulary refers to the number of words known and depth refers to the richness of knowledge about the words known (Anderson and Freebody, 1981; Stahl and Nagy, 2007). Whereas "breadth is thought of as the number of words that a student knows (…), depth is thought of as a continuum spanning from some familiarity with a word to full ownership of a word's various meanings and the associated ability to use it appropriately in varying contexts" (Kieffer and Lesaux, 2012: 349, also see Tannenbaum et al., 2006). There is broad evidence that vocabulary is sensitive to targeted instruction throughout the course of childhood (Stahl and Fairbanks, 1986; Beck et al., 1982; Stahl and Nagy, 2007; Beck and McKeown, 2007; Kieffer and Stahl, 2015). Interventions that provide opportunities for independent reading are emphasised from an early age (Kieffer and Stahl, 2015), based on Stanovich's (1986) Matthew effect hypothesis which posits that individual differences in reading comprehension lead to differences in vocabulary development which perpetuates further reading gaps as well as gaps in other cognitive areas (Stanovich, 1986). Vocabulary is also learned within the contexts in which children develop, through communication with parents and peers, classroom lectures and discussion, reading, and television (Nagy and Herman, 1987). Learning words from oral contexts is particularly important in early childhood, before a child has begun to read (Nagy and Herman, 1987), i.e. through conversation or from listening to books read aloud (Leung, 1992; McKeown, 1985; Werner & Kaplan, 1950).

There has been growing criticism that the measures used to assess vocabulary do not capture the complexities of the construct. Moreover, the most frequently used assessments tend to measure receptive vocabulary rather than productive vocabulary (Kieffer and Stahl, 2015). Partly due to the fact that vocabulary as a construct is much more difficult to assess than other foundational reading-related skills such as knowledge of letters, phonemes, and spelling rules (Snow and Kim, 2007). Nagy and Scott (2005) identify five aspects of vocabulary, illustrating the challenge of constructing operational vocabulary measures and providing a basis for research that seeks to understand the complex nature of vocabulary development:

> (1) Incrementality: Knowledge of a given word varies along a continuum from no knowledge through partial knowledge to full, productive command of the word's meanings, connotations, and uses. (2) Multidimensionality: Knowledge about a given word includes several different types of knowledge (e.g. knowledge of the word's semantic and morphological relationships to other

> words, appropriateness to specific contexts, connotations, and metaphoric uses). (3) Polysemy: Many words have multiple meanings and shades of meaning that vary by context. (4) Interrelatedness: Word meanings are not independent from each other but intricately related and interdependent. (5) Heterogeneity: Different types of words are known in different ways and vary in their utility.
>
> (Nagy and Scott, 2005 cited in Kieffer and Stahl, 2015, p. 121)

Nagy and Scott's (2005) five aspects paved the way for empirical research in this area and new directions for the development of vocabulary assessments (Nagy, 2007; Scott et al., 2008; Stahl and Bravo, 2010). One such study is Kieffer and Lesaux (2012), which investigates the multidimensional aspects of vocabulary knowledge among minority English language learners (L2) in the U.S in comparison to students who speak English as a first language (L1). The sample consists of 583 sixth-grade students in an urban district of California (413 L2 learners and 170 L1 learners). The authors use reading-based measures as opposed to oral or picture-based measures partly due to the challenges of assessing multiple vocabulary dimensions simultaneously, but also in view of the fact that "word meanings that are important to reading comprehension in upper elementary and middle school grades are abstract representations of processes and relationships… that appear more frequently in written text than oral conversation and cannot be easily represented with pictures" (Kieffer and Lesaux, 2010: 354). Thirteen measures of vocabulary were administered across two time points based on three sub-tests of the Stanford Achievement Test, 10th Edition (SAT-10) Intermediate Level II as well as four researcher-developed tools. Using multiple-group confirmatory factor analysis techniques, three dimensions of vocabulary knowledge are identified: breadth, contextual sensitivity, and morphological awareness, illustrating the complexity of vocabulary as a construct. The authors conclude that although understanding dimensionality is not sufficient for addressing difficulties in language development among learners, it is a necessary starting point (Kieffer and Lesaux, 2012).

2.2.5 Summary

This section discussed theoretical advancements across academic disciplines acknowledging contextual and cultural understandings of child development. These theoretical transformations have important implications for the current study, namely that the relationship between children's development and their wider contexts is a complex process. Recognising the foundational role played by context and culture within a study of child development that takes a quantitative analytical approach is not a matter of accounting, i.e. specifying cultural factors within a conceptual framework or controlling for related variables in statistical models (e.g. ethnicity, language, religion, etc…). Rather, recognition of its importance is exhibited through culturally meaningful and contextually informed research processes.

For the present study, three separate cohorts within childhood are identified following Graber and Brooks-Gunn (1996). The first is the early childhood period spanning birth to six years. This period of life is considered crucial for the development of the basic functionings, sensori-motor, cognitive-language and socio-emotional, that shape future learning (Masten and Powell, 2003). This is followed by late childhood, which corresponds to seven to 10 years of age, and adolescence, age 11-19 years. This conceptualisation seems appropriate for addressing the current research questions given the official age of entry into formal primary school in Ethiopia is seven years. The transition to junior secondary school is expected during early adolescence.

The stability of cognitive abilities is explored in the study. The characterisation of cognitive processes depends on the specific cognitive construct, the context, and the stage of childhood. The contextual factors that are important for vocabulary development are also explored. Individual differences in change over time are analysed to pinpoint the most appropriate statistical model for describing how poverty shapes vocabulary development, and advances the premise that to improve children's life chances, the factors that promote better outcomes for children should be explored throughout childhood. Consistent with the literature, the researcher recognises the critical early years period, and the role played by early protective interventions in fostering longer term development and learning. The researcher further recognises the importance of maintaining a nurturing and stimulating environment over the course of childhood for early cognitive gains to promote achievement later on. For example, the term "Head Start Fade" was used to reflect the circumstance in which many children who initially benefited from the Head Start school readiness programme, did not experience any longer term cognitive or academic gains once they left pre-school and entered primary school due to a lack of continuity in the programme (Puma et al., 2010; Barnett and Hustedt, 2005; Lee and Loeb, 1995; Currie and Thomas, 1999; Barnett, 1995).

The next section of the review examines evidence of the effects of poverty on children's cognitive development. It documents the effects of poverty in early childhood on development and establishes whether changes in children's socioeconomic conditions further help to explain developmental paths. The pathways through which poverty shapes trajectories are then examined to identify known contextual influences that vary with living standards and shape cognitive outcomes. As a secondary aim, this section of the review considers the quality of the measures used in the reviewed studies. The final section examines different concepts of poverty within academic disciplines to identify an appropriate construct and operational measure of poverty for the research.

2.3 Researching the effects of poverty on children's cognitive development

With a theoretical framework for researching context in relation to children's cognitive development now established, the next stage of the literature review is to explore scholarly work on how living standards affect children's progress, with a focus on research in low-income countries. Sections 2.3.1 and 2.3.2 look at empirical evidence of the effects of early childhood poverty on development, and of poverty experienced at different stages of childhood. Research on the mechanism underlying the poverty-growth relationship is examined in 2.3.3. Section 2.3.4 considers different concepts of living standards to identify an appropriate construct and operational measure for the present study.

2.3.1 Early childhood poverty and cognitive development in low-income countries

> The child's early experiences… create the base for all subsequent learning. Strong early childhood foundations – including good health, nutrition and a nurturing environment – can help a smooth transition to primary school, a better chance of completing basic education, and a route out of poverty and disadvantage.
>
> UNESCO, 2007: 11

The first stage of the review documents the empirical evidence on the effects of early poverty exposure on children's cognitive development, with a focus on the findings as well as the measures used to assess poverty and cognitive outcomes. The review is modelled on traditional perspectives within developmental psychology which emphasise the primacy of early childhood (Gewirtz, 1969; Williams, 1972; White et al., 1979; Jencks et al., 1972), and that individual differences in development are set early on in life (White et al., 1979; Gewirtz, 1969; Kindermann and Skinner, 1992). This literature is reinforced by evidence from the field of developmental science that biological factors and early experiences set the stage for future development and learning throughout the course of childhood (Nelson, 2000; Engle et al., 2007; Grantham-McGregor et al., 2007; Walker et al., 2007). There is vast empirical evidence from Western countries to support this reasoning (Pianta and McCoy, 1997; Duncan and Brooks-Gunn, 1997; Duncan et al., 1998; Currie and Thomas, 1999; La Paro and Pianta, 2000; Feinstein, 2003; Duncan et al., 2007; Farkas and Beron, 2004). However, due to the absence of quality longitudinal data on children's development in low-income countries, evidence of such associations in the developing world is scarce.

Among the few longitudinal studies in developing countries that follow the same sample of children over time, strong associations have been found between early cognitive outcomes and socioeconomic factors such as wealth (Table 2.1). One of the earliest studies, Sigman et al. (1991), used data from a small sample of 83 children in a Kenyan village. The findings showed that verbal ability and reasoning at age 5 years were best predicted by nutritional factors during toddlerhood, socioeconomic status between 18-30 months, and prior ability (Sigman et al., 1991). More recent studies have sought to examine individual differences in cognitive growth by socioeconomic factors using large scale nationally representative survey data. In Brazil, data from the Pelotas birth cohort study 1982-2001 was used to show that children in the lowest socioeconomic class where at a higher risk of developmental delay (Victora et al., 2003). Wealth in infancy predicted the achievement gap at both 12 months and 2 years in nearly 4,000 children (Moura et al., 2010a; Moura et al., 2010b). In Guatemala, parental wealth at birth was strongly associated with grade attainment in over 1,400 adults, which in turn predicted adult reading comprehension, vocabulary, and reasoning (Stein et al., 2005). Grantham-McGregor (2007) found that wealth quintiles at birth were related to intelligence at 8 years in the Philippines, and to cognitive scores at seven years in South Africa and nine years in Indonesia (Grantham-McGregor, 2007). The difference in the z-scores between the richest and poorest wealth quintiles was substantial, ranging from 0.70 to 1.06 SD across the three samples of children (Grantham-McGregor, 2007). In Ecuador, household wealth at three years predicted receptive vocabulary development in over 3,000 children between 3-6 years, with children from the wealthiest or most educated families achieving higher vocabulary tests scores than children from the poorest or least educated families (Paxson and Schady, 2007). Furthermore, poorer children were shown to fall increasingly behind (Paxson and Schady, 2007). Similar trends were found in Nicaragua and Madagascar, where initial differences in ability in the highest and lowest wealth categories worsened with age during early childhood (Macours et al., 2008; Fernald et al., 2011), in Cambodia (Filmer and Naudeau, 2010), Mozambique (Bruns et al., 2010), and across Latin America and the Caribbean (Schady et al., 2015; Rubio-Codina et al., 2015; Rubio-Codina and Grantham-McGregor, 2019). These findings reinforce research within developmental neuroscience that early cognitive gaps persist, and potentially widen over time (Nelson, 2000).

Table 2.1 Evidence of associations between wealth and ability in early childhood

	N	Outcome variable	Assessment instrument	Independent variable	Effect measure	Estimate
Kenya	83	Reasoning and vocabulary	Ravens progressive matrices and four vocabulary items at 5 years	Socioeconomic status composite index at 18 months	Multiple correlation coefficient	0.31 ($p < 0.05$, two-tailed)
Brazil*	3,907	Socio-emotional, communication, motor	Battelle Screening Developmental Inventory at 12 months and 2 years	Wealth quintile at birth grouped into tertiles	Relative risk ratio (lowest to highest tertile)	3.95:1 (1.68 – 9.29)
Guatemala†	1,469	Reading and vocabulary	Inter-American Series reading and vocabulary modules at 26 to 41 months	Wealth tertile in early childhood	Mean difference between top and bottom tertiles	9.9 points ± 1.7 SD for men and 6.2 points ± 1.6 for women
Philippines	2,485	Cognitive ability	Philippines Non-Verbal Intelligence Test at 8 years	Wealth quintile at birth	Difference from the richest quintile in z-scores	0.35 SD to 0.84 SD
South Africa	1,143	Reasoning	Ravens progressive matrices at 7 years	Wealth quintile at birth	As above.	0.34 SD to 0.70 SD
Indonesia	371	Reasoning and arithmetic	12 reasoning items and 5 arithmetic items at 9 years	Wealth quintile at birth	As above.	0.31 SD to 1.06 SD
Ecuador‡	3,153	Vocabulary ability	Peabody Picture Vocabulary Test (Spanish) at 36-53 months and 54-71 months	Wealth index at 36-71 months	Significant regression coefficient	1.56 (0.23) younger cohort; 3.80 (0.31) older cohort
Nicaragua§	1,066	Receptive vocabulary	Peabody Picture Vocabulary Test (Spanish) at 36 and 83 months	Cash transfer treatment (dummy = 1)	Multivariate regression coefficient	0.10 (0.040) at 36 months; 0.06 (0.043) at 83 months
Madagascar‖	1,232	Visual spatial processing, working memory, reasoning, executive function, and receptive vocabulary	Stanford Binet Intelligence Scales; Leiter International Performance Scales; Woodcock-Munoz assessment; PPVT at ages 3 and 6 years	Wealth index	Mean difference between top and bottom	10.8 points ± 0.72 SD for vocabulary; 9.1 ± 0.61 SD for working memory; 8.9 ± 0.59 SD memory of phrases
Mozambique¶	2,000	Receptive vocabulary	Peabody Picture Vocabulary Test	Wealth index	Multivariate regression coefficient	1.29 points (SE=0.51), p<0.10 for

	N	Outcome variable	Assessment instrument	Independent variable	Effect measure	Estimate
			(translated from Spanish) at 36-59 months			the highest quintile
Cambodia	4,072	Receptive vocabulary	Peabody Picture Vocabulary Test (translated from Spanish) at 36-59 months	Wealth index	Multivariate regression coefficient	2.42 (SE=0.34) P<0.50 for the highest quintile
Colombia	1,330	Cognitive processing, socio-emotional, fine and gross motor,language	Bayley Scales of Infant and Toddler Development (Bayley-III) at 6-42 months	Wealth quintile	Mean difference between top and bottom	0.5 SD

*95% confidence interval reported in parentheses; †significance level p <0.01; ‡standard errors reported in parentheses; §standard errors reported in parentheses and significance level p <0.05; ‖significance level p <0.01; ¶mean scores from the sample are also compared with mean scores obtained from the reference group used to norm the test.

The studies demonstrate the importance of monetary-based measures of living standards for contextualising children's development. Nearly all the studies use constructed wealth index-based approaches, highlighting the absence of datasets on actual consumption or expenditure that can be linked to data on childhood outcomes. Each of the studies document the process by which the wealth measures were constructed but do not include statistical information on the reliability of the wealth index data, i.e. a comparison of the distribution of the index scores to those obtained by population-based surveys. In Brazil and Guatemala, researchers used wealth indices derived from household expenditures and household characteristics based on national practice (Stein et al., 2005; Maluccio et al., 2005; de Moura et al., 2010b). National measures were also adapted in the Philippines (Mendez and Adair, 1999) and in South Africa (Yach et al., 1991). In Kenya, the authors developed a composite measure of living standards, consisting of indicators of access to "land, farming and household assets, use of social services and community involvement" (Sigman et al., 1991: 310). In Ecuador, Madagascar, Cambodia, and Mozambique, constructed wealth indices drew upon the methodology proposed by Filmer and Pritchett (2001), using evidence of asset/commodity ownership rather than actual expenditures.

Regarding the cognitive assessment instruments, the main weakness in the research designs lies in the lack of documented evidence of the reliability and validity of the measures used to assess cognitive ability. Different aspects of statistical reliability are reported. Validity arguments, where provided, are generally limited to a description of test translation procedures. None of the studies share information about how the developmental domains are conceptualised and defined within each of the study countries, or how the constructs are adapted and operationalised to suit the research settings. Acknowledgement of possible cultural sources of bias, or “cultural-loading” in instrument adaptation (Jensen, 1974b), is not explicit. The focus is on establishing evidence of statistical reliability. Moreover, validity issues relating to the assessment process are not documented. This is consistent with much of the developing world assessment-based research (see Nag et al., 2017). Apart from a description of test translation procedures and piloting processes, arguments are rarely provided to justify the quality of the measures.

However, documenting test validity is particularly important when instruments are applied in a multicultural setting or have been developed and validated for use in some other cultural context. As argued by Yach et al. (1991):

> Comparisons of cognitive and social development across cultures and social classes are problematic because standard psychological tests are biased against children who fall outside the framework…
>
> (Yach et al., 1991: 218)

Whereas the cognitive tests administered to the Indonesian and the Philippines samples were developed nationally (Cheung, 2006; Mendez and Adair, 1998), international tests were adapted for use in all other study countries. In Brazil, a Spanish version of the Battelle Screening Developmental Inventory was translated into Portuguese (de Moura et al., 2010b). In Guatemala, researchers applied the Inter-American Series (SIA) to assess reading comprehension and vocabulary, originally developed for Spanish-speaking American children (Stein et al., 2005). The authors state the “SIA demonstrated adequate test-retest reliability… internal consistency … and validity in previous studies in this population" (Stein et al., 2005). Only four Peabody Picture Vocabulary Test (PPVT) items were used to assess verbal ability in the Kenya study community (Sigman et al., 1989). As the PPVT normally consists of several sets of items covering a range of ability levels, this four-item measure is not a meaningful proxy. Needless to say, the authors report that the test showed high inter-rater reliability based on correlations between scorers and internal consistency reliability across time points (Sigman et al., 1989), as well as evidence of predictive validity, i.e. good correlation between test scores and school examination results (Sigman et al., 1989). In Ecuador and Nicaragua, children were given the Spanish version of the PPVT (Paxson and Schady, 2007; Filmer and Naudeau, 2010), which has been internationally validated and is now widely used in Latin America. In Madagascar, the English version of the PPVT was translated into Malagasy. While the test showed high inter-rater reliability and good internal consistency (Fernald et al., 2011), no qualitative or statistical evidence of construct validity is provided. In Cambodia and Mozambique, the Spanish version of the PPVT was translated into local languages as it was considered more appropriate for children in low-income settings than the English version (Naudeau et al., 2011).

A wide range of developmental assessment tools have been adapted from Western contexts and applied to children in Low and Middle-Income Countries (LMICs); however, international instruments are often applied less critically in low-income countries. Table 2.2 provides a summary of some of the main tests as well as pending questions regarding their reliability and validity when applied in developing countries. The main finding from this review is that while the instruments have shown good psychometric properties in high income countries, these tests are not always subject to the same level of conceptual reflection, adaptation, and robustness checks in developing countries. Comprehensive diagnostic tests, such as the Bayley Scales of Infant Development (BSID) and the Woodcock-Johnson/ Munoz assessments are expensive, technical skill intensive and require substantial investment in training for successful administration (Fernald et al., 2009). Administrators normally require expertise in early childhood education and experience in assessing/working with children. However, in developing country contexts such resources and competencies are often lacking, raising questions about the quality of the data collected using these tools.

Developmental screening tests, such as the Ages and Stages Questionnaire, are often used as alternatives to comprehensive ability tests due to the challenges of adapting and administering comprehensive assessments. The resulting data are then used to characterise populations of children. While developmental screeners are inexpensive, take less time to administer and require much less training time, they are not diagnostic tests. As noted by Fernald et al. (2009) “because screening tests only contain a sample of items per domain (i.e., they do not assess the full range of ability) they do not yield continuous scores, but are used to classify children into categories, such as “Delayed,” “At Risk for Delay” or “Within Normal Limits” for age” (Fernald et al., 2009: 32). Thus, applying a developmental screener as a diagnostic tool may not be appropriate for addressing certain research questions. Moreover, the use of a screener would first require the establishment of meaningful categories for the population in that setting. In countries with no population-based cut-offs, it has been recommended that the use of screening tests should be limited to examining how one group of children performs on the test relative to another group, and then using the tests as a “guage” of developmental outcomes (Fernald et al., 2009: 32).

Table 2.2 Summary of the main cognitive instruments applied in LMICs

Instrument (Developer)	Age range	Developmental domain(s)	Characteristics of the test	Validity and reliability issues
Ages and Stages Questionnaire - ASQ (University of Oregon-Brookes Publishing)	1 to 5 years	Problem-solving, language, socio-emotional, fine and gross motor	Developmental screener, not a full diagnostic test; cheap; short enough for use as a population-based measure; less technical skill required; widely used in LMICs; requires home administration	Evidence of low reliability and low concurrent validity with the BSID, which is considered the gold standard; unknown predictive validity
Bayley Scales of Infant Development - BSID (Nancy Bayley)	1 mo.to 4 years	Cognitive, language, socio-emotional, fine and gross motor	Comprehensive diagnostic test; considered a 'gold-standard'; expensive; long to administer; technical skill intensive; validated in several languages	Focus on infant/toddler development implies potential ceiling effects among older children; limited use in LMICs due to the high level of technical skill required.
Battelle Developmental Inventory screener – BDI (Jean Newborg)	Birth to 7 years	Cognitive, language, fine and gross motor, socio-emotional	Developmental screener; expensive; long to administer; technical skill intensive; requires home administration	Uncertain predictive validity; shows increasing concurrent validity in age with BSID
Denver Developmental Screening Test - DDST (University of Colorado)	2 months to 6 years	Language, fine and gross motor, socio-emotional	Developmental screener; cost effective; short enough for use as a population-based measure; low technical skill requirement; home administration;	Uncertain predictive validity; evidence of increasing concurrent validity in age with BSID; does not measure cognitive areas
Early Development Index – EDI (McMaster University, Canada)	4 to 7 years	Cognitive, language and communication, socio-emotional, health	School readiness focus; cost effective; short enough for population-based monitoring; easy to administer	No direct assessment; not widely validated in low-income countries (Mozambique, Malawi)
International Development and Early Learning Assessment – IDELA (Save the Children)	3 to 6 years	Literacy and language, emergent numeracy, fine and gross-motor, socio-emotional	ECD focus; operationally free; short to administer; self-training required; caregiver module measures the quality of the home environment	Designed for use in rural, impoverished communities and thus requires substantial adaptation and piloting for use in urban, less poor settings
Measuring Early Learning Quality and Outcomes – MELQO (UNICEF, World Bank, Brookings)	3 to 7 years	Literacy and language, mathematics, executive function, socio-emotional	School readiness focus; operationally free; low technical skill required; short to administer; caregiver module measures the quality of the home environment	Increasingly field tested in LMICs; uncertain feasibility for population-level monitoring and monitoring programme reform; focused on inter-country comparisons
Peabody Picture Vocabulary Test – PPVT (Dunn and Dunn)	2-90+ years	Receptive vocabulary	Comprehensive vocabulary assessment; expensive; technical skill intensive; long to administer; often used as a proxy for cognitive ability and screening instrument for school readiness	Uncertain construct validity when used with young children and in multicultural or multilingual settings
Regional Project on Child Development Indicators -PRIDI (Inter-American	2 to 5 years	Language, cognitive, socio-emotional, motor	ECD focus; operationally free; low technical skill required; short to administer; includes the quality of the home environment and access to	Contextualised in 4 Latin American countries only; focused on inter-country comparisons

Instrument (Developer)	Age range	Developmental domain(s)	Characteristics of the test	Validity and reliability issues
Development Bank)			ECE; requires home administration	
Woodcock-Johnston/ Munoz Early Cognitive and Academic Achievement (HMH)	2 to 6 years	Cognitive, language and communication, numeracy	Comprehensive diagnostic test; expensive; long to administer; technical skill intensive	Limited application in LMICs due to the level of technical skill required to administer the test and analyse the test data

Authors own analysis based on a review of the international instruments and secondary sources: Rubio-Codina et al. (2016), Frongillo et al. (2014), and Nag et al. (2017).

This section explored empirical research on the effects of early childhood poverty on children's cognitive development in low-income countries, with a focus on the measures used to assess cognitive outcomes. A large proportion of studies do not provide evidence of sound psychometric properties. The studies that do mention concepts of reliability and validity, emphasise statistical evidence of the overall test reliability. Fit-for-purpose, contextualisation and cultural sensitivity are not addressed as main priorities (see Nag et al., 2017; Rubio-Codina, et al., 2016; Frongillo et al., 2014). This finding is reinforced by results from a recent systematic review of early literacy and language assessments in developing countries which showed that the reliability and validity of learning assessments is poorly documented, and that the psychometric properties of tests are rarely analysed (Nag et al., 2017). However, language assessments with demonstrated good psychometric properties are essential for developing effective interventions (Dockrell and Marshall, 2014; Denman et al., 2017).

2.3.2 The time effects of poverty on children's cognitive development

Poverty effects on children's cognitive development could also manifest at different stages of childhood. Such longer-term relationships cannot be investigated using cross-sectional data. Whereas the previous section looked at evidence of how poverty slows cognitive growth in early childhood, this section considers longitudinal research, with attention paid to both the findings and the measures. The rationale stems from the theoretical reasoning within developmental psychology that although a nurturing early environment is vital it “may not be sufficient to support a positive trajectory” (Skinner et al., 1998: 31). Investment and support may be discontinuous over time, with implications for children’s development and learning. For example, it is well established that poverty experiences may not remain constant over the life course (Bane and Ellwood, 1986; Duncan and Rodgers, 1988; Duncan et al., 1994; Haveman and Wolfe, 1995). Poverty may be persistent – experienced over a long duration, short term or periodic. Moreover, children often experience poverty in surges, in terms of movements in and out of poverty (Duncan and Rodgers, 1988; Jenkins, 2011), and in terms of the strength of the relationship between poverty and child outcomes (Teachman et al., 1997; Guo, 1998).

Several longitudinal studies have documented the negative effects of poverty on children’s learning and schooling in high income countries (reviewed in Chen, et al.,2010). There is a large body of work documenting the positive association between family income and achievement (Hill and Duncan, 1987; Corcoran et al., 1992; Haveman et al., 1991; Brooks-Gunn and Duncan, 1997; Duncan and Brooks-Gunn, 1997; Haveman and Wolfe, 1995). Family income in the first five years of life strongly predicts later academic achievement (Duncan et al., 1994; Miller and Korenman, 1994; Smith et al., 1997). However, in adolescence, family income has been shown to have a weaker effect (Haveman and Wolfe, 1995; Guo, 1998), due to the growing influence of motivation and other factors at this later stage of childhood (Henderson and Dweck, 1990; Zimmerman, 1990; Schunk et al., 2014). Using data from the Panel Study of Income Dynamics, Duncan, Yeung, Brooks-Gunn and Smith (1998) examined the effects of the timing of family income on education-related outcomes. They use four measures of poverty based on family income at different stages of childhood and three measures of childhood outcomes in adolescence: years of schooling, school completion and premarital childbearing. Their findings indicate that the income effect on school completion is more pronounced than its effect on the other outcomes, and that family income in early childhood has a greater impact on adolescent outcomes than income in middle childhood.

The main limitations of this study lie in the decision to exclude direct measures of learning achievement and to measure child outcomes only in later childhood. As such, the study cannot disentangle early poverty effects from prior educational effects.

Poverty in the early years predicts deficits in test scores across a range of cognitive outcomes, with persistent or long duration poverty having a more detrimental effect than short term poverty (Dickerson and Popli, 2012; Brooks-Gunn and Duncan, 1997; Duncan et al., 1994; Korenman et al., 1995; Smith et al., 1997). Using data from the National Longitudinal Survey of Youth, Guo (1998) analysed the effects of poverty on the cognitive ability and achievement of children in early childhood and early adolescence. Three constructed poverty measures were used based on family income: the number of years living below the poverty line, the average income-to-needs ratio, and average annual family income, all of which were measured prior to the assessments. Two tests were used to measure cognitive ability: the Peabody Picture Vocabulary Test-Revised (PPVT-R), which assesses verbal ability or scholastic aptitude and the Memory for Digit Span Assessment. Three subsets of the Peabody Individual Achievement Test (PIAT) were used to measure academic achievement: reading recognition, reading comprehension and mathematics. For the PIAT, childhood poverty was found to be less important than poverty experienced in adolescence, The early childhood poverty effect was stronger for the PPVT-R and the Memory for Digit Span Assessment. Guo (1998) argues that the findings are consistent with the premise that ability is a more stable outcome than achievement and that factors such as motivation and opportunity, which “tend to be affected by an impoverished environment from early adolescence on” (Guo, 1998: 265), become more important predictors of outcomes in later childhood (Henderson and Dweck, 1990; Breen and Goldthorpe, 1997; Guo, 1998; Ogbu, 1986; Mickelson, 1989, 1990). The main shortcoming has to do with the measures available in the dataset. The PPVT-R is a measure of receptive vocabulary. The PPVT is a measure of receptive vocabulary. The use of vocabulary assessments as indicators of generalised intelligence (e.g. the

NICHD Early Childhood Care Research Network, 2005) is based on the Aptitude hypothesis, which suggests,

> …students who score higher on vocabulary tests have learned more word meanings from exposure to oral and written language because they have higher levels of generalised verbal aptitude. This aptitude, in turn, contributes to their development of reding comprehension skills, independently of knowledge of specific word meanings.
>
> (Keiffer and Stahl, 2015, p. 126)

However, Nagy (2007) hypothesises that individual differences in vocabulary are primarily influenced by differences in metalinguistic ability such as phonemic and morphological awareness, and sensitivity to context (Nagy, 2007), abilities that are not stable but learned over time.

Hart and Risley (1995) provide one of the first pieces of evidence of how children's growth trajectories are shaped by changes in family living standards over time, as shown in Figure 2.4. They use small sample data collected from 42 American children who were followed from one to three years of age: 13 children from high SES backgrounds, 23 middle/lower SES children and 6 children whose families receive welfare assistance. Children and families were assessed monthly. The authors observed the oral language interactions between the children and their caregivers, i.e. the number of words heard by the child, and recorded children's verbal language demonstrations, i.e. the number of different words spoken by the child. They plotted the trajectories of oral language acquisition for children living in three subgroups of families: largely Caucasian-American professionals, working-class families, and largely African American welfare recipients. One African American family was upper SES, three were middle, and seven were lower. The authors found notable differences between the sub-groups in terms of the number of words spoken by the child. Differences in vocabulary size emerge immediately as the children begin to speak, with high SES children outperforming the children in the other two groups. From around 18 months onward, the rate of vocabulary growth for high SES children exceeded that working-class and welfare children. Children from working-class families begin to develop their vocabularies more rapidly than welfare children from about 22 months of age. The vocabulary gap between social classes continued to widen across the first three years of life. The researchers went on to analyse variation in the quality of children's conversations with their parents. High SES parents were found to have more conversations with their children and to use more wide-ranging vocabulary in their conversations, in comparison with working class and welfare families (Hart and Risley, 2003). Of the 42 children followed in the research, a sub-sample of 29 were followed-up at age 9-10 with assessments of language and literacy: the Peabody Picture Vocabulary Test (PPVT-R), the Test of Language Development-2 Intermediate (TOLD) and the Comprehensive Test of Basic Skills (CTBS/U). Oral vocabulary at age 3 years predicted receptive vocabulary, language and reading comprehension at nine years (Hart and Risley, 2003), showing that initial gaps widen in early childhood and lead to future gaps in other areas of development and learning. The Hart and Risley "word gap" study was immediately influential among policy makers despite substantial methodological criticisms among academic researchers on grounds of biased sampling, poor data coding and generalisability issues (Sperry et al., 2019). Moreover, a more recent longitudinal study has failed to replicate the results, showing no correlation between socioeconomic status and the number of words a child hears and revealing substantial variation in vocabulary environments within each socioeconomic stratum (Sperry et al., 2019).

Figure 2.4 Differences in children's vocabulary trajectories across income groups

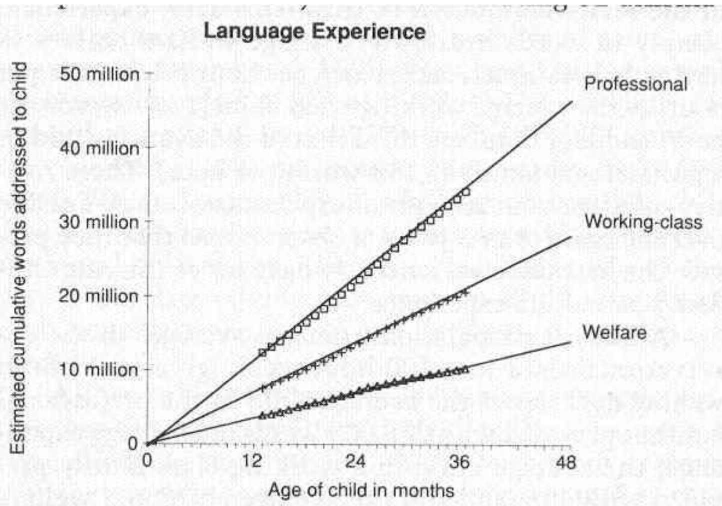
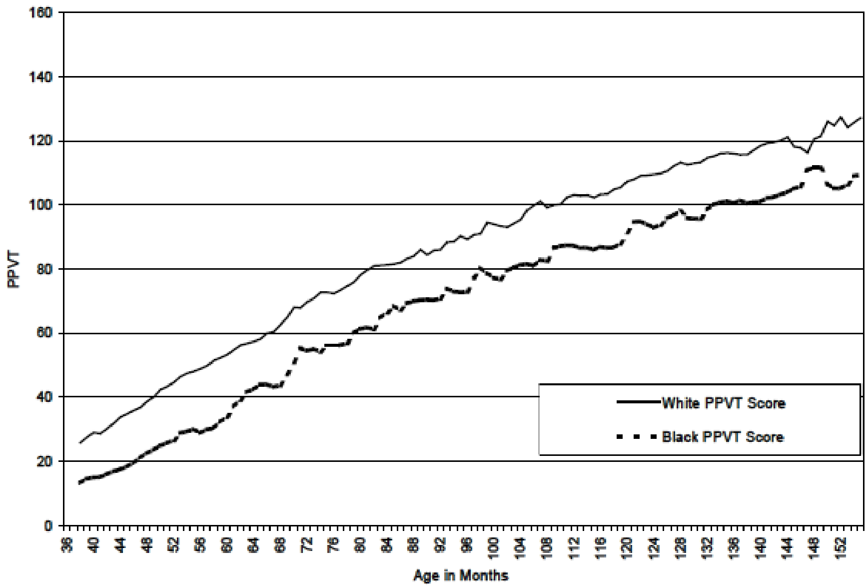


Figure 19. Estimated cumulative differences in language experience by 4 years of age. (See Appendix B for a detailed explanation of this figure.)

Source: Hart and Risley, 1995: 234.

This early pioneering work was followed by Farkas and Beron (2004), using data from the Children of the National Longitudinal Survey of Youth (CNLSY) to estimate the oral vocabulary trajectories of American children between the ages of three and 13 years. The study examines how race and class differences in learning evolve over childhood, as shown in Figure 2.5. The national survey tracks children's receptive vocabulary development at irregular intervals over the period, based on the Peabody Picture Vocabulary Test (PPVT). The authors do not consider vocabulary knowledge a stable characteristic in contrast with Guo (1998). They regard it "as every bit as acquired an ability as reading and mathematics" (Farkas and Beron, 2004: 467). A distinction between ability and achievement is not made, rather, the premise is that there is "a developmental sequence of skills and knowledge whose successful performance must be achieved in an ongoing learning effort that extends from birth throughout the life cycle" (Farkas and Beron, 2004: 467). Data obtained from the repeated vocabulary test are stacked for each child. They use multilevel growth modelling techniques to estimate individual learning trajectories for all children in the sample, controlling for a range of contextual factors. The authors found that by three years of age, racial and social class disparities were already evident. The gap continues to widen until age five, after which time it stabilises and remains constant throughout middle childhood. Social class differences stabilise at around three years of age for Caucasian-American children and at five years for African American children. Family linguistic instruction in early childhood explains some of the group difference, based on measures of mothers' vocabulary and mothers' cognitive and emotional support to their children. The finding that individual differences in vocabulary development are established during the preschool years and then stabilise supports previous research. Receptive vocabulary, which is a building block for literacy, develops rapidly in the early years. Related research has shown that progress in other areas of learning such as mathematics, which build on prior knowledge and skills may also demonstrate a widening gap well into middle and late childhood (DiPrete and Eirich, 2006). The main limitation of the research is the focus on a single cognitive domain despite the breadth of test score data available in the survey. Moreover, hypotheses about what drives racial and social class differences are not explicit. It has been established that neighbourhood, peer and social factors such as motivation become increasingly influential in later childhood (Ogbu, 1986; Mickelson, 1989, 1990; Duncan and Brooks-Gunn, 1997; Breen and Goldthorpe, 1997). Such contextualised analyses would help to identify opportunities for intervention at different stages of childhood.

Figure 2.5 PPVT scores from 3-13 years of age, by race



Source: Farkas and Beron, 2004: 477.

Fewer longitudinal studies have examined individual differences in child development in low-income countries. This is largely due to the challenges of collecting repeated measures data in developing countries, particularly the high rates of attrition and the lack of funding for participant follow-up (Campbell and Rudan, 2011). Recent work in the field of international development economics has sought to improve the evidence base in this area. There is a small body of research showing that cognitive delays in very young children tend to increase during early childhood (Fernald et al., 2011; Filmer and Naudeau, 2010; Bruns et al., 2010; Macours et al., 2008; Macours and Vakis, 2010; Schady and Paxson, 2007). The results build on earlier pioneering longitudinal research in low- and middle-income countries (Richter et al., 1998; Walker et al., 1990; Lozoff and Teal, 2004), and are consistent with theories of brain development (Nelson, 2000) and the sociological hypothesis of cumulative advantage, also referred to as “the Matthew effect”, (Jensen, 1966 and 1974a; Merton, 1968; DiPrete and Eirich, 2006), which posits that because later learning depends on earlier learning, “learning deficits are cumulative” (Guo, 1998: 283). The theory has been applied to understanding individual differences in educational outcomes, most notably reading (Stanovich, 1986) and science achievement (Walberg and Tsai, 1983). This strand of literature has further shown that wealth disparities in cognitive outcomes emerge early on in life, and are persistent throughout early childhood (Hamadani et al., 2014; Schady and Paxson, 2007; Fernald et al., 2012; Filmer and Naudeau, 2010; Bruns et al., 2010). The findings echo research from developmental psychology and developmental neuroscience documenting huge disparities between young children in high and low wealth quintiles in terms of language development and executive function (Bradley and Corwyn, 2002; Farah et al., 2006; Hackman and Farah, 2009; Noble et al., 2007; Siddiqi et al., 2007).

The evidence from low-income countries on whether the wealth gap in early childhood remains constant, widens, or decreases through to middle childhood and adolescence is mixed. Fernald et al. (2011) analyse the effects of wealth on children's cognitive development between ages 3 and 6 years in Madagascar. Five instruments are used: the Stanford Binet Intelligence Scales for Early Childhood which assesses visual-spacing processing, working memory and reasoning; the Leiter International Performance Scales which assess attention; the Woodcock-Munoz assessment for memory; a locally developed test of the Stroop effect which measures inhibitory control, an important construct for self-regulation; the Peabody Picture Vocabulary Test, 3rd edition (PPVT-III) which assess receptive vocabulary and is used as a proxy for scholastic achievement. Based on a constructed measure of initial wealth status measured when the children were three years old, the findings showed that 3-year olds in the top wealth quintile performed significantly better across nearly all developmental domains compared with the lowest wealth quintile. By 6 years of age, the wealth gap in executive function, language and cognition doubled, eventually levelling off at age nine (Galasso et al., 2017). The composite cognition score is derived from assessments of visual spatial processing, working memory and reasoning. Similar deficits are shown in the Young Lives study countries, where wealth disparities at age five where found to persist through to eight and

12 years (Lopez-Boo, 2016; Reynolds et al., 2017). The size of the wealth gap decreased in Vietnam and Peru, increased in Ethiopia, and remained constant in India (Reynolds et al., 2017). The results from a Latin American multi-country study were also inconsistent. In Peru, the size of the wealth gap in vocabulary ability stabilised at four years of age and remained unchanged when the children were retested at 8-9 years (Schady et al., 2015). In rural Ecuador, the gap was constant up to 12-13 years (Schady et al., 2015). In Nicaragua, the wealth gap in vocabulary ability widened between age four and seven years (Schady et al., 2015). The reason for the discrepancies across and within studies may be due to assessment design and data collection strategies, specifically: “the limited measurements of child development; or the use of maternal reports in early childhood” (Rubio-Codina and Grantham-McGregor, 2017: 2). In each of the studies, measures of developmental and learning contexts are particularly limited. A second limitation in the Latin American and Madagascar studies is that wealth is measured only once in early childhood, thus the effect of changes in children’s socioeconomic contexts cannot be investigated.

2.3.3 Mechanisms through which poverty affects cognitive development

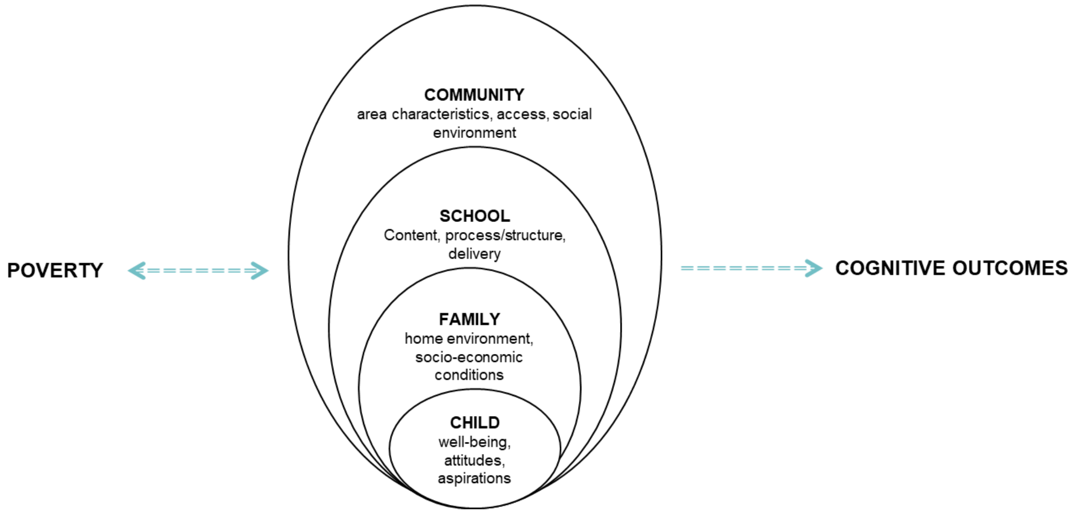
This third section examines three main pathways through which poverty influences children’s cognitive development: children’s health and nutrition, the quality of the home learning environment, and access to preschool. It considers research that seeks to explain individual differences in the poverty effect, i.e. what drives socioeconomic disparities in cognitive outcomes over time? However, the mechanisms mediating the effects of *wealth* on cognitive development is not widely studied in human developmental research. In this field, wealth is recognised as one of many monetary-based measures of living standards, such as: family income, the number of years living in poverty and the income-to-needs ratio. Wealth status is commonly used as a proxy measure of household living standards in human developmental research in low-income countries due, in part, to the challenges of collecting more direct living standards information (Filmer and Pritchett, 2001). The argument that basic needs are more significant in low-income countries also partly justifies the prominence of monetary-based measures (Sen, 1983). In high-income countries, attention is more often focused on multidimensional constructs of living standards such as *socioeconomic status (SES)*, which extends beyond monetary aspects to include noneconomic indicators of social status. Typically, operational measures of SES include indicators of family income or wealth, educational qualifications and occupational status, or some composite measure encompassing various aspects of a family's social and economic position (Dutton and Levine, 1989).

It is well-established that childhood socioeconomic status (SES) is associated with academic achievement, and there is a voluminous literature showing that the achievement gap tends to widen with age through a range of mechanisms. Feinstein's (2004) seminal study linked UK census data with educational administrative data between 1958 and 1970 to examine mobility in pupil achievement by socioeconomic class (Feinstein, 2004). The results demonstrate the stability of the social class effect, with 65 per cent of the children in the bottom socioeconomic quartile who had low scores at age 7 still showing poor performance at age 11, compared with 35 per cent of children in the highest socioeconomic quartile (Feinstein, 2004). For the poorest children, the pattern of change in achievement over time is associated with increasing exposure to negative individual, family and neighbourhood conditions that systematically vary with SES. Feinstein (2004) shows that the probability of falling from among the top scoring quartile to the bottom scoring quartile was only 20 per cent for the highest socioeconomic class and 74 per cent for children in the lowest class (Feinstein, 2004). In a previous study, Feinstein (2003) shows that children of educated or wealthy parents who scored poorly on cognitive tests at 22 months were able to catch up, whereas children of disadvantaged backgrounds who scored poorly were unable to do so (Feinstein, 2003). Factors that predict the widening achievement gap include disparities in the quality of home interactions, particularly during the summer months (Alexander et al., 2007), institutional arrangements which place high SES students in more advantaged learning environments (Maaz et al., 2008); and evidence of cumulative advantage in which earlier success produces future advantage (DiPrete and Elrich, 2006).

SES is an important construct for cognitive ability as well. The mechanisms through which SES predicts cognitive growth differ from the ways in which SES predicts *academic* achievement growth. To reiterate, cognitive ability refers to the primary cognitive functions that begin to form in early childhood, such as language and executive function. It differs from academic/educational achievement which is linked to formal schooling. Research on the cognitive effects of SES indicate that it strongly predicts early cognitive outcomes (Hackman and Farah, 2009; Farah et al., 2006). Particularly noteworthy is that "multiple family, psychosocial and neighbourhood experiences and characteristics that influence development negatively, systematically vary with SES" (Hackman and Farah, 2009: 66). Drawing on Bronfenbrenner (1979), Figure 2.6 provides a visual illustration of poverty, interacting with other environmental influences, to shape children's cognitive outcomes.

The literature on mediated SES effects is dominated by cross-sectional studies or short-term longitudinal surveys covering periods of early childhood, with few studies investigating such relationships over a long timeframe. Drawing on the theoretical foundations discussed in section 2.2, the premise here is that the child, his/her environment and the relationship between the two changes over time, and thus attention should be given to both individual change and on contextual change.

Figure 2.6 How poverty interacts with environment to shape children's development[2]



Proximal influences in early childhood are crucial to the formation of cognitive skills (Hart and Risley, 1995; Roberts, Jurgens, and Burchinal, 2005; Caldwell and Bradley, 1984). Proximal processes are the "engines of development" and relate to interactions in the child's immediate environment, such as biological characteristics and family experiences (Mayes and Lewis, 2013: 19). In early childhood, interactions with peers, the school and community are less frequent (Bornstein, 2012), and would therefore be considered as distal influences in the early years. As a child's social life evolves to adolescence, such influences become more proximal (Harris, 1995; Aber et al., 1997; Chase-Lansdale et al., 1997; Klebanov et al., 1997). It is well-established that the home learning environment in early childhood mediates the effects of family socioeconomic risk on cognitive development (Chase-Lansdale et al., 1997; Klebanov et al., 1997; Bradley and Corwyn, 2005; Evangelou et al., 2008; Melhuish et al., 2008; Galindo and Sheldon, 2011; Bornstein and Putnick, 2012). Home factors most widely studied include the availability of

[2] Author's own illustration, drawing on Bronfenbrenner (1979) The Ecology of Human Development.

stimulating materials, conversations and activities with adults, and parental warmth and affection, with children facing greater socioeconomic risk showing higher gains from positive experiences (Kagitcibasi et al., 2001; Jordan et al., 2000; Desforges et al., 2003; Siraj-Blatchford, 2010).

Farkas and Beron (2004) build on the cross-sectional work of Hart and Risley (1995), Duncan et al. (1994) and Bradley and Caldwell (1984), to estimate the longitudinal relationship between the home learning environment and children's cognitive development over a 10-year period. Up to this point, the vast majority of research used cross-sectional data to explore these relationships. They use data obtained from the US Children of the NLSY79 (CNLSY) to estimate vocabulary growth curves. The results of the assessment show an initial gap in vocabulary knowledge at three years between the highest and lowest social classes and between African-American and Caucasian-American children, as well as a widening gap until the age of five years (Farkas and Beron, 2004). The social class effect was also found to grow among African American children until age five (Farkas and Beron, 2004). Three home environment factors: mothers' vocabulary, cognitive stimulation and emotional warmth partially explains the widening developmental gap. Mothers' vocabulary, which is included in the survey to capture maternal literacy, is measured using test scores obtained from the administration of the Armed Forces Qualification Test (AFQT) at the first sweep. To measure cognitive stimulation, a range of questionnaire items were included to capture the availability of learning materials and opportunities to learn at home (e.g. books, support for learning). The emotional warmth scale, derived from the Home Observation for Measurement of the Environment HOME (Bradley and Caldwell, 1984), measures affect in mother-child conversations and the extent to which mothers encourage children's contributions to discussions. The justification for using these three measures together is based on the premise that while the acquisition of children's vocabulary in the early years is highly dependent upon parents' linguistic skills, language skills can only be transferred to children if parents interact with their children in a stimulating and affectionate way.

There is a sizeable literature on the importance of nutrition for early cognitive functioning and development. In a systematic review of 140 published studies on this topic, Walker et al. (2007) identified that the most "proximal" biological and psychosocial risks factors encountered in the first five years of life affecting 20 to 25 percent of children in the developing world were mainly health and nutritional related, specifically: stunting, inadequate cognitive stimulation, iodine deficiency and iron deficiency anaemia (Walker et al., 2007: 145-150). There is further evidence that nutritional deficiencies in the early years have long term implications for children's future ability to learn (Pollitt et al., 1993; Grantham-McGregor, 1995; Gomes-Neto 1997; Grantham-McGregor et al., 1999a; Grantham-McGregor et al., 1999b; Glewwe et al., 2001; Walker et al., 2005; Paxson and Schady, 2007; Sanchez and Dercon, 2011; Outes-Leon and Porter, 2011). Using cross-sectional data at two separate time points, Fernald et al. (2012) analyse the mediating effect of child nutrition on the wealth-cognitive growth relationship using test scores from the Extended Ages and Stages Questionnaire (EASQ), administered to children between age three and 23 months in four low-income countries: India, Indonesia, Peru and Senegal. The EASQ assessment is indirect in that it seeks information from parents and does not assess children's ability directly. The survey also addresses child health and nutrition, using the length-for-age z-score (LAZ), which is a World Health Organisation growth standard for wasting. The Family Care Index, developed for UNICEF Multiple Indicator Cluster Survey (MICS), is used to measure the home learning environment. The index is a composite derived from children's access to stimulating materials and opportunities to engage in stimulating activities with adults. The wealth effect was found to increase with age in India and Indonesia 0.4–0.5 SD between the ages of 0-7 months and 16-23 months. The gap remained constant in Senegal and Peru. The maternal education effect increased in India and Peru, by 0.2–0.7 SD. The results show that 18-37 percent of the wealth effect was mediated by the home learning environment, and child nutrition explained 12-31 percent of the maternal education effect. The limitations of the study lie in the use of cross-sectional data, and the use of a child developmental screener rather than a comprehensive, diagnostic assessment.

High-quality pre-schooling has been shown to promote improved cognitive development among children living in families with low SES. One of the most notable research initiatives is the Effective Pre-school, Primary and Secondary Education (EPPSE) project, which tracked a group of 3,000 pre-school aged British children at irregular intervals from the start of pre-school until 16 years. A nationally representative sample of children with no pre-school experience was used as a comparison group to examine the effectiveness of the pre-school intervention. The EPPSE pre-school quality instruments, which are based on three internationally recognised observation instruments: the Early Childhood Environment Rating Scale-Revised (ECERS-R, Harms et al., 1998), the Early Childhood Environment Rating Scale-Extension (ESERC-E, Sylva et al., 2003), and the Child-Care Interaction Scale (Arnett, 1989), quantify personal care routines, space and furnishings, use of language and reasoning, activities, interactions, programme structure and staff/parent engagement (Sammons et al., 2014). The main finding of the study overall is that throughout the primary school years and up to the secondary level, children with pre-school education consistently perform better in mathematics and English than their counterparts who are not exposed to pre-school (Sylva et al., 2014). Children who experienced high-quality early years education benefited from additional positive behavioural outcomes, namely: higher self-regulation abilities, more positive social behaviour, and less hyperactivity (Sammons et al., 2014). Moreover, the most positive effects on cognitive outcomes were found among children whose parents have low educational qualifications (Sylva et al., 2014). For children experiencing family risk (e.g. poverty) or individual risk (e.g. low birth weight) in early childhood, quality pre-schooling was found to moderate the effects of experienced risk on cognitive outcomes (Hall et al., 2009), in turn reducing the developmental gap at the start of primary school between children who are at-risk and children who are not.

There is a small number of longitudinal studies that seek to explain socioeconomic effects on cognitive development in low-income countries due to the difficulties of obtaining longitudinal data on children's development. Using the Young Lives longitudinal survey data, Lopez-Boo (2016) examined the mediated effect of SES on vocabulary achievement at age 5 and then again at 8 years, in four low-income countries: Ethiopia, India, Peru and Vietnam. Though the research does not apply longitudinal data analysis techniques, the study is important for the current research as it sheds light on the mechanisms through which SES and cognitive achievement are associated. SES is proxied by a measure of household expenditures. Children's cognitive ability is measured by the Peabody Picture Vocabulary Test (PPVT). A range of contextual factors are modelled as mediator variables in the regression models. The choice of variables is largely based on what is available in the survey data structure: urban/rural stratum, mothers' educational level, nutrition status (height-for-age z-score), pre-school attendance and primary school attendance. The results showed that urban residence, caregiver's education, and children's nutrition most strongly account for the SES-ability relationship. Consistent with developmental theory which posits that ability is largely formed in the early years and thus shaped mainly by proximal contextual influences, primary school attendance was not found to be statistically significant. Preschool attendance mediates a small proportion of the SES effect in Ethiopia and Vietnam. In contrast with the EPPSE study, pre-school attendance is hypothesised as a mediator variable in Lopez-Boo's Young Lives research design. EPPSE models pre-schooling as a moderator of the effects of low SES on development and learning. While there is some uncertainty in social science research surrounding the distinction between these terms, moderator variables generally refer to variables that specify the "conditions" under which a pre-established causal relationship is observed (Hall and Sammons, 2013: 268). Mediator variables are hypothesised “mechanisms” through which an underlying relationship is observed (Hall and Sammons, 2013: 268). In EPPSE, the research hypothesis is that among children living in families with poor socioeconomic conditions, access to quality pre-school mitigates the SES effect on children’s developmental trajectories. Lopez-Boo hypothesises pre-schooling as a pathway through which SES shapes children’s development (Lopez-Boo, 2016).

Additional evidence from the Young Lives longitudinal study suggests that stunting, i.e. low skeletal growth or linear growth retardation, as measured by children’s height-for-age, may be influenced by children’s socioeconomic conditions through to middle childhood (Lundeen et al., 2014; Schott et al., 2013). The research further showed that children who experience growth recovery in early childhood demonstrate improved cognitive outcomes than children who are persistently stunted (Crookston et al., 2013). Findings from related Young Lives research showed that one-year old children in the top wealth quartile were taller than children in the bottom quartile and that this disparity persisted over the course of childhood, at age five, eight and 12 years in all study countries (Reynolds et al., 2017). Similarly, children in the top quintile achieved higher scores on the PPVT vocabulary test than children in the bottom quartile. However, the results were mixed across study countries: wealth disparities in cognitive outcomes were constant in India, widened in Ethiopia and narrowed in Vietnam and Peru. Hamadani et al. (2014) showed that the widening wealth effect in early childhood in rural Bangladesh was mediated by parental education and home stimulation at ages 18 and 64 months, reinforcing similar findings in Madagascar (Galasso et al., 2017), as well as previous research from high-income countries.

A recent study in Bogota, Colombia (Rubio-Codina and Grantham-McGregor, 2019) administered the Bayley Scales, which is a comprehensive diagnostic developmental assessment, to a sample of over 1,000 children at age 6-42 months. The results showed a large and significant wealth effect on cognitive development which increased in age. At age 6-8 years, the children were assessed on three additional constructs. IQ was assessed using the Wechsler Intelligence Scale for Children (WISC-V), school achievement was assessed using the Woodcock-Munoz Test of Achievement (WM-III) and the Peabody Picture Vocabulary Test (PPVT-R), and the Strengths and Difficulties Questionnaire (SDQ, Goodman, 1997) was used to assess behaviour. The purpose was to investigate the evolution of the wealth effect and to identify potential mediating factors at different stages of the childhood. Bearing in mind the instruments used to assess development differ across sweeps, the results showed an increase in the wealth effect on cognitive outcomes by 6-8 years and demonstrate the importance of early developmental delays for later cognitive deficits. The gap in cognitive outcomes between the richest and poorest increased more rapidly in early childhood compared to middle childhood. A large and significant IQ gap was found between wealth quintiles at 6-8 years. Parental education and the home learning environment, measured in early childhood, mediated the wealth effect on cognitive outcomes at 6-42 months, IQ at 6-8 years and school achievement at 6-8 years. Home stimulation in middle childhood also mediated the wealth effect on IQ and school achievement, although the mediation effect was stronger in early childhood.

2.3.4 Living standards measurement

While there is a global consensus that poverty negatively affects children's growth and development (Grantham-McGregor et al., 2007; Walker et al., 2007; Engle et al., 2007), there is disagreement on how poverty should be defined and measured for meaningful investigation of this relationship. Effective targeting of programmes and services to alleviate poverty effects requires some way of quantifying poverty and identifying poor families, and there are different ways of achieving this. The first section of this review provides an overview of the theoretical foundations underpinning quantitative poverty research, with a focus on the differences between the prevailing living standards constructs and the corresponding measures used to identify the poor. The aim is to consider the main issues arising in the literature that apply to the present study.

The field of living standards measurement has experienced rapid theoretical and methodological transformation in recent years, yet traditional monetary approaches remain the most dominant method of quantifying living standards, and in turn the incidence of poverty. Although adopted mainly within the economics paradigm, the origin of monetary poverty is generally attributed to Benjamin Seebohm Rowntree, a British sociological researcher who sought to formalise an objective way of quantifying poverty in Britain. Rowntree defined poverty as the minimum income level required for subsistence and operationalised this definition by calculating the amount needed to maintain a family in a state of 'physical efficiency' based on estimates of a minimum food, clothing and shelter requirement (Rowntree, 1901: viii, 87). His publication, *Poverty. A study of town life*, provided the first statistical description of living standards in England, contributing some of the earliest quantitative evidence of poverty. However, Rowntree's research was not necessarily guided by the establishment of an absolute poverty line, or minimum income threshold. Rather, Rowntree was concerned with the living conditions of the poor in relation to the non-poor and sought to demonstrate that this disparity could be attributed, in part, to differences in income (Veit-Wilson, 1986a; Veit-Wilson, 1986b). As such, some scholars have argued that his poverty construct should be viewed as a relativistic model rather than absolutist one (see Veit-Wilson, 1986a).

Nevertheless, Rowntree's work formed the basis for methodological advancements towards the characterisation of an income/consumption-based poverty line in traditional welfare economics, i.e. the normative utilitarian framework. The utilitarian model describes individuals as maximising their utility, a construct representing wellbeing, measured either by income or by monetary expenditures on food and non-food commodities. Within this paradigm, poverty is defined by a lack of utility, or more concretely, "a shortfall in consumption below some minimum level of resources," termed the poverty line (Laderchi et al., 2003: 248). Several approaches have been advanced to estimate this minimum threshold, with the 'cost of basic needs' approach being the most common. The method generally involves specifying a basket of goods deemed adequate for satisfying basic consumption needs and using local price data to estimate the cost of those needs (Haughton and Khandker, 2009). Individuals or households whose income/expenditure falls below the poverty line are deemed poor. In application, there are various challenges to constructing reliable poverty figures once the choice of a poverty line is justified, such as: deciding what the minimum basket should contain in order to cover the needs of all household members, including children (Pradhan et al., 2000; Bradbury and Jantti, 1999); tackling the problems of collecting accurate income/expenditure and market price data (Grosh and Glewwe, 2000); and ensuring trustworthy data interpretation (Haughton and Khandker, 2009; Deaton and Grosh, 1998). For example, in non-Western societies, income and expenditures may not necessarily reflect whether individuals/households meet their minimum requirements, e.g. in the case of subsistence agriculture. Moreover, in the absence of functioning markets, prices of certain commodities may be ambiguous. Subsequent imputations of monetary expenditures may be substantial, affecting the reliability of statistical inferences based on the data (Deaton and Grosh, 1998).

Whereas in Western contexts household income or expenditures may be a good indicator of consumption patterns, this may not be true in non-Western societies. In recent years, quantitative poverty research in developing countries has seen a rise in the application of poverty measures that are based on what is actually consumed by households, rather than on household earnings and spending patterns. The rationale underlying this movement is partly based on practical experience that income and expenditure data are rather difficult and expensive to collect, but also justified on the basis that while income is only received intermittently, consumption is smoothed over time and is a more direct indicator of actual living standards (O'Donnell et al., 2008). This is particularly relevant in societies where income fluctuates substantially in the short term, "as in rural agriculture –

the ranking of households by income will usually be much less stable than the ranking by consumption" (Deaton and Zaidi, 2002: 13).

In settings where income/expenditure or consumption data are not available or difficult to collect, summary wealth indices based on household ownership of durable goods and housing characteristics, are constructed as alternative indirect or "proxy" measures of living standards (for example, Filmer and Pritchett, 2001). From these data, wealth quintiles are computed and assessed to identify the poor and to characterise gaps in human developmental outcomes (Rutstein and Johnson, 2004). Thus, unlike poverty incidence measures (e.g. poor vs. non-poor), constructed quintiles enable comparisons of the relative standing of households across a continuum of wealth index scores. Wealth-related indicators are included in the Demographic and Health Survey (DHS) to monitor wealth effects on population, nutrition, and health outcomes and in UNICEF's Multiple Indicator Cluster Surveys (MICS). However, wealth and income/consumption are often weakly correlated (O'Donnell et al., 2008). Whereas income and consumption are flow variables, wealth "represents a more permanent status" (Rutstein and Johnson, 2004: 4), and is not expected to change significantly over time. The choice of poverty measures then becomes important for interpreting relative poverty rankings. For example, while wealth may be important for understanding inequalities in various outcomes, it may not efficiently describe how such outcomes are affected by changes in living standards.

In focusing on solving the methodological problems, the utilitarian framework has not given priority to debating critically the validity of the monetary poverty measures and the assumptions underpinning them. Within the economics paradigm, a large literature has amassed based on Sen's seminal critique of utilitarianism, and in turn, monetary poverty constructs (Sen, 1983, 1985a, 1985b). For Sen, a poverty measure should not be based on utility but on a person's capability: the freedom a person has to achieve valuable functionings (Sen, 1985a: 3-4; Sen, 1985b: 200-201), which, it is argued, better captures the notion of wellbeing. The capability construct taps two sets of goals: the first relates to the goals of individual agency, or the ability to choose. The second concerns the achievement of valued functionings, defined by Sen as the ‘beings and doings’ that a person may value and have reason to value (Sen, 1985b: 197-198; Sen, 2001: 18). However, the vast majority of empirical work in this area of research focuses on the measurement of achieved functionings, using direct indicators of wellbeing. Operationalising the ‘freedom’ dimension of the capability construct has proven difficult. Therefore, in practice, existing measures do not directly account for agency objectives, and are thus similar to the basic needs approach. The Multidimensional Poverty Index (MPI), published in the United Nations Human Development Report, provides an example of such a measure. It is based on Sen’s reasoning that capability measures should be developed around the “irreducible absolutist core” set of capabilities (Sen, 1983: 159) – “the need to meet nutritional requirements, to escape avoidable disease, to be sheltered, to be clothed, to be able to travel, and to be educated” (Sen, 1983: 162-3).

The MPI conceptualises poverty as multiple deprivations that are simultaneously experienced, based on direct indicators of achieved core functionings (Alkire and Santos, 2010; Klasen and Dotter, 2013; Calderon and Kovacevic, 2014). An individual who is below the threshold in one-third or more of the weighted indicators is identified as MPI poor (Alkire et al., 2011). UNICEF has adopted a similar index for assessing child poverty, the Multiple Overlapping Deprivation Analysis (MODA), which is conceptualised based on the MPI (de Neubourg et al., 2013). MODA has been widely used to analyze child well-being. The methodology defines multi-dimensional child poverty as non-fulfilment of basic rights contained in the UN Convention on the Rights of the Child and concludes that a child is poor if he or she is deprived in three to six age-specific dimensions. Such measures have been criticised on methodological grounds, and on the basis that they are difficult to interpret and may lack value-added over monetary measures (Ravallion, 2010: 15, 29; Rippin, 2011). However, these criticisms have not gained much momentum as evidenced by the mainstreaming of the Global MPI as a complement to monetary poverty statistics and the widespread application of MODA worldwide (World Bank, 2018; Alkire et al., 2016).

Solving the complexities initially raised by Rowntree, followed a very different path in sociological research. Theoretical transformations in sociology were guided by a separate set of normative assumptions in that the formalisation of the poverty question was not rooted in understanding if people were happier (utility-based) or freer (capabilities-based), but rather on the extent to which people were included in society. Thus, in the post-Rowntree era, the problem for sociologists was to understand disparities in living standards in the context of the broader distributional aspects of wealth in that society. The sociological researcher Peter Townsend, who initiated the conceptualisation and measurement of poverty in this way, defined people living in poverty as those who,

> lack the resources to obtain the types of diet, participate in the activities, and have the living conditions and amenities which are customary, or at least widely encouraged or approved, in the societies to which they belong. Their resources are so seriously below those commanded by the average individual or family that they are, in effect, excluded from ordinary living patterns, customs and activities.
>
> (Townsend, 1979: 31)

In later writings, Townsend further reasons that the inability to attain the prevailing conventional living standard (e.g. nutrition, health, shelter, clothing and so on) may prevent an individual from being able to "play the roles, participate in the relationships and follow the customary behaviour which is expected of them by virtue of their membership in society" (Townsend, 1987: 130). Townsend's definition of poverty is therefore distinct in two ways. First, there is the primary objective of social participation as an end itself. Second, there is the notion of relative deprivation, or the failure to achieve a socially defined minimum standard of living. For Townsend, achieving a minimum standard provides information about how well a person can integrate into society. Therefore, while Townsend's notion of poverty is also operationalised by direct indicators of poverty, for each indicator the threshold is defined based on the prevailing social standard, i.e. the mean level of achievement (see Townsend et al., 1988). Townsend's theory has inspired various applications and forms the basis for the European Union 'at-risk-of-poverty' threshold defined as the share of individuals with a disposable income below 60% of the national median level (Eurostat, 2014: 125). The approach has also been adapted for assessing poverty among children in industrialised countries and in the developing world (Bradshaw et al., 2000; Gordon et al., 2003).

The main criticisms of relative deprivation are both conceptual and methodological. Sen argues that the relativist definition is not actually one of poverty but a rather "imperfect representation of inequality" (Sen, 1983: 156). A moving poverty line, adjusted for changing living standards, implies that poverty cannot be eradicated as a proportion of the population will always worse off "in the sense that they receive below-average incomes" (Fiegehen et al, 1977: 2-3; Sen, 1983: 156). Moreover, the effect of major economic shocks, causing general adversity for all and extreme hardship for some would go unnoticed "if the relative distribution is unchanged" (Sen, 1983: 157). Thus, while relative measures may have applicability in highly industrialised societies, where most people can achieve basic necessities and "sharp declines [in prosperity] are most unlikely" (Sen, 1983: 157), they may be less appropriate in developing countries where such challenges do exist. From a methodological standpoint, scholars have argued that the process of identifying indicators of deprivation and persons falling below the minimum standard is arbitrary and fails to account for individual choice in achievement, i.e. social or cultural decisions (Piachaud, 1987). Mack and Lansley's (1985) *Poor Britain* sought to address these issues by engaging the public in the selection of "socially perceived necessities" (Mack and Lansley, 1985: 173-174) and by distinguishing between those who could afford a particular item/ activity but did not want it, from those who could not afford it (Mack and Lansley, 1985: 174-178). In their report, *Poverty and Social Exclusion in Britain*, Gordon et al. (2000) expanded this investigation using a similar approach to evaluate material deprivation and social exclusion (Gordon et al., 2000). However, neither studies have resolved the theoretical criticisms made by Sen and others.

Table 2.3 summarises the main features of the different approaches to poverty measurement discussed in this review.

Table 2.3 Characteristics of key living standards measures

Theoretical framework	Poverty construct	Poverty threshold
Utility	Monetary poverty	Min. income, consumption, wealth
Capability	Multidimensional poverty	Proportion of weighted indicators
Social participation	Relative poverty	Median income
Social inclusion	Domains of social exclusion	At risk of exclusion

2.3.5 Summary

This section explored poverty effects on children's cognitive development from theoretical, conceptual, and empirical standpoints. Section 2.3.1 documented evidence of the effects of poverty on children's cognitive development. A key finding from the review is that socioeconomic disparities in cognitive development emerge early in life and grow rapidly. While there has been good progress towards understanding the effects of poverty on children's development and learning over time in industrialised countries, the time effects of poverty are less well researched in low-income countries. The challenges of collecting longitudinal data and measuring cognitive outcomes restrict the possibility of understanding longer-term growth trajectories and the pattern of change in the poverty-cognitive growth relationship. The studies that do exist have shown that initial gaps in development widen in early childhood, and lead to gaps in other areas of development and learning. However, the findings on whether initial gaps continue to widen through to later childhood are mixed.

Section 2.3.2 discussed some of the main factors identified in the literature as important for understanding the poverty-child development relationship, specifically: family background, the home learning environment, children's health and nutrition, and exposure to quality pre-schooling. Other influences on cognitive development that have been highlighted in the literature have not been discussed here. These influences can be broadly classified as sociocultural factors (e.g. gender inequity and access to services), psychosocial factors (e.g. maternal depression and exposure to violence) and biological factors such as prenatal and postnatal growth (Walker et al., 2007). The focus of the current research will be on the contextual factors that are known to vary with socioeconomic status and that are included in the Young Lives data structure. Understanding the pathways through which poverty contributes to developmental deficits over the course of childhood is necessary for the effective design and targeting of policies and programmes in Ethiopia. However, as this review has shown, there are few longitudinal studies that sufficiently measure developmental outcomes over the full span of childhood in low-income countries. Without such information, cognitive trajectories remain poorly understood, and thus strategies for mitigating the effects of disadvantage in later childhood are not prioritised.

In section 2.3.3, concepts and operational measures of poverty across social science disciplines are discussed with a view to identify a way forward for the current study. Based on the researcher's understanding of the primacy of basic needs in Ethiopia and the importance of wealth for early cognitive performance, a monetary-based measure of poverty is deemed appropriate. The research questions are centred on understanding how poverty shapes individual developmental trajectories and the mechanisms by which poverty exerts its influence. The wealth index is the primary instrument used in the Young Lives survey to measure living standards. Despite oversampling districts with a food-deficit status, and hence a higher representation of poor families in the sample, variability in the wealth data is comparable to the variability in living standards outcomes found in other Ethiopian population-based surveys (Outes-Leon and Sanchez, 2008: 22). The second measure of living standards available in the data structure is a consumption aggregate of household food and non-food expenditures. As wealth represents a more permanent state than consumption, it may be more important for understanding inequalities in outcomes and less efficient in describing how such outcomes are affected by changes in living standards. Both measures are discussed in more detail in the methodology chapter.

A final aspect of the review relates to the measures used to assess child outcomes. Documentation of test design and adaptation efforts is rarely explicit in child developmental research in low-income countries. To describe, explain and subsequently improve cognitive outcomes among poor children, the educational and psychological measures used need to be psychometrically sound, designed with rigour, and adaptation processes need to be systematic and transparent. This issue is explored in the next section.

2.4 Measurement of children's cognitive development in context

The aim of this final phase of the literature review is to better understand the properties of good measures to identify a framework for assessing the reliability and validity of the PPVT instrument used in the current research. The review explores the main measurement issues in cross-cultural educational and psychological assessment that pertain to the present study, specifically: (i) cultural sources of test bias and the validity of the claims made using biased test data; (ii) test translation vs. test adaptation as a process of decomposing broad concepts into culturally meaningful constructs and operational measures; (iii) transparency and documentation of test development, adaptation and administration processes; (iv) the focus on solving methodological problems rather than substantive considerations about the appropriateness of measures; and related to this point (v) the reliance of psychometric research on statistical theory ahead of educational and psychological theory.

2.4.1 Reliability, validity, and test bias in cross-cultural assessment

One of the main conclusions to emerge from the previous section of the review is that although good progress has been made towards documenting the effects of poverty on children's cognitive development in low-income countries, arguments are rarely provided to support the appropriateness of the measures used to assess development and learning. Researchers should be aware of certain challenges when testing children across cultures or in unfamiliar contexts. However, the absence of documentation on test development and administration processes suggests that addressing these challenges is not prioritised in child developmental research in low-income countries.

This section draws on the concepts of *reliability* and *validity* as criteria for evaluating the performance of assessments. In educational research, "… [reliability] is the name given to one of the properties of a set of test scores – the property that describes how consistent or error-free the measurements are" (Frisbie, 1988: 25). A consistent test yields a set of scores that are reproducible, i.e. that can be obtained by the same group of individuals, within the same timeframe and under the same testing conditions. Linked to reliability is the notion of test validity which has traditionally been defined in terms of four interrelated ideas: content validity, concurrent validity, criterion validity and construct validity. More recent literature refers to the primacy of construct validity, which incorporates the other three aspects. As stated by Messick (1986), the construct validity of an assessment instrument is "an overall evaluative judgement, founded on empirical evidence and theoretical rationales, of the adequacy and appropriateness of inferences and actions based on test scores" (Messick, 1986: 2). Similarly, for Cronbach, "validation of an instrument calls for an integration of many types of evidence. The varieties of investigation are not alternatives any one of which would be adequate. The investigations supplement one another" (Cronbach, 1971: 445).

Related to concepts of reliability and validity is the notion of *test bias*, which is generally defined as "deficiencies in a test itself or the manner in which it is used [that] result in different meanings for scores earned by members of different identifiable subgroups" (AERA, APA and NCME, 1999: 74). An unreliable or biased test would therefore be indicated by differences in the test scores between individuals that are not explained by differences in their level of content knowledge. Sources of bias/error might include factors external to the respondent, such as inconsistencies in test administration and scoring bias. Motivation, health, and guessing are all examples of sources of measurement error that are internal to the respondent. Another way to think about measurement error is in terms of whether the source of error is systematic or random. Sources of systematic error include sampling errors which occur when there is a difference between the sample and the population from which it is selected (Bryman, 2008); non-response, which happens when individuals who are being sampled do not cooperate or cannot supply the required information; participant attrition; and errors on the instruments themselves (Bryman, 2008). Examples of random error include factors internal to the respondent or to the data collector, situational glitches in data collection, errors in data entry, among many others. Systematic errors produce distortions that affect the validity and generalisability of research results, in contrast with random error which affects the reliability (Malmberg, 2011).

While an instrument may be proven reliable and valid in its source context, retaining these properties in a different socioeconomic or cultural setting is a challenge. With the expansion of international comparative assessments in recent decades, the adaptation of a test from one context to another has rapidly grown in practice. Reliability and validity issues relating to cross-cultural educational assessment have been discussed widely in the literature (Harkness et al., 2002; Prais, 2003; Goldstein, 2004; Hopmann, Brinek & Retzl, 2007). The most prominent criticism is that during the test development stage, emphasis is usually placed on test translation with inadequate attention given to cultural sources of influence on instrument adaptation and score interpretation (Hambleton and Patsula, 1998: 154). Moreover, multiple cultural sources of error can occur during an assessment process, from the test design phase through to administration and results interpretation. Errors are often linked. In a multicultural testing situation where children from different language groups are being assessed on a specific ability in different parts of a country, language bias might occur alongside biases in testing procedures, in that administrators follow different local testing standards as opposed to a single common standard. For example, researchers involved in the *WHO Motor Development Study* which sought to construct standards to assess growth and development in children from birth to five years of age, found that sample heterogeneity, differences in test administration procedures and scoring differences explained part of the high variance in the age of milestone attainment (Lansdown et al., 1996). The authors state: "although there was high intertester reliability within [testing] centres, constraints of time and money prevented any attempts at achieving reliability checks between centres" (Lansdown et al., 1996: 288). The author's further point out the risk of researcher assumptions about children's exposure to certain item wording or stimuli. For instance, test items may give certain urban children an advantage over those living in rural areas, and vice versa.[3]

Concerns about test bias have grown alongside advancements in developmental psychology which saw a movement towards a more nuanced understanding of culture as a system that is constantly shaping and being shaped by the process of human development. This involved a shift from accounting for the role of culture in human development as a broad level of external influence (e.g. ethnicity, religion) towards modelling culture as a complex process embedded within a social, environmental and historical context (Whiting, 1980; Super and Harkness, 1997 and 1999; Harkness and

[3] Other studies that are attempting to standardise early childhood developmental milestones in developing countries include: the Kilifi Developmental Inventory (Kenya); the Grover-Counter Scale of Cognitive Development (South Africa); the Philippines Early Childhood Care and Development (ECCD) Checklist; ICMR Psychosocial Developmental Screening Test (India).

Super, 2002; Rogoff, 1995). Consequently, methodological developments in cross-cultural psychology evolved on the basis that,

> The use of culture as an explanatory variable is not satisfactory, and culture must be decomposed into a set of psychologically meaningful constructs, which are then used to explain the differences observed.
>
> (Van de Vijver and Leung, 1997: 260)

2.4.2 Cross-cultural test adaptation and transparency in reporting

Test adaptation involves delving deep into the culture under study to ensure that the test format, its content, and the testing process are familiar and meaningful for the respondents (see Van de Vijver and Leung, 1997; Hambleton and Patsula, 1998). For a test to provide valid and reliable information, adequate resources are needed, i.e. financial, human and time, to design, adapt, field-test, revise instruments and train fieldworkers, ensuring that the tools are working as intended. The notion of test adaptation has gained momentum in the literature and has become an important priority for organisations involved in international educational assessment. Institutions such as the International Association for the Evaluation of Educational Achievement (IEA) and the Organisation for Economic Cooperation and Development (OECD) have developed methodologies for identifying potential sources of cultural bias in comparative testing, mainly at the test development stage (OECD, 2014) and for addressing cultural differences in response behaviour (OECD, 2014). However, the withdrawal of several countries from the 2012 Programme for International Student Assessment (PISA) suggests that cultural bias remains a significant issue.[4] As noted by a representative of one of the countries that dropped out of the PISA project, “if the child hasn’t heard of airbags, hot air balloons and ATMs, he won’t even attempt those questions. Also, such detailed questions would be confusing; straight answers would work better” (Rebeiro,

[4] Countries that did not participate in the 2012 learner assessments are: Azerbaijan, Georgia, India, Kyrgyz Republic, Malta, Mauritius, Miranda (Venezuela), Moldova, Panama.

2012). The comment illustrates concerns over content bias and bias in testing conventions.

While the international organisations involved in educational and psychological testing are continuously revising guidelines and methodologies for reducing bias in international assessments (OECD, 2018), in developing world research, advancements are taking place at a much slower rate. Test localisation refers to the process of adapting assessments to "suit the context in which they are administered in order to ensure that the data they produce is valid and reliable" (Nag et al., 2017: v). However, in low-income countries, where research funds are overstretched, test adaptation is often sidestepped or under-resourced (Nag et al., 2017). This is a particular concern in countries where learning assessments are not only used to evaluate the impact of programme interventions but are also being scaled up and used for population-level monitoring of developmental and learning outcomes to inform educational systems reform.

Recent evidence from a systematic review of literacy and learning assessments in developing countries showed that despite a growing awareness of the influence of language characteristics and other cultural factors on test performance, test localisation efforts are rarely documented in research (Nag et al., 2017). The authors found that while a large proportion of the studies mention translation, they do not offer details about the translation procedures used, or about how local experts were mobilised to support the process, i.e. local psychologists, educationalists, linguists. The scope of the review comprised studies using assessments of language and literacy skills administered to children between ages three and 14 years over the period 1990-2014, covering 55 languages in 53 countries. All studies reviewed were rated as 'Moderate' or 'High' in methodological quality. However, most of the studies were found to have adopted only "token localisation" procedures if any:

> Where translation is mentioned, the use of back translation is not, and it is difficult to decide if this important step was bypassed or overlooked in the

> reporting. Where local experts have been used, their qualifications have typically been university-based academics in the field of language, education or linguistics, or they are school teachers. The guidance note for quality checks of items is not reported and again it is unclear if there was consensus among those consulted, and when there wasn't, how the differences were reconciled. A further attempt at localisation is to align the material to the local school textbooks. Here, there are a range of localisation attempts, ranging from calculating type and token frequencies of words in textbooks to gathering impressionistic data on whether items reflect curricular targets.
>
> (Nag et al, 2017: 35)

Test reliability information was not provided for a large majority of the reviewed studies (60% for foundation literacy and 58% for language),[5] raising questions about the robustness of the measures. Of the studies that did report reliability information based on the Cronbach's alpha statistic, which is a measure of internal consistency reliability, in many cases the values were well below the minimum acceptable level of 0.60, although 0.70 or higher is the recommended criterion (Nag et al., 2017: 83-86; Nunnally and Bernstein, 1994).

The systematic review concluded that there is significantly better reporting of localisation/ adaptation among researcher-developed tools, in terms of test design, field-testing, field-worker training, etc…, compared with commercial test adaptations. Researcher-developed tools can be defined as those which are generally used in smaller samples and in response to specific research questions. Commercial tests are those that have been developed in Western countries that can only be acquired by paying a procurement fee to the test developer (see Table 2.3). Commercial tests are increasingly being used at-scale in developing countries to monitor child development and learning across populations. While these tests may have demonstrated good psychometric properties in their source context, less is known about their reliability and validity when used in other cultural contexts (see section 2.3.2). One exception is the Measuring Early Learning Quality and Outcomes (MELQO) instruments, which is intended to be operationally free-to-use and

[5] Author's own calculation based on the data provided in Annex B of Nag et al., 2017.

flexible for local adaptation according to the test developers. The MELQO is derived from widely validated international measures of learning quality and outcomes. It consists of two sets of linked instruments, a direct child assessment which is used to measure literacy, maths and executive function and a classroom observation instrument, which is used to assess the features of the early learning setting, teaching/learning activities and child-teacher interactions. The core tools are complemented by caregiver and teacher questionnaires, which can be used to better understand the characteristics of the home learning environment, family socioeconomic contexts, children's social-emotional wellbeing, as well as aspects of teaching and learning that cannot be efficiently captured by observing the class/teacher. The instruments are accompanied by detailed manuals and guidelines for instrument adaptation, piloting, and fieldworker training (UNESCO, UNICEF, Brookings Institution and the World Bank, 2017).

2.4.3 Methods for addressing bias in development and learning assessments

There is a plethora of strategies for detecting bias or "systematic error" in an assessment. In cross-cultural research, the use of various statistical methods to verify the *equivalence* of different versions of a test is a common way of establishing the extent of bias and the comparability of scores. These methodological advancements are due in part to a handful of ground-breaking studies that revealed how the performance of instruments can vary widely across cultures (Poortinga, 1995; Van de Vijver and Leung, 1997; Van de Vijver and Tanzer, 2004; Van de Vijver and Poortinga, 1997). However, there is some disagreement about what constitutes sufficient statistical evidence of instrument equivalence, or *invariance*.

The most widely applied test of invariance in the social and behavioural sciences involves analysing the equivalence of covariance structures to assess whether the underlying factor structures of the scores obtained from two separate groups are equal (Schoot et al., 2012). The argument is that a good-fitting factor structure across sub-samples, or factorial invariance, indicates measurement equivalence.

However, the approach has been critiqued based on evidence that, “although the factorial structure of a measuring instrument may yield a similar pattern when tested within each of two or more groups, such findings represent no guarantee that the instrument will operate equivalently across these groups” (Byrne and Watkins, 2003: 156). In other words, the factorial structure may indicate measurement equivalence within two contexts (i.e. countries or cultural groups), but that does not guarantee equivalence across these contexts. Byrne and Watkins (2003) define measurement invariance as consistency in the operation of the items on a test, or the extent to which the content of each item has the same meaning across samples, which differs from structural invariance, defined as consistency in the definition of the underlying construct (Byrne and Watkins, 2003). The authors draw on a sample of 497 Australian and 493 Nigerian adolescents to assess equivalence of the Self-Description Questionnaire (SDQ-I), a commonly used measure of self-concept. The questionnaire examines physical and social self-concept (Marsh, 1992). Using confirmatory factor analysis techniques to study the factor structures, and ANOVA to investigate item bias, the findings showed that despite well-fitting factor structures within each of the sampled cultural groups, the instruments showed evidence of structural non-equivalence. Specifically, the correlation between latent factor structures of the physical self-concept subscales: ability and appearance, was higher for Australian adolescents. The authors conclude that the difference is likely due to differences in the social construct of *physical attractiveness*.

The study underscores concern about “whether instruments can ever be totally equivalent when used in cross-cultural comparison”, even when constructs are similarly defined, and the factor structures fit well for both cultural groups (Byrne and Watkins, 2003: 155). The results therefore raise questions about whether comparisons of relative standings between cultural groups can ever be valid and reliable without first verifying structural and measurement invariance. It further highlights the pivotal role of robust adaptations, the need for procedures to identify and address potential sources of bias at each stage of the assessment cycle, and the importance of documenting the efforts taken to minimise bias in cross-cultural assessment research. While statistical procedures can be useful for spotting non-equivalence and identifying problematic items, these methods cannot account for the source of the bias once detected. Such qualitative judgements would need to be based on researcher experience and records of the assessment process.

Drawing on previous work (Van de Vijver and Lenug, 1997; Van de Vijver and Poortinga, 1997), Van de Vijver and Tanzer (2004) propose a systematic approach to identifying and addressing bias at each stage of the assessment cycle. *Construct bias* occurs when a measure fails to capture the same construct in different cultural groups, either due to differences in how the construct is defined across the cultures, or differences in how construct is operationalised (also see Helms, 1992; Hui and Triandis, 1985; Harkness, 1998). The definition of the constructs and the specific aspects that are selected to represent the construct may not fully overlap between cultures. The notion of children's *social-emotional competence* provides an illustrative example of how this conflict might manifest. While it is well established that social-emotional health is central to children's development and learning, "cultural expectations for how relationships are formed, what appropriate social interactions look like, and how children should regulate their emotions means that these constructs may not look the same in all cultures" (UNESCO et al., 2017: 38). The example of *intelligence* has also been used in the literature to demonstrate this phenomenon. While intelligence is a very broad concept covering many abilities, standardised measures of intelligence generally assess the cognitive aspects of intelligence only, and neglect to consider the social domains which may be more important for non-Western societies (Van de Vijver and Poortinga, 1997: 30; Van de Vijver and Tanzer, 2004). Concepts like "happiness" and "depression" provide additional examples of constructs that require substantial contextualisation for meaningful measurement. Across subgroups within a population, there may be substantial differences in the way a construct is defined, or it may be the case that different test items are needed in different cultures (Vijver and Tanzer, 2004).

Method bias occurs when a cultural factor that is irrelevant to the construct studied affects most of the items of a test in a differential way across the cultures of study. Cultural factors such as acquiescence and social desirability may affect group response styles and, subsequently, the comparison of test scores across groups (Hambleton and Patsula, 1998: 159). Test administration processes may produce unintended biases due to differences in the way enumerators ask questions. Related to this is differential effect of certain testing formats on assessment outcomes,

> … even simple geometric figures are culturally entrenched and not equally familiar across cultural groups. Exposure of children to such figures will substantially differ across cultures; differential exposure will influence the test results and challenge score comparability more in cross-cultural than in intracultural research.
>
> (Van de Vijver and Poortinga, 1997: 30)

Picture stimulus familiarity is an important source of methods bias in assessments like the Peabody Picture Vocabulary Test (PPVT), as evidenced in Kierscht and Vientze's (1975) study of pre-school aged American children. When asked to label objects, coloured pictures and black-white line drawings, the children were able to label the objects more successfully than the pictures and the discrepancy was more pronounced for low-income children. The findings reinforce complementary evidence that highly representational picture stimuli require a higher level of cognitive skill, and that children with less exposure to such forms will be at a greater disadvantage (Coleman et al., 1980). Similar findings were found in Deregowski and Serpell's (1971) cross-cultural study of Scottish and Zambian children performing an object and picture sorting task. While both groups of children were equally able to sort the objects, the Scottish children performed better on picture sorting (Deregowski and Serpell, 1971).

Finally, *item bias* refers to anomalies in the data that relate to specific individual items, operationally referred to as differential item functioning (DIF) in the literature (Holland and Wainer, 1993). DIF approaches are used to assess group bias at the item level (Holland and Wainer, 1993; Ercikan et al., 2004; Sireci, 2011; van de Vijver and Leung, 1997; 2011), and bias over time when repeated measures are used (Gebhardt and Adams, 2007; Kreiner and Christensen, 2013; Robitzsch, 2010). To establish comparable versions of a test across cultures, items that are least effective in assessing the construct are usually eliminated or modified (Tenant et al., 2004; Tenant and Conaghan, 2007; Hagquist and Andrich, 2004; Lawton et al., 2006).

The focus on solving the methodological problem of detecting and addressing the effect of bias on results has produced a range of statistical techniques. Procedures are generally classified according to the source of measurement error (see Van de Vijver and Poortinga, 1997; Van de Vijver and Tanzer, 1997, 2004; Hambleton, 2005). Differential item functioning approaches are used to assess item level bias by group or across time (Holland and Wainer, 1993; Ercikan et al., 2004; Sireci, 2011; van de Vijver and Leung, 1997; 2011; Gebhardt and Adams, 2007; Kreiner and Christensen, 2013; Robitzsch, 2010). Exploratory Factor Analysis and Confirmatory Factor Analysis approaches are commonly applied to test the equivalence of tests (Van de Vijver and Hambeleton, 1996; Hambleton and Patsula, 1998, 1999; Gierl, 2000; Hambleton, 2005; Konishi et al., 2009; Loner and Peri, 2009; Byrne and van de Vijver, 2010; Hambleton and Lee, 2013). Overall test performance is examined using test reliability coefficients, such as the Cronbach's alpha, which is a measure of internal consistency, and the McDonald's omega, which further takes into account item-specific measurement error and the extent to which each item on the tests contributes to the measurement of the underlying construct (Oliveri and Kadriye, 2011; McDonald, 1999). Item response models are used to examine the properties of the items, i.e. item targeting and item misfit (Birnbaum, 1968; Rasch, 1980).

Methodological debates have not given as much attention to qualitative judgements that cannot be drawn through statistical analyses of the test data (Van de Vijver and Poortinga, 1997), thus overlooking theoretical concerns in educational or psychological assessment. An assessment process should begin with a clear definition of the core concepts, of how operational constructs are understood across the cultures under study, and of the meaning underlying each of the test items from the perspective of both respondents and test administrators. From there it is possible to design different versions of tests and to establish whether the tests tap the same construct. Similarly, minimising bias in testing methods requires investments in administrator training and records of administration procedures, test papers, and any stimulus used during testing to shed light on irregularities.

2.4.4 Paradigm developments in psychometric research

As the current study involves both educational measurement (i.e. vocabulary ability) and data modelling (i.e. developmental trajectories), the specific aim of this section of the review is to identify a strategy for constructing a valid and reliable measure of cognitive ability that is suitable for a growth analysis within a multicultural sample of children. The review does not attempt to summarise the historical origins of measurement theory. Rather, the main differences between the main competing research traditions are presented to highlight both the conceptual and practical implications for current research.

Within Classical Test Theory (CTT), the focus of analysis is on the total test score and emphasis is placed on the overall quality of the test, rather than the quality of the items on the test. CTT follows a deterministic model such that the level of a given construct is indicated by the total score with some error of measurement. A total score always indicates the same level of the construct regardless of which items are answered correctly. For any given test in which the items are indicators of the same construct, each person v has an observed score on the test, X_v which is composed of the person's true score on the construct, T_v and the error of measurement on the test, E_v.

In the classical paradigm, a reliable test is one in which the observed-score variance is largely attributable to the true score variance and minimal error variance. The overall reliability of the test, $r_{xx,}$ refers to the extent to which the set of observed test scores is error-free. It is assessed as a proportion of true-score variance to observed-score variance and has a range of zero to one. A value of one indicates no measurement error or, in other words, that the observed-score variance is attributable to the true scores (high reliability); low reliability is indicated by a small true score variance relative to the observed variance (Traub and Rowley, 1991). There are ways to estimate reliability in CTT, the Cronbach's alpha, a measure of internal consistency, is the most widely reported statistic. However, more recent thinking points out that "[the alpha coefficient] covers only a small perspective of the range of measurement uses for which reliability information is needed… [the alpha coefficient] is now seen to fit within a much larger system of reliability analysis" (Cronbach and Shavelson, 2004: 416).

With the focus on the overall performance of the test, item analysis within the traditional framework is significantly limited, which has implications for test validity. How the items contribute to the measure is not central to the assessment of how reliable the test is (Christensen et al., 2012). In CTT, three simple indicators are used to assess how well the items of a test function. The *difficulty index*, which is the proportion of people who answer an item correctly. Extreme items with values close to zero (very difficult) or values close to one (very easy) do not adequately differentiate between individuals and should be discarded (Matlock-Hetzel, 1997: 6). The *discrimination index* is used to measure the degree to which an item discriminates between persons showing high- or low-test performance, which assesses whether the items target individuals as expected. It is computed as the difference between the proportion of respondents who answered the item right in the upper scoring group and the proportion of people who answered correctly in the lower scoring group. There is no specific guideline for what constitutes a well-discriminating item apart from more discrimination is better (Masters, 1988: 17). Finally, the *non-response rate*, the proportion of respondents who do not answer an item, may also be reported. This rudimentary descriptive statistic can be used to identify items which may have been problematic for respondents and thus should be modified or discarded. With the focus on the overall reliability of the test, construct validity tends to be argued in terms of factor analytic evidence that items in the scale contribute to the measurement of a single construct.

Whereas in the traditional paradigm inferences are made about an individual's true score given a set of items, modern psychometric approaches use probabilistic models to measure the level of the construct underlying a set of items, such that the ordinal score on an item (better/worse, more/less, etc…) is transformed into a cardinal variable. The underlying construct is referred to as a latent trait. The difference between traditional and more modern approaches to measurement is highlighted by the example of two respondents with the same total score on a test but with different values on the latent trait underlying the set of items on the test. This occurs when there is a difference in the level of difficulty among the items that are answered correctly by the respondents. In modern psychometrics, the individual who answered more of the harder items is assumed to demonstrate a higher level of the latent trait. The total latent score is derived by estimating the difficulty level of each item based on the person responses and weighting the items accordingly on an interval scale, thereby legitimizing the summing of scores. The construction of this continuum reveals the differences in scores. If the data adhere to model expectations, it follows that the items of the test contribute to the formation of a single *unidimensional* construct; item parameters are constant across groups of respondents, i.e. they are unbiased; and pairs of item responses are independent. Inferences can then be made about the underlying construct being measured.

Modern psychometrics, or Latent Trait Theory (LTT), is therefore considered an advancement on the traditional paradigm, in that interpreting aggregate latent scores is considered superior to reporting raw scores, particularly when examining the relative standing of individuals or groups or when studying change over time (Bond and Lang, 2013; Embretson, 1996; Fraley et al., 2000; Kang and Waller, 2005). The additional advantage of latent trait theory lies in the possibility of linking overlapping tests either by including a set of common items across different tests, referred to as anchor or link items, or by repeating the same version of a test over time (for example, Looveer and Mulligan, 2009; Das and Zajonc, 2010).

The first paradigm within LTT is traditional item response theory (IRT), initially developed by Lord and Novick (1968) and Birnbaum (1968). Within this framework, emphasis is placed on accounting for measurement disturbances that may compromise the construction of a unidimensional scale. The requirement of a single dominant dimension is downplayed (see Goldstein, 1979, 2000, 2004; Dickson and Kohler, 1996). Rather, anomalies in the data are controlled for by the inclusion of model parameters and the choice of model is generally determined by how well it accounts for the data. The rationale underpinning this approach is argued by Traub (1983):

> … A priori considerations suggest strongly that items measuring educational achievement are unlikely to satisfy the assumption of unidimensionality. It must be granted, however, that the issue of unidimensionality may be in doubt for item sets related to narrowly defined achievement domains, and perhaps for other special situations. When this is the case, it will be necessary to seek empirical evidence as to whether or not the assumption of unidimensionality is tenable... If unidimensionality can be assumed, then the number of item parameters to include in the model becomes a matter of concern.
>
> (Traub, 1983: 63)

The two parameter IRT model (Birnbaum, 1968) estimates the probability of a person answering an item *x* correctly (e.g. *x* = 1) as a function of the difference between the person's ability level, β_v and the item's difficulty level, δ_i. The parameters β_v and δ_i are the locations of the persons and items on the scale continuum, expressed in logits. The parameter α_i is the discrimination parameter and takes a value of one in the 1-parameter logistic model. Discrimination, the steepness of the function/curve, is a measure of how well a given item distinguishes between respondents of different ability levels. Therefore, the inclusion of a discrimination parameter controls for the observed variation among test items in their power to discriminate. The three-parameter logistic model accounts for guessing on an item (Birnbaum, 1968). The guessing parameter, *c*, provides the lower bound for the probability of a correct answer on a given item for a respondent with 'no ability'. That is, it controls for the possibility of answering correctly without any trait knowledge.

The main criticism of IRT models is that while the addition of model parameters may improve the statistical description of the test data, it reduces the possibility of relating test scores directly to the underlying latent trait of study. Within an IRT framework, the choice of model: single, two- or three-parameter, is determined based on how well the data fits the model. However, it is argued that adding parameters to account for anomalies in the data may affect validity judgements. For example, it has been pointed out that the 2PL discrimination parameter is not used to model item characteristics but to provide a more detailed description of observed item responses (Ryan, 1983). Similarly, since guessing is more likely to be a characteristic of students rather than items (Ryan, 1983), the 3PL guessing parameter yields information about observed responses rather than the items, and thus cannot inform on the quality of the measures. In other words, including model parameters is akin to modelling observations in the data which are unrelated to the "fundamental properties of good measures" (Andrich, 1989: 14). Such parameters hide problems that should be addressed rather than controlled for and inadvertently undermine the validity of test score interpretations (Andrich, 2004).

Whereas traditional IRT seeks to fit the data to the model, Rasch Latent Trait theory approaches the problem the other way around in that emphasis is placed on adherence to the fundamental requirements of objective measurement (see Thurstone, 1931; Thurstone and Chave, 1929). Based on this foundation, the model requires that the data structure satisfies three criteria. The first is unidimensionality, the presence of a dominant dimension with the possible presence of other minor dimensions (Panayides et al., 2010).

The second is the requirement of local independence of item responses. Only a person's ability, and no other factor (i.e. experience or environment), should influence the pattern of responses (Marais and Andrich, 2008a, 2008b). Related to this, a person's response to one item should be unrelated to his/her response to any other item on the test, such that earlier items should not provide hints to subsequent ones, i.e. uncorrelated residuals (Marais and Andrich, 2008a, 2008b). The third requirement has to do with invariant comparisons. Person ability estimates should be invariant to the different subsamples of items and item difficulty estimates should be invariant to different subsamples of the population. As Linacre (1996) writes,

> Constant item parameters imply a constant construct. Different item parameters across samples of the relevant population imply that the construct has changed. Then measures cannot be compared across samples, and we are reduced to a vague notion of what we are measuring.
>
> (Linacre, 1996: 513)

Within the Rasch paradigm, the reason for misfitting data should not be addressed through parameterization. Data-model misfit needs to be explained qualitatively regardless of whether an item is retained, modified, or discarded. Even unusually high discriminating items should be inspected, contrary to the traditional approach in which over-discrimination is desired (Masters, 1988). Andrich elegantly summarises the distinction,

> When misfit … is identified, the main challenge in the traditional paradigm [.] to those with expertise in statistics or data analysis [is] to identify a model that accounts better for the given data, notwithstanding that they may find other problems in the data; the main challenge in the Rasch paradigm is for those with expertise in the substantive field of the construct to understand the statistical misfit as substantive anomalies and, if possible or necessary, to generate new data that better conform to the model while also enhancing substantive validity of the variable…
>
> (Andrich, 2004: 15)

Although conceptually different, in application the Rasch model for dichotomous responses takes the same functional form as the 1-parameter logistic IRT model. The probability of getting an answer correct is determined by person ability and item difficulty, and no other parameters. The discrimination parameter is fixed at a value of one for all items such that items vary only in their level of difficulty. Guessing is not a property of the Rasch model but considered an anomaly in the data that should addressed to improve the construct validity of the instrument. However, it is possible to investigate the extent of guessing as a measurement disturbance based on the premise that it is more likely to occur when a person has a low ability level in relation to the item's difficulty level such that more correct responses are generated than would have been the case if guessing had not been a factor (Andrich et al., 2012).

The main criticism of the Rasch model is that its properties are too restrictive to meaningfully interpret real world educational data (Traub, 1983; Goldstein, 1979). While traditionalists agree that simplest model is preferred, mainly for its ease of application and interpretation (Fischer and Molenaar, 1995), a less restrictive model "allows more scope for the operation of educational criteria" (Goldstein, 1979: 218). In other words, real world data does not always fit the assumptions of the Rasch model. Other scholars critique the model with more severity, arguing that it may not be possible to quantify cognitive abilities through latent trait models at all, in what is now referred to as the Rasch paradox (Mitchell, 2008).

Several concerns about the robustness Rasch and IRT scores have been raised in the literature that are relevant for the current research. Typically, little is said about the extent to which results reflect real outcomes or artefacts of the procedure (Brown et al., 2007). First, there is growing evidence that different scaling/weighting strategies, i.e. the process by which the scores are transformed, will produce different person/item estimates, affecting the relative rankings of individuals and the measure of change over time (Goldstein, 2004; Brown et al., 2007; Bond and Lang, 2013; Kreiner and Christensen, 2014; Gebhardt and Adams, 2007).

Bond and Lang (2013) show that the test score gap between African- and Caucasian Americans in the United States, is partly explained by sensitivity of the test scores to the scaling method. The authors use four different procedures to transform the raw scores of students between kindergarten and grade three. The results show that the gap at different time points varies from 0.10 to 0.30 standard deviations, depending on the transformation used (Bond and Lang, 2013). The authors also find that the gap between kindergarten and third grade ranges from zero to 0.6 standard deviations, and that some scales show a decrease in the gap over time (Bond and Lang, 2013). Consequently, the authors state that “researchers should be more circumspect in their use of test scores and other ordinal scales as dependent variables” (Bond and Lang, 2013: 1479).

Other methodological decisions that may undermine the robustness of test score results have been reported in the literature. Kreiner and Christensen (2014) showed that the country rankings obtained by the Programme for International Student Assessment (PISA) were sensitive to the use of different subscales of items, to alternative strategies for treating differential item functioning and to variations of the Rasch model (Kreiner and Christensen, 2014). Gebhardt and Adams (2007) showed that repeated measures trend data obtained by PISA between 2000 and 2003 were sensitive to the choice of linking items, test equating procedures and to different model alternatives for estimating mean trends (Gebhardt and Adams, 2007).

This important analytical work has led to a call for a more cautious approach to interpreting results from assessments and a more critical approach to assessment design, one that is rooted in substantive educational and psychological theory (Hopfenbeck, et al., 2017; Baird et al., 2017; Rutowski and Rutowski, 2016; McGrane, 2015a; McGrane 2015b). As noted by Stenner et al. (2013), “everything important to instrument justification cannot be subsumed under the validity banner. Validity is not constituted by a bucket of correlations” (Stenner et al., 2013: 11). A similar point has been made by Borsboom (2005) who writes: “… the problem of validity cannot be solved by psychometric techniques or models alone. On the contrary, it must be addressed by substantive theory” (Borsboom, 2005: 151). Thus, depending on the purpose of the assessment, some relationship between theory and test design is necessary, or put more compellingly:

> A rationalist might anticipate that theories of learning influence assessment theory and assessment design: at a minimum, learning theories shape our notions of what aspects of performance need to be included. Assessment design would both reflect and influence the outcomes of learning and therefore, theories of learning and assessment would ideally benefit from the information generated by each. Moreover, assessment theory and learning theory might also be expected to have reciprocal effects upon each other.
>
> (Baird et al., 2017: 317)

2.4.5 Summary

The first section 2.4.1 of this review explores the concepts of reliability, validity, and test bias in cross-cultural assessment research. The current study adopts a unified notion of validity, which suggests that validity needs to be argued based on evidence of the different aspects of validity rather than proven according to specified criteria. In other words, “validity does not only apply to the instrument … nor just to the numbers produced by these instruments but rather to the claims and decisions that are made by the users based on these numbers” (Stenner et al., 2013: 11). It follows that ensuring educational and psychological constructs are clearly defined and meaningfully operationalised strengthens the conclusions drawn based on the resulting test score data.

Section 2.4.2. addresses poor investment in test adaptation and documentation of adaptation processes in low-income countries. Assessment design does not receive the same level of caution and critique. Retaining test items that do not work as intended in a certain population, or overlooking items that might have worked well, weakens a test by producing data that does not provide a true picture of children’s performance. If the test data cannot accurately describe individual differences, it will not be possible to accurately explain or address them. Nag et al. (2017) cites the pioneering Young Lives International Longitudinal Study of Childhood Poverty (2002 – 2017) as an example of a good practice in test adaptation and documentation of the adaptation process. The study proposed methodologies for enhancing the cultural appropriateness of the adapted tests, with a focus on linguistic issues. Young Lives does not seek to draw inter-country comparisons of children's achievement. Such relative rankings are discouraged by the research team, who recommend that only comparisons of associations between achievement and contextual factors should be made (Cueto et al., 2009: 10; see Crookston et al., 2014 for an example of this approach). The researchers further argue that in the case of within-country comparisons, "it may be that different language groups require different combinations of items to establish their ability…" (Cueto et al., 2009: 10).

The issue of bias in cross-cultural assessment is discussed in sections 2.3.3. The main finding is that efforts to understand bias is largely driven by technical concerns, with less attention paid to substantive theories of learning and educational measurement. It was reported in section 2.2 of this literature review that studies rarely report on the psychometric quality of assessment instruments. Among the studies that do report psychometric results, none provided substantive justification for the choice of model, i.e. measurement vs. structural equivalence, Rasch vs. 2 or 3PL IRT models.

Theoretical advancements in psychometric research are discussed in section 2.3.4 with a view to set up an appropriate framework for the present study. The review also highlights the need for assessment to be guided by substantive theories of education and psychology. The first step in ensuring that a measure is appropriate for context involves establishing the purpose of the measure as a first step, then deciding what developmental or learning outcomes would need to be measured and what type of test would be most effective to achieve these goals. As noted by Fernald et al. (2009), deciding "why" to measure children's development, "what" to measure and "how" to measure it are crucial steps in the evaluation of interventions and programs targeting young children (Fernald et al., 2009: 22). Moreover, "every test should be carefully examined and piloted within the given cultural context, and with the collaboration of a local psychologist, to be sure that it is measuring what it intends to measure" (Fernald et al., 2009: x).

2.5 Research framework for the current study

This final section draws on the lessons learned from child development theory, measurement theory, and evidence on the effects of poverty on children's development to design a conceptual framework for the current study. The motivation driving the research is to better understand the pathways through which poverty shapes vocabulary development in Ethiopia, and how changes in children's living conditions influence developmental paths. There is an abundance of research documenting the devastating and persistent effects of poverty on children's development. However, much less is known about the dynamics of poverty and children's growth over time.

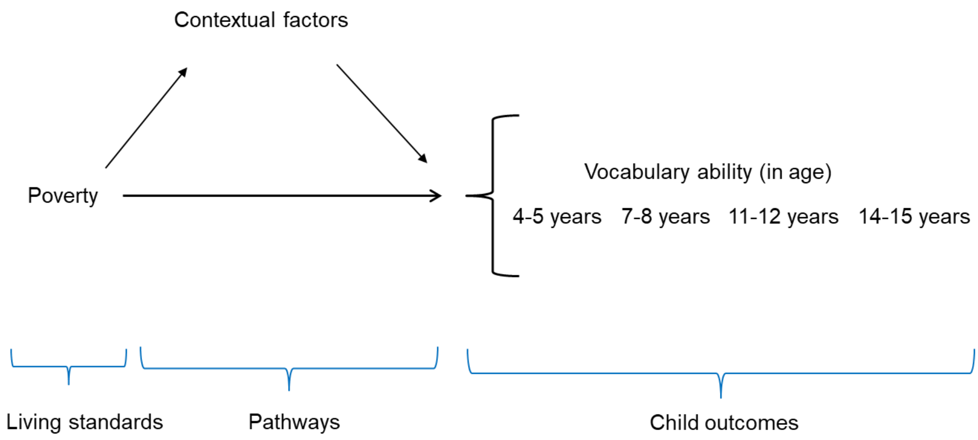
The researcher's perspective is that children's development cannot be understood without careful consideration of the context and culture in which children live their lives, and that this should be reflected in the measures and models used to assess development (i.e. psychologically meaningful constructs). This principle, discussed in the first section of the literature review, guides subsequent stages of the research. Accordingly, the measures used to assess both living standards and children's cognitive outcomes are not unquestioned but examined critically. A key message emerging from the review is that ensuring cultural sensitivity, and in turn, reliability and validity in a quantitative educational or developmental study is not about accounting for cultural factors, i.e. controlling for language, ethnicity, religion, etc… It is about developing contextually and culturally meaningful concepts, constructs, and operational measures. An important finding from the review is that the use of cognitive tests is not always based on substantive theories of child development. Moreover, test design, adaptation, and analytical psychometric processes are not always explicit, raising questions about the validity and reliability of the data produced. Thus, the third and final section of the literature review considers debates within educational and psychological assessment theory to identify a way forward for assessing the reliability and validity of the measure used in the current analysis.

Recognising the importance of cultural and linguistic influences on assessment outcomes (Berry, 1969; Hambleton and Patsula, 1998; Van de Vijver and Poortinga, 1998), careful consideration is paid to understanding how well the tests performed across linguistic subsamples. The importance of test adaptation is heightened when an instrument is used in a multicultural or multilingual setting, or when it is transferred from one context to another (Nag et al., 2017). Whereas for construct measurement a more restrictive Rasch model is generally preferred (e.g. for item banking, test construction and test validation), more flexible 2- and 3-parameter IRT models may be used for statistical modelling of test score data. As stated by Panayides et al. (2010),

> If the purpose is to construct a good measure then the items and the test should be constrained to the principles of measurement. If, on the other hand, the purpose is to model some test data then the model which fits the data best should be chosen.
>
> (Panayides et al., 2010: 7).

The current research has to do with both measurement and modelling. Thus, the more restrictive Rasch model is preferred due to its focus on item functioning and the foundational properties of measurement. The Rasch scores are then used to model cognitive developmental trajectories over time. The robustness of the derived scores is argued in terms of: (i) the relationship to external measures of achievement and conformity with theoretical expectations; (ii) the stability of item calibrations over time; (iii) how well the test works across language groups; and (iv) on the basis of statistical data-model fit.

Figure 2.7 Framework for analysing the effects of poverty on vocabulary trajectories



The conceptual model for the analysis is depicted in Figure 2.7. It is expected that poverty will impact on cognitive development differently at different stages of childhood. In early childhood, which is the most critical period for brain development, being poor is expected to have a profound effect on development. Poverty status in later childhood is expected to be less predictive of vocabulary outcomes, due to the increasing importance of other factors such as motivation to learn/read, classroom lectures and discussions, peer group and other social factors. However, inequalities formed early on in life are expected to persist over the long term. What is uncertain is how the initial gap evolves over time. Do developmental inequalities increase, narrow, or remain constant with age? Do changes in children’s living conditions influence the shape of their developmental trajectories? Moreover, what are the specific pathways underlying this relationship, i.e. the contextual factors that mediate the relationship between poverty and children’s vocabulary development?

Chapter 3 Methodology

3.1 The research questions

The study seeks to address the following overarching research question: how does poverty shape children's vocabulary development in Ethiopia? Three sets of supplementary descriptive and explanatory questions are taken up as follows:

3.1.1 Measurement of children's cognitive development

1. To what extent is the Peabody Picture Vocabulary Test (PPVT) a valid and reliable measure of vocabulary ability among the sample of children?
2. How consistently does the PPVT perform across children from different linguistic backgrounds and across time?
3. To what extent can growth trajectories be reliably estimated, considering the cultural differences across the sample vis-à-vis language?

3.1.2 Wealth effects on children's cognitive development

1. Based on the constructed metric of children's abilities, to what extent do individual developmental trajectories differ in relation to the overall mean trend, in terms of initial status, the growth rate and the shape of the trajectory?
2. What is the effect of wealth on vocabulary developmental paths, i.e. how does wealth status in early childhood predict initial ability and the growth rate? How do patterns of development differ between the top and bottom wealth quintiles?
3. To what extent is change over time in wealth associated with patterns of development (i.e. change in ability)?

3.1.3 Mediated wealth effects on children's cognitive development

1. How is the relationship between wealth and children's vocabulary development is explained by the children's nutritional status, their home environment and preschool exposure?

To address these questions, the current research draws on different strands of literature, discussed in Chapter 2, to identify common threads that link research on the developmental effects of poverty and environment. The study builds on relevant themes emerging within different academic debates to establish a rationale, theoretical foundation, and methodological approach for addressing the research questions. To illustrate this multidisciplinary approach, the box below summarises the literature review search strategy criteria, including the keywords used to search for academic papers. An excerpt of the literature review mapping exercises is included in Appendix 1.

Box. Literature review search strategy criteria used for the current study

The following databases were searched for studies, systematic reviews and meta-analyses that were reported in English from 1990 to October, 2015: 3ie, Cochrane Library, ERIC, PsycINFO , Web of Science, SCOPUS and PubMed along with published documents from the World Bank, UNESCO and UNICEF. References cited in retrieved papers were further examined.

Two searches were conducted. Keywords that were used for search 1 are: “developing countries” or “low income countries” or “developing nations” or “third world” **and** “cognitive development” or “language development” or “vocabulary development” or “socio-emotional development” or “psychosocial development” or “school readiness” or “early childhood” or “preschool” or “schooling” or “school enrolment” or “school attendance” or “school dropout” or “grade retention” or “grade progression” or “grade attainment” or “grade completion” or “educational achievement” or “educational attainment” or “educational progress” or “numeracy” or “literacy” or “non-cognitive skills” or “soft skills”. For search 2, search 1 keywords are used in addition to “poverty” or “income” or “economic status” or “poor” or “deprived” or “disadvantaged”.

To identify important studies, i.e. latest research, missed through the database searches, hand searches were conducted using Google Scholar and bibliographies of studies reviewed.

3.2 Addressing the research questions

3.2.1 The Young Lives International Longitudinal Survey

The research uses survey data from the Young Lives longitudinal study of children and youth in Ethiopia. The data structure includes repeated measures of children’s developmental and educational outcomes together with extensive information on the children’s local environments. The survey can therefore be used to better understand relationships between patterns of change in children’s development and learning in relation to their wider contexts and experiences. The Young Lives data have been collected to investigate topics related to the antecedents and outcomes of childhood poverty. Data collection began in 2002 in four study countries – Ethiopia, India (Andhra Pradesh), Peru and Vietnam. In each country, the study follows 3,000 index children over a fifteen-year period (2002-2017).

Thus, the current study uses secondary data analysis techniques, where secondary data analysis refers to the analysis of data “that has already been gathered or compiled in some way” (Dale et al., 1988: 3), or “for purposes that were not envisaged by those responsible for the data collection” (Bryman, 2008: 296). While there are many benefits to using secondary data, particularly large-scale data sets, the limitations need to be identified. Having not been involved in the Young Lives data collection process, the final set of research questions were designed to take into account the strengths and weaknesses of the data. First, efforts were made to understand the context in which the data were collected – including the background, aims and research questions that underpin the Young Lives study, the theoretical framework informing instrument design, and how the data were piloted, recorded and coded. This was achieved by engaging in background research and by working as a temporary researcher within the Young Lives team. Secondly, substantial time has been spent assessing data quality vis-à-vis the assessment instruments, child, and household questionnaires. Whilst the availability of vast amounts of data permits the examination of a wide range of relationships, the data are not a perfect match for the current research questions, with some key measures missing and others only partially assessed. Finally, as with all longitudinal surveys, there is the often the problem of participant non-response, discussed in the next section.

3.2.2 Sampling, attrition, and missing data

The target population in Ethiopia includes two age cohorts of children, one born in 1994/1995 and the other in 2001/2002. The current study focuses on the younger cohort of 2,000 children who are followed from one to 15 years of age, or more precisely: 0.5/1.5

years of age through to 15.5/16.5 years of age. Five rounds of survey data have been collected covering different aspects of the children's lives, their households and communities, and thus developmental and contextual data are available at ages one, five, eight, 12 and 15 years. The Young Lives survey is not a representative sample survey. The small sample size within each cohort makes it difficult to examine characteristics of groups who make up a small proportion of the total population. As noted by Young Lives, the survey data allows for the "detection of moderate-sized differences… between subgroups when the subgroups are at least 20 per cent of the overall sample" (Young Lives, 2002: 2).

The Young Lives sampling strategy follows a sentinel site surveillance methodology, involving a multi-stage sampling procedure (Outes-Leon and Sanchez, 2008). First, sites were selected based on pre-defined criteria. For diversity and national coverage, five out of nine regions were selected accounting for 96 per cent of the population. Then within each region, three to five districts (*woredas*) were chosen, ensuring variation in poverty status and rural-urban representation. Insecure zones and remote areas were excluded to maximise security and tracking. Within each district a sentinel site was chosen corresponding to the lowest level administrative unit (a peasant association in rural areas and a *kebele* in urban areas). Finally, within each sentinel site a village was randomly selected and all households within the village and surrounding area were interviewed until 100 households with a 1-year old and 50 households with an 8-year old were located.

Young Lives highlights that attrition rates in the sample are low "in absolute terms… [and] when compared with attrition rates for other longitudinal studies in less developed countries" (Outes-Leon and Dercon, 2008: 4). Sample attrition, summarised in Table 3.1 has been minimised by tracing all children even if they changed location (Outes-Leon and Dercon, 2008). Refusal rates have been kept low by retaining the same data collectors between rounds, where possible, to build stable relationships with the children and their families (Outes-Leon and Dercon, 2008).

Table 3.1 Sample attrition among the younger cohort children between R1-R5

Attrition	n	%
Deaths	85	4.3
Refusals	10	0.5
Untraceable	14	0.7
Living abroad	6	0.3
Total	115	5.8

Source: Young Lives, 2018: 3.

Understanding the pattern of attrition is important as non-random attrition can lead to biased inference and undermine research. Outes-Leon and Dercon (2008) found that among the younger cohort children, those who died between Rounds 1 and 2 had very poor nutrition, as measured by their height-for-age and weight-for-height z-scores, whereas untraceable children had better anthropometric z-scores than non-attriting children (Outes-Leon and Dercon, 2008). Survey non-response refers to non-sampling error that happens when sampled individuals "refuse to cooperate, cannot be contacted, or for some reason cannot supply the required data" (Bryman, 2008: 696). It covers both data that are missing due to the loss of participants over the sampling period and data that are missing for particular study variables. Initially, the extent of missing data points in the Young Lives pooled sample was computed for the core study variables: PPVT test scores and household wealth, which revealed very few missing data points. However, missingness among potential mediator variables ranged from very low to over 50%. Examining the pattern of missingness among these analytical variables will be important for minimising biased statistical inference (Little et al., 1988).

3.2.3 Data collection

The current study draws on the main household survey for information. Table 3.2 presents an overview of the data collection framework for Ethiopia. The research protocols involve both quantitative and qualitative tools. Three are a total of five rounds of household survey data. Data on children's reasoning, vocabulary, literacy and mathematics are available for all children in the sample. Contextual data are also available across a range of topics outlined in Table 3.3, including the home environment, children's socioeconomic contexts, health, nutrition, and education.

Table 3.2 Young Lives data collection framework in Ethiopia (younger cohort)

Household Quantitative Survey				
Round 1	Round 2	Round 3	Round 4	Round 5
2002	2006	2009	2013	2017
6-18 mo.	4.5-5.5 yrs.	7.5-8.5 yrs.	11.5-12.5 yrs.	14-15 yrs.
Qualitative Data Collection				
Round 1	Round 2	Round 3	Round 4	
2007	2008	2011	2014	
5.5 to 6.5 yrs.	6.5 to 7.5 yrs.	9.5 to 10.5 yrs.	12.5 to 13.5 yrs.	
Quantitative School Survey				
Round 1	Round 2			
2009-2010	2012-2013			
8-9 yrs.	Grades 4 & 5			

Table 3.3 Content of the child, household and community questionnaires

Household	Child	Community
Household composition Caregiver background Living standards Household livelihoods Economic and social events Child health, nutrition, education Caregiver mental health Pregnancy and delivery Breastfeeding Early childhood care	Measures of cognitive and learning achievement Perceptions of wellbeing Social networks Schooling and work Violence Reproductive health Mental health	Infrastructure Services Water access Sanitation services Waste disposal Economy Local prices Social protection

Although not used in the analysis, the research also involves qualitative case studies tracking a sub-sample of 50 children. These data are intended to be used to investigate in-depth changes in the children's lives from the perspective of the children themselves (Young Lives, 2014a). The project also includes two rounds of school survey data. The school survey addresses questions related to understanding the relationship between poverty and children's educational outcomes, as well as the role of the school in reinforcing or compensating for disadvantage (Young Lives, 2014b). The first survey (2009-2010) was very broad, designed to capture a wide range of detailed information about the children's schooling. It covered a sub-sample of 952 younger cohort children and 692 older children. Data collection took place at the pupil, class, teacher, and school levels. However, educational achievement data was not collected and only the Young Lives index children were selected. Class peers were not part of the sample which restricts the type of analysis that could be performed. The second school survey (2012-2014) covered approximately 11,000 children in grades four and five, including class peers. The second survey adopted a school effectiveness approach in that curriculum-based literacy and maths achievement data were collected at the start and the end of the academic year, and school environment items were more focused on school inputs, processes, and delivery.

The research protocols for the quantitative household surveys were developed through a process that involved working with local teams of researchers to adapt the instruments to the national context (Morrow, 2009). Each measure was designed to maintain inter-country comparability, while at the same time ensuring appropriateness at the country level. The quantitative questionnaires were implemented through structured interview-based techniques to minimise the risk of misunderstanding, efficient collection of such extensive data and to ensure standardisation. To explore children's perceptions and attitudes towards more sensitive topics, self-administered questionnaires were also given to the older cohort children and the younger cohort children in later rounds.

3.3 Research design: psychometric analysis

3.3.1 Background to the PPVT instrument

To measure vocabulary ability among the sample of children, at each round of data collection the Peabody Picture Vocabulary Test (PPVT-III, Dunn et al., 1997) was administered in the children's choice of language. The test was given to the younger cohort children at ages five, eight, 12 and 15 years. For the first administration, language was classified into four categories: Amharic, Oromifa, Tigrinya and Other. For subsequent administrations, the "Other" category was disaggregated into component languages, eight in total. Table 3.4 presents the frequency of responses by testing language at each administration. Nearly half of the children chose to respond in Amharic, a result of the sampling strategy in that Amharic is the dominant language used in the capital, Addis Ababa, and in the Amhara region. For 8% of 12-year olds, PPVT scores are reported but the test-taking language is not, potentially limiting the sample size available for the analysis. The test-taking language is unreported for 14.3% of 15-year olds.

Table 3.4 Language used by the child during PPVT administration, younger cohort

	5 years		8 years		12 years		15 years	
Language	N	(%)	N	(%)	N	(%)	N	(%)
Amharic	811	43.4	768	41.1	913	49.0	902	47.8
Oromifa	289	15.5	352	18.8	336	18.0	342	18.1
Tigrinya	371	19.9	377	20.2	377	20.2	373	19.8
Other	351	18.8	366	19.6	89	4.8	0	0
Missing	46	2.4	5	0.3	150	8.0	270	14.3
Total	1912	100.0	1868	100.0	1865	100.0	1887	100.0

The PPVT is a widely used measure of receptive vocabulary achievement. It is standardised in English and has been adapted to many other languages. The test consists of 17 sets of items, with the level of item difficulty increasing by set. It is individually and orally administered, untimed and is designed to be used from 2.5 years throughout adulthood. For each item, the respondent is asked to choose one of four pictures that best describes the word spoken by the administrator. If a set of items is completed successfully, the next set is offered until the child reaches a ceiling (eight or more errors in a set). Each child receives a total score based on the number of correctly answered items. Raw scores are then computed by subtracting the total number of errors made from the ceiling item. In the Young Lives study, examiners were required to choose an appropriate start item and starting set for the first two administrations according to the child’s chronological age. The starting set for a child is found when a child makes no more than one error in a set of 12 items (Dunn et al., 1997). In Young Lives, if the child received more than one error in the starting set then the examiner was required to drop back to the previous set beginning with the first item in that set, which continued until the start set was established (Young Lives, 2009). The ceiling set is found when eight or more errors are made in a set of 12 items (Dunn et al., 1997).

The PPVT-III was standardised on a nationally representative English-speaking US sample and has not been norm-referenced for the Ethiopian context. In the Young Lives study, the PPVT is used as a proxy measure for cognitive ability, since the test has been shown to correlate strongly with varied measures of intelligence in other samples (Dunn and Dunn, 1997; Campbell, 1998; and Campbell et al., 2001; Gray et al.,1999). The use of vocabulary assessments as indicators of generalised intelligence (e.g. the NICHD Early Childhood Care Research Network, 2005) is based on the Aptitude hypothesis, which suggests,

> …students who score higher on vocabulary tests have learned more word meanings from exposure to oral and written language because they have higher levels of generalised verbal aptitude. This aptitude, in turn, contributes to their development of reding comprehension skills, independently of knowledge of specific word meanings.
>
> (Keiffer and Stahl, 2015, p. 126)

However, Nagy (2007) hypothesises that individual differences in vocabulary are primarily influenced by differences in metalinguistic awareness such as phonemic and morphological awareness, and sensitivity to context (Nagy, 2007), abilities that are not inherited and can be learned. Moreover, research has shown that the concurrent validity of the PPVT with other language measures is not high (Dockrell and Marshall, 2014).[6] Comparisons between the performance of the Young Lives PPVT measure with other measures of intelligence or language have not been made. However, correlations between the test scores and home background characteristics have been shown to support theoretical expectations (Cueto et al, 2012; Leon and Collahua, 2015).

[6] Findings relate to the British Picture Vocabulary Scale III (BPVS-III), which is based on the PPVT, and is norm-referenced for the UK population.

Concerns have been raised over the validity of the PPVT when used with very young children and in multicultural contexts. Since vocabulary is related to individual differences in literacy attainment, the measure is commonly used with young children as a screening instrument for school readiness (Chien et al., 2010) and achievement (Williams and Wang, 1997). The use of the PPVT also dominates vocabulary assessments in developing countries, well over half of those identified in a systematic review (Nag et al., 2017). However, it has been argued that the highly representational black-and-white line drawings may be inappropriate for pre-school aged children as they require a higher level of cognitive skill than other representational forms such as objects and coloured pictures (Kierscht and Vietze, 1975). Experts have argued that poor performance among sub-groups of children may be due to cultural differences in exposure or experience rather than real differences in ability (Stockman, 2000; Ukrainetz, 2000 and 2002; Champion et al., 2003). For example, while the English version of the PPVT was developed on the basis of a nationally representative US sample including a large ethnic minority norm group, Liang and Kamhi (2003) showed that for African-American children, poorer performance was not only due to lower receptive vocabulary ability but also “a lack of familiarity with the interaction pattern of pointing to pictures…” (Liang and Kamhi, 2003: 45). Furthermore, the selection of the PPVT pictures themselves may be poorly contextualised and may not capture the complexity of vocabulary knowledge. As documented by Nag et al. (2017):

> The first challenge is that pictures may be ambiguous and the reason for this could be several: pictorial representations may be alien for children with exceptionally low print experience, the pictorial idiom may be difficult for children to understand or the visualisation may be outside their lived experience. A further challenge is to create picture items that appear equivalent, perform equally across subgroups, and do not elicit word knowledge around unintended constructs. Finally, picture-based assessments are constrained by the medium; those words that are not easily captured in pictures need to be set aside.
>
> (Nag et al., 2017: 29)

In the Young Lives survey, the PPVT is the only common assessment across rounds of the survey. It is not curriculum benchmarked, and meaningful for both those who attend school and those who do not. Other educational measures are available. Table 3.4 presents a list of the assessment instruments used in the survey, excluding the single-item measures. At each round, simple tests were carried out to assess basic reading and writing ability. The items involved reading a sentence from a card and writing a sentence spoken out loud. Reading levels were assessed on a 4-point scale: not able to read, able to read letters, words, or sentences. Writing ability was measured on a 3-point scale: not able to write, able to write even with errors and able to write without difficulty. The data structure also includes a repeated measure of basic numeracy based on a single arithmetic calculation, at each round. The younger cohort children also completed a 15-item Cognitive Developmental Assessment (CDA) at 5 years. The Early Grade Reading Assessment (EGRA) was included at eight years, followed by more substantive literacy tests at 12 and 15 years. To assess mathematics, the children were given a basic maths test at eight years and more comprehensive tests at 12 and 15 years

Table 3.5 Child developmental assessment instruments

Round	Age	Cognitive	Literacy	Mathematics
Round 1	1 year	-	-	-
Round 2	5 years	PPVT CDA quantity subscale	-	-
Round 3	8 years	PPVT	Early Grade Reading Assessment (EGRA)	30 maths items
Round 4	12 years	PPVT	24-item literacy test	28 maths items
Round 5	15 years	PPVT	27-item English test	30 math items

3.3.2 Localisation of the PPVT instrument in Ethiopia

The Young Lives project has been cited as an excellent example of test localisation, demonstrating sensitivity to establishing fairness in test design, administration, and results interpretation (Nag et al., 2017). To use the PPVT in Ethiopia, several changes were made to the instruments. For the first administration, a national panel of experts was set up to select and translate items from Forms A and B of the PPVT-III (Cueto and Leon, 2012). The panel tried to ensure that the items were of increasing difficulty and that there were enough difficult items to adequately assess higher ability children in later rounds (Cueto and Leon, 2012). Administration procedures were standardized, documented in the fieldworker manual and examiners were trained accordingly. Examiners had to record whether participants experienced acceptable testing conditions. If not, their responses were recorded as invalid (Cueto et al., 2009).

A ceiling effect was observed in the test data for the older cohort children at 12 years, suggesting that the problem may arise among the younger cohort sample.[7] Many of the items did not function consistently across gender and language subgroups (Cueto et al., 2009). Excluded from the total score computation were those items showing low item-test correlation; an out of range in-fit statistic (i.e. items could not be answered correctly by individuals for whom they were targeted); and/or items showing differential item functioning for gender (Cueto et al., 2009). Items showing DIF for language were retained, and thus Young Lives recommends that international comparisons of test outcomes should not be made. For intra-country comparisons, “it may be that different language groups require different combinations of items to establish their ability…” (Cueto et al., 2009: 10).

For the second administration, test translation was revisited. All languages spoken by the child were identified. Two translators fluent in both English and the target language were required to re-translate all items independently. The translators then compared translations and agreed on a final version to be used in the pilot (Cueto and Leon, 2012). The test was later revised based on the pilot and the findings of the psychometric analysis (Cueto and Leon, 2012). For total score computation, misfitting items were defined as those with a low item-test correlation statistic or an out of range in-fit statistic (Cueto and Leon, 2012). Items showing DIF for gender were retained. Based on experience from previous rounds, the older cohort was not given the test in Round 4 (at 19 years), and

[7] Item targeting is investigated in Chapter 4.

roughly two-thirds of the scale items, those which functioned least well across the whole sample, were discarded.

3.3.3 Test performance among the older cohort and implications for the analysis

As part of a preliminary study for the D.Phil project, the performance of the full set of 204 items was analysed for older cohort children at age 12 years and 15 and the younger cohort children at ages five and eight years. A large majority of the items targeted an easy to moderate difficulty level. The finding suggested that future rounds of the test would not adequately measure vocabulary ability among the older cohort children unless additional harder items were included. The finding also implies that for the current analysis which focuses on the younger cohort children, the test may not perform as expected in later rounds of the survey. Response dependency was also found to be an important reliability and validity issue, with a large proportion of items showing very high residual correlations. There are ways of handling deviating items, including removing them. However, eliminating items without substantively understanding the reason for their misfit would further undermine scale validity (Andrich, 2004). In the Young Lives study the full scale of 204 items was reduced to 55. The psychometric analysis of the reduced set of 55 items is presented in Chapter 4. The technical outputs and key findings of the older cohort analysis (the full 204 items) are presented in Appendix 2.

Further investigation of the raw test data is therefore needed to better understand the validity of the construct in light of the observed ceiling effect among the older cohort sample and the removal of so many items from the test. The first aim is to identify any inconsistencies in the performance of the PPVT items across language groups. The second aim is to identify any inconsistencies in the performance of the test across rounds with a view to construct a reliable metric of scores for the growth trajectory analysis.

It is possible to estimate a single growth trajectory for each child by calibrating the scores from the repeated test onto a single metric using Rasch/IRT. The procedure for constructing this metric involves first analysing the psychometric properties of the test at each time point separately, to observe how well the test functions across language groups. The next step is to establish the robustness of the measure over time by identifying unstable items in the response structure. The final step is to look at the validity of the resulting Rasch scores by correlating them with the raw scores and other measures of achievement available in the data set. The data can then be pooled across time points or link items can be used to calibrate the repeated test onto a single metric.

3.3.4 Assessing the reliability and validity of the PPVT data

In reiteration, the rationale underlying the analysis of test performance relates to the first set of research questions on the PPVT's psychometric properties, examining how the test functions at each time point, across cultural/linguistic groups and over time. The section describes the process of validating the test and generating a set of comparable test scores across rounds. The study attempts to contribute to the emerging literature on the robustness of test scores in cross-cultural educational research. Addressing cultural and linguistic biases will be important for reliable estimation of the children's growth trajectories. This component of the research is limited to statistical analysis of overall test performance and individual item functioning. The lack of qualitative inputs on potential cultural sources of error limits the scope of this investigation. There is limited information about the test administration process and about how the vocabulary construct has been operationalised across cultures. Translated words and child/translator descriptions of the meaning underlying the items have not been recorded (as done in Haitana et al., 2010). The picture stimuli used during testing are not included in the test booklet or administrative documents.

Extending the analysis to investigate the stability of the test over time will help to determine whether the scores from the repeated test can be reliably calibrated onto a single metric. The Rasch model can be used to assess person change using repeated measures data by treating persons at different time points as distinct, known as stacking or pooling the data. The measure of person change would then be reflected in the change in a person's location on the Rasch metric. However, the accuracy of the change measure is likely to be affected by response dependency which occurs when a person's response to one item governs their response to a subsequent one, obscuring or amplifying the change measure (Smith, 2005: Marais, 2009). Wright (1996) further points out that item functioning may not remain constant over time even when identical data collection protocols are used on the same sample of respondents (Wright, 1996). Building on this literature, to analyse the stability of the Rasch scores over time, person-time dependencies are investigated, and item calibrations obtained at each time point are compared. An alternative approach to stacking the data would be to identify the most stable items across rounds and to use them as anchor items for linking the tests. In this approach, the item difficulties are fixed on the items that produce the most consistent difficulty estimates across time. Anchoring calibrations on the most stable items is thought to produce more reliable person ability estimates and a more accurate description of person change (Wright, 1996). The method is most applied when separate tests are used, each containing a set of overlapping items to link the tests so as to place the scores obtained from the separate tests onto a comparable scale.

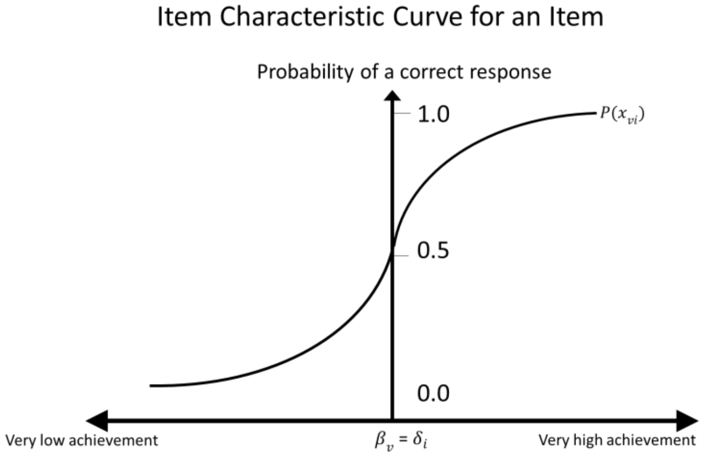
The psychometric properties of the PPVT data are assessed according to the Rasch measurement model for dichotomous items, also known as the rating scale model. The model ranks items according to their difficulty and ranks persons according their ability. The difficulty of each item is estimated based on the person responses. Person ability scores are obtained by estimating the probability of a person answering an item correctly, which is a function of their ability level and the item's difficulty level. Item difficulties and person abilities can then be placed on the same scale continuum. The more restrictive Rasch framework is considered better suited for establishing the cross-cultural validity of the test compared with alternative paradigms due to its emphasis on item functioning and adherence to the fundamental properties of measurement. Multiple choice item responses have been recoded from a polytomous into a dichotomous response structure so that for each item there are only two response categories, incorrect ("x_{ni} = 0") and correct ("x_{ni} = 1").

The Rasch model for dichotomous responses is given by Equation 3.1. The model estimates the probability of a person answering an item *x* correctly (e.g. *x* = 1) as a function of the difference between the person's ability level, β_v and the item's difficulty level, δ_i. The parameters β_v and δ_i are the locations of the persons and items on the scale continuum, expressed in logits. The parameter α_i is the discrimination parameter and takes a value of one in the Rasch model.

$$P\ \{x_{vi} = 1|\beta_v, \delta_i\} = \frac{e^{(\beta_v - \delta_i)}}{1+e^{(\beta_v - \delta_i)}} \qquad \dots 3.1$$

The logistic probability function is known as the item characteristic curve (ICC). This theoretical curve has an "S" shape (Figure 3.1). Person and item parameters are located on the same scale. The item locations have an arbitrary mean of 0 which is imposed by the model. The mean of the person locations is empirically derived. Each item is located at the point on the curve where a person's ability equals the item's level of difficulty, or when the probability of success on the item is 0.50. If the person's ability is above the item's difficulty level, the probability of that person getting the item correct is greater than 0.50. If the person's ability is below the item's difficulty, the probability of a correct response would be less than 0.50.

Figure 3.1 The Rasch model (simple logistic probability function)



The dataset involves a large proportion of structurally missing data (i.e. “missing by design”) clustered at the start and at the end of the test, relating to items that were not administered as a result of being located beyond the basal/ceiling range. Ludlow and O’Leary (1999) have shown that missing data due to the test taker, i.e. items that have been skipped by respondents or not reached in time, are legitimate responses and should be factored into the estimation of person ability and item difficulty (Ludlow and O’Leary, 1999: 626). These are items that were not administered or reached as a result of being too hard for the student. In contrast, item responses that are missing by design, i.e. the respondent was not expected to give an answer as in the case for adaptive testing, will have minimal impact on the parameter estimates and can be ignored when calibrating the items (Wu and Adams, 2002; OECD, 2005). However, a main feature of the PPVT is the assumption that for a given child, any item below the basal set would have been correctly answered, and any item beyond the ceiling set would have been answered incorrectly. Therefore, for the current data, items appearing below the basal set can be replaced with 1s, and items above the ceiling can be replaced with 0s (Leon et al., 2015; Embretson and Harshberger, 1999).

The RUMM 2030 statistical software is used to perform the test of fit analysis (Andrich, Sheridan and Luo, 2014). Stata 2014 is used to recode systematically missing responses and to remove cases with inconsistent language choice (StataCorp, 2015).[8] Based on the findings, decisions are made regarding which scale items are least effective in assessing cognitive development. The results are organised as follows:

- Overall fit of the items to the model;
- Item targeting;
- Order and location of the items;
- Differential item functioning by language;
- Individual item fit;
- Response dependency and multidimensionality;
- Scale stability;
- Score validation and transformation.

3.3.4.1 Overall fit

RUMM provides four statistics which give a general overview of the fit of the data to the Rasch model. The *mean (M) and standard deviation (SD)* of the item and person fit residuals provide some preliminary information. If the data fit the model well, the residuals should have a mean that is close to zero and a SD close to one. The *Person Separation Index*, r_β, is an index of reliability, similar to an internal consistency measure in traditional test theory. The difference is that in the Rasch model, reliability is calculated using person ability estimates and is interpreted in terms of variation in person locations. Internal consistency measures, such as the coefficient alpha or the KR-20, are interpreted in terms of variation in total scores. High person separation indicates that persons spread well across the continuum of items, and that the test reliably distinguishes between test-takers (Andrich, 1982: 104). The *item-trait interaction* is an approximate chi-square statistic which show whether the expected and observed scores are significantly different overall. Recalling from literature review, in Latent Trait Theory, the underlying construct being measured by a set of items is referred to as a latent trait. The item-trait interaction

[8] In 33 cases, the testing language was coded as missing. The way in which these missing data points were handled is discussed in Chapter 4.

should be interpreted cautiously with large samples which can produce very high expected-observed score deviation.

3.3.4.2 Item targeting

The person-item location distribution plots person and item locations on the same continuum. The distribution shows whether there are sufficient items for the range of student proficiencies. A test is well targeted if the average of the item difficulties is about the same as the average of the students' abilities and the item difficulties are evenly spread across the ability distribution (OECD, 2012: 222). Floor and ceiling effects can also be identified here.

3.3.4.3 Order and location of items

The item map presents the rank ordering of items in terms of difficulty level and illustrates whether this ordering is consistent with the ordering of the items on the test, providing additional evidence of construct validity. For a well-constructed test, test item ordering should be increasing in difficulty in that easier items should be asked earlier in the test and harder items later on. If items are not in the appropriate rank order, the scale may not be working as expected.

3.3.4.4 Differential Item Functioning (DIF)

Differential Item Functioning (DIF) is a form of item bias which occurs when different response patterns emerge for different subgroups of respondents with the same level of proficiency. It relates to the measurement property of invariance, which requires items to be unbiased and work in the same way across different classifications of individuals. If DIF is identified in the data, the scale is not invariant/unbiased across groups and the items in the scale should not be treated as the same for different subgroups. DIF may be uniform such that groups show a consistent difference in their response pattern across the trait, e.g. the slopes of the Item Characteristic Curves (ICCs) are parallel. It may be non-uniform such that the difference between groups varies across the trait, e.g. the slopes of the ICCs are non-parallel. DIF can be detected graphically by examining the ICC and confirmed statistically using ANOVA. Uniform DIF is present when there is a significant main effect for the group factor. Non-uniform DIF is indicated by a significant interaction effect between the group factor and the class interval. When detected, items showing DIF may be excluded from score computation when establishing relative standings, such as country rankings (Kirsch et al., 2002). Alternatively, the approach taken by PISA 2003 is that an item demonstrating "sound characteristics in each country but that shows substantial item-by-country interactions may be regarded as a different item (for scaling purposes) in each country…" (Adams, Wu and Cartensen, 2007: 274). Such items may be split in sequential order beginning with the item having the largest Mean Square (Andrich and Hagquist, 2012). The resulting data structure would include a new item for each group, with missing data cells for non-responses. However, if DIF is widespread then group comparisons should not be made, and separate scales would need to be constructed for each subgroup.

3.3.4.5 Statistical fit of the items to the model

RUMM provides two residual-based tests of item fit, which are intended to be interpreted relatively. The *standardized item-fit residual*, z_{gi}, given by Equation 3.2, compares the observed number of correct responses with the expected number of correct responses for each class interval (CI). Class interval refers to a group of individuals with similar scores. The number of persons in each CI should be roughly equal across groups. T_{gi} is the observed number of people who answered item *i* correctly in class interval g. $N_g P_{gi}$ is the expected number of people who answered correctly. N_g is the number of people in each class interval, and P_{gi} is the probability of getting item *i* correct in each class interval. The difference between observed and theoretical correct responses is divided by the standard deviation of the number who are likely to get the item correct, σ_{gi}. The values of z_{gi} should fall in the range of ±2.5. When the residuals are squared and added up this gives an approximate chi-square distribution on G-1 class intervals for the item as a whole.

This second statistic is the *item-trait interaction test of fit*. It tests whether there is a significant difference between the expected and the observed mean scores.

$$z_{gi} = \frac{T_{gi} - N_g P_{gi}}{\sigma_{gi}} \qquad \ldots 3.2$$

For each of the flagged items, the *Item Characteristic Curves* (ICCs) are then examined to identify the source of misfit. The ICC plots the observed mean scores of each class interval against the predicted model proportions, providing a visual display of how well the data for the item fit the model. The Bonferroni adjustment is used to identify the most misfitting items. The correction reduces the risk of a Type I error of finding a significant difference when there is none (Bland and Altman, 1995). A distinction is made between overfitting items (i.e. observed means are steeper than the theoretical curve) from those which underfit (i.e. flatter than the theoretical curve). High discrimination may indicate response dependence and poor discrimination may suggest multidimensionality (Marais and Andrich, 2008b).

3.3.4.6 Detecting violations of local independence

Once item fit and DIF have been resolved, a Principal Component Analysis (PCA) of the item residuals is carried out to detect violations of local independence and multidimensionality. For a measure to be robust, the person ability parameter should fully account for the variation in item responses, and there should be no further relationship between responses once the "Rasch factor" is extracted (Marais and Andrich, 2008a: 201). There is substantial evidence that Rasch/IRT scores are not robust to violations of this property (see Monseur et al., 2011).

The property of local independence involves two sets of goals, which are frequently taken for granted in the literature. The first has to do with dimensionality, items should contribute to the formation of a unidimensional construct. If some other person parameter influences the pattern of responses then the data are not unidimensional and reliability may be decreased (Marais and Andrich, 2008a). It should be noted that from a theoretical Rasch perspective, unidimensionality is regarded as a continuum, and a researcher must determine the point at which multidimensionality threatens the interpretation of item and person estimates given the aim of the research (Smith, 2002). If it is asserted that the assumption does not hold, it can be concluded that both the reliability of the scale and the variance in person estimates is reduced relative to the case for a unidimensional scale (Marais and Andrich, 2008a). A principal component summary of the residuals will provide some clues (Smith, 2002). The pattern of the residuals may show that the factor loadings on the first principal component load in opposite directions (e.g. a set of positive and a set of negative items with high factor loadings), suggesting the presence of a second dimension. This can be confirmed by a paired t-test to determine whether the person ability estimates obtained from the two subsets significantly differ from each other (see Pallant and Tennant, 2006 for an illustrative example). It is also possible to establish the magnitude of multidimensionality in terms of how it affects the accuracy of person estimates (RUMM Lab, 2009; Marais and Andrich, 2008a), and it may be necessary to analyse the subsets of items separately if the magnitude is large (Marais and Andrich, 2008a). Essentially, the decision regarding how to handle multidimensionality is "a relative issue and depends on the purpose of the scale or assessment" (Marais and Andrich, 2008a: 2012).

The second goal relates to independence between item responses. Summed item responses are more valid and reliable than responses to only one item only if multiple items provide unique information. If for the same person, the response to one item governs the response to another then some information is redundant. This has been shown to inflate scale reliability and the variance of person/item estimates (Marais and Andrich, 2008b; Smith, 2005; Marais, 2009). However, Smith (2005) points out that "the amount of redundancy needs to be considerable before such interpretations would be adversely impacted" (Smith, 2005: 147). Item residual correlations will reveal dependency between pairs of items. If suspected, its effect can be quantified (see Andrich and Kreiner, 2010 for the case of dichotomous items; Andrich et al., 2012 for the polytomous case). Depending on the magnitude of dependence, problematic items can be dealt with using the item split procedure: namely, the response to the dependent item is conditioned on the response to the independent one (see Andrich, 1985; Andrich and Kreiner, 2010; Zenisky, Hambleton and Sireci, 2002). However, if dependency is widespread or if a more complicated dependence structure is observed (e.g. a single item governs the response to many others), the procedure may not be optimal.

3.3.4.7 Scale stability over time

Extending the analysis to investigate item stability over time will help to determine whether the scores from the repeated test can be reliably calibrated onto a single metric. Person-time dependencies may affect the measurement of change due to response dependence over time (Smith, 2005: Marais, 2009). Moreover, item functioning may not remain constant over time (Wright, 1996). The item difficulty estimates should not 'drift' but should remain fairly constant over time (Wells et al., 2002; Miller and Fitzpatrick, 2009). In the current analysis, the first step is to investigate person-time dependency by conducting a Differential Item Functioning (DIF) analysis of each item by time point. The data are stacked to examine DIF of the same item by time point (Marais, 2009; Wright, 1996; also see Barley and Jones, 2006). Large DIF effects by time indicate scale instability. Then, item calibrations are compared across time to pinpoint inconsistencies in item functioning. Very unstable items can be removed from the scale to improve scale stability (Barley and Jones, 2006). Alternatively, if DIF is significant, it is possible to anchor the scales on the most stable items (Wright, 1996). Anchoring item calibrations in this way is thought to produce more reliable person ability estimates and thus a more accurate description of person change (Wright, 1996). The final scaling procedure will depend on the results of the DIF for time analysis. The "stacking" approach will only be used if the DIF effect is minimal. If dependency is significant, the most stable items will be used as anchor items to equate the tests.

3.3.4.8 Score validation and transformation

To verify that the Rasch scores are meaningful, the distribution of the Rasch scores are compared with the distributions of the raw scores. The scores are then correlated with other measures of achievement, available in the data set: i.e. maths and literacy (Chapter 4, Tables 4.8 and 4.9). The correlations between the scores and home background factors that are known to vary with vocabulary are examined. The extent of guessing is also investigated. A guessing parameter is not included in model estimation, rather it is considered an anomaly to be addressed in the data. Since guessing is more likely to be a characteristic of students rather than items (Ryan, 1983), including parameters to control for guessing is akin to hiding problems that should be addressed and may undermine the validity of score interpretations (Andrich, 2004).

The Rasch scores, which are expressed in logits are transformed onto a scale with a mean of 500 and a standard deviation of 100. The purpose of the transformation is to avoid negative ability scores. The transformed scores, which will be referred to as scaled scores, are precise measures of where children sit along the PPVT scale. The scores can be used for statistical analyses of children's developmental outcomes. The formula used to transform the Rasch logit scores is as follow (Equation 3.3):

$$\theta_n^* = [(\theta_n - m)/sd] * 100 + 500 \qquad \ldots 3.3$$

where θ_n is the ability for child n in logits and θ_n^* the transformed ability for child n, m is the weighted mean score of all children and sd is the weighted standard deviation of all children.

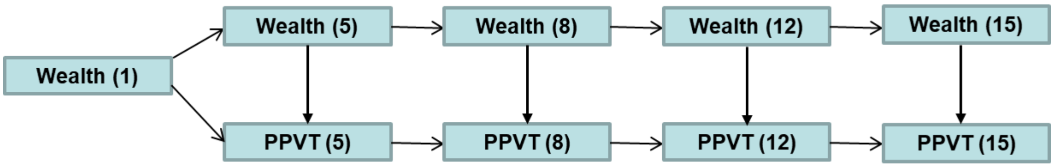
3.4 Research design: wealth effects on vocabulary development

3.4.1 Conceptual framework for analysing the timing of wealth effects

The research is rooted in understanding patterns of change in the relationship between living standards and children's developmental trajectories from early childhood to adolescence, and specifically how inequalities evolve over time. Due to the primacy of basic needs among the Ethiopian sample of children, and the subsistence-based economies in which the children live, a monetary-based measure of living standards is deemed most appropriate for relating changes in the children's outcomes to their socioeconomic contexts. In low-income countries, household consumption is considered a more direct indicator of living standards than income as it captures both food and non-food items consumed by the household (O'Donnell et al., 2008; Deaton and Zaidi, 2002). Patterns of change in living standards are also more efficiently reflected by household consumption than wealth (Rutstein and Johnson, 2004). However, human developmental research in low-income countries generally applies wealth-based measures of living standards, as wealth data are much easier to collect and interpret than expenditure data (Deaton and Grosh, 1998).

The Young Lives data set includes two measures of living standards: the household wealth index score and total household expenditure. The current study focuses on wealth as the main predictor variable. Whereas wealth is measured at each round of the survey, including when the children were aged one year, household expenditure data are only available at age five, eight and 12 years. Since developmental assessments were also carried out when the children were five, eight, 12 and 15 years old, the wealth variable is the preferred antecedent measure of living standards. It captures children's socioeconomic contexts prior to the administration of the PPVT and throughout the course of the testing process, as illustrated below.

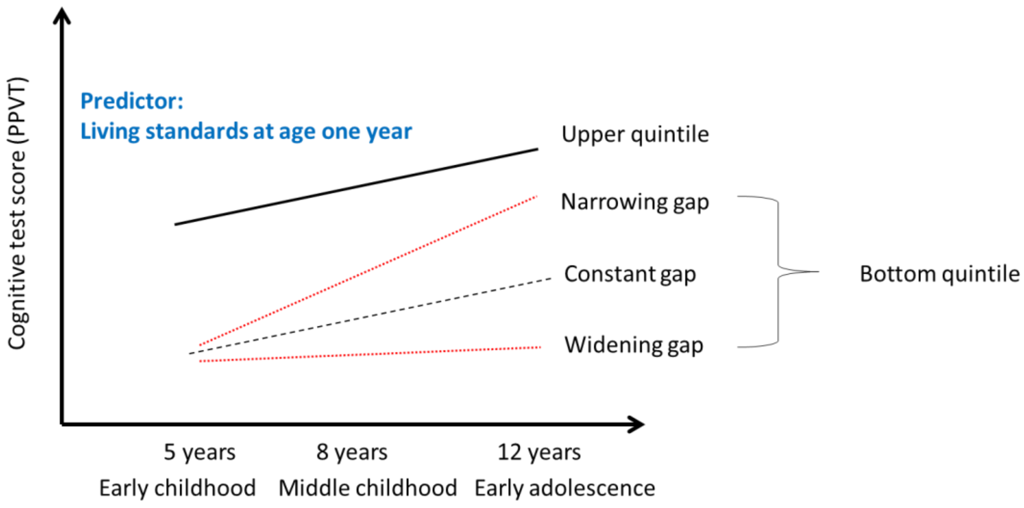
Figure 3.2 Wealth as a predictor of the PPVT developmental path



The study considers two conceptual and statistical approaches to examine the timing of poverty and children's development. First, it is expected that at five years of age there will be an initial gap in vocabulary ability between children living in the richest and poorest households, and that disparities will persist over the course of childhood. This is supported by previous studies in developing countries which have shown that poverty predicts developmental disadvantage in early childhood (Paxson and Schady, 2007; Fernald et al., 2011).

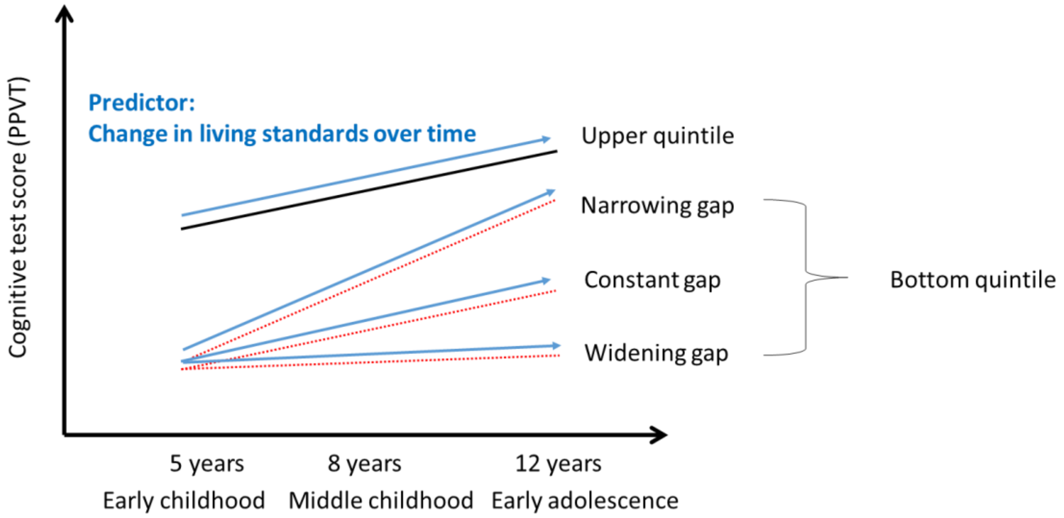
The first statistical model analyses how wealth at one year predicts onset vocabulary scores at five years and the rate of vocabulary growth through to the age of 15. The overall mean trends in initial status and the growth rate are computed and differences in individual trajectories are compared in relation to the means. Is there a gap in ability between the highest and lowest quintiles at age five? If so, does this gap persist or widen with age, narrow, or remain constant throughout middle childhood and early adolescence? Figure 3.3 illustrates the conceptual model for this investigation. Using the scaled PPVT scores, the analysis will show whether initial wealth, measured in the first year of childhood, predicts vocabulary test scores at age five years (i.e. the ability intercept) and the growth rate between five and 15 (i.e. the ability slope). If the intercept and slope variance is significant, the analysis will further examine the extent to which initial wealth explains this variance. The slope variance may be insignificant or significant but not sizeable, pointing to a launch model of developmental change. This would support the theoretical premise that cognitive ability is largely formed in early childhood and that interventions designed to alleviate poverty effects are needed earlier in childhood to prevent developmental disadvantage. Alternatively, the slope variance may be significant, implying that developmental change is more efficiently described by some other model that captures patterns of change in wealth over time.

Figure 3.3 Diagram of how wealth launches the developmental trajectory



If individual differences in the growth rate are observed, a second model will be estimated to investigate whether changes in living standards over time further help to explain patterns of development. Research from industrialised countries has shown that poverty exposure in early childhood has a more detrimental effect on ability than poverty experienced later in childhood or in adolescence (Alwin and Thornton, 1984; Guo, 1998; Duncan and Brooks-Gunn, 1997). However, few studies have looked at the pattern of change in the poverty-growth relationship, namely the effect of changes in living standards on changes in ability. For example, do increases in wealth over the course of childhood improve the developmental trajectory, or conversely, does a decrease in wealth predict developmental stagnation? Developmental theory suggests that, as cognitive abilities develop more rapidly in early childhood and learning deficits are cumulative, such increments and decrements in living standards would have a small effect on scores. The conceptual model for this investigation is presented in Figure 3.4.

Figure 3.4 Diagram of how changes in wealth influence the developmental trajectory



The findings may have important educational policy implications. If individual differences in vocabulary development are consistent with the launch model of developmental change, this would support theoretical expectations and previous empirical evidence. Such findings suggest that in order to promote future learning in Ethiopia, early interventions are needed to address initial gaps. The next step would then be to identify the pathways through which living standards influence cognitive ability in early childhood, and to develop early interventions accordingly. However, if the developmental trajectory is more efficiently predicted by changes in living standards over the course of childhood, then it will be important to identify the specific pathways through which living standards influence development at different childhood stages, ensuring that effective interventions are in place in early childhood, middle childhood and early adolescence.

3.4.2 Additional measures used for the study

The Young Lives study includes two composite measures of living standards. The first is a measure of household consumption based on consumed food items and non-food expenditures in the two weeks prior to the survey. Food items are those that are purchased, home produced, and received as gifts, in-kind payments, or transfers. Food groups can then be aggregated in different ways such as the approach recommended by the World Health Organisation dietary diversity measure: e.g. grains, fruits and vegetables, meat, dairy products, fish, and seafood (World Health Organisation, 2010). Non-food expenditures include spending on health, education, and clothing. The construction of the variable involved allocating a market value to each of the food items consumed, based on local market prices, and aggregating the food and non-food items. Nominal expenditures were converted into monthly terms before aggregation and adjusted for the cost of living (i.e. to obtain real expenditures) and then for household size. In Ethiopia, real expenditure is divided by household size adjusted for the number of equivalent adults (Azubuike and Briones, 2016: 8), to reflect “the fact that children typically consume less than adults, so that deflating [expenditures] by household size will understate the welfare of people who live in households with a high fraction of children” (Deaton and Zaidi, 2002: 45).

The second measure is a household wealth index, which is a composite measure comprising the average of three equally weighted sub-indices. The first is a housing quality index which is computed as the average of the number of rooms per person and the type of flooring, roofing, and walls (dummy = 1 if the minimum requirement is met). The second is an asset index which is the average of ownership of a set of ten consumer durables (e.g. radio, refrigerator, bicycle, television, phone, etc.). The third is a services quality index, which is the average of access to safe drinking water, electricity, sanitation, and adequate cooking fuel. The index is derived from the measures used by the World Bank and Macro International to develop the internationally comparable wealth index cited in the UNICEF Multiple Indicator Cluster Surveys (Young Lives, 2002). A wealth index score of less than 0.2 indicates a *very poor* household; households with index scores greater than or equal to 0.2 and less than 0.4 are classified as *poor* and households with index scores greater than or equal to 0.4 are *less poor* (Seager and de Wet, 2003). These arbitrary cut-offs can be adapted depending on the frequency distribution of the wealth data. As noted by Seager and de Wet (2003), “in the case of non-normally distributed data, arbitrary equal divisions may be less appropriate” (Seager and de Wet, 20013: 13).

For example, a positively skewed distribution would require a higher cut-off to ensure a sufficient number of individuals in the bottom category, and possibly fewer categories. In the case of a multimodal distribution, they argue that cut-off points should correspond to the “natural division” rather than arbitrarily chosen values (Seager and de Wet, 20013: 13). The Young Lives wealth data is normally distributed, as discussed in Chapter 5. As such, the cut-points suggested by Seger and de Wet (2003) can be applied. However, to address the research questions, wealth quintiles are constructed to compare outcomes between the poorest 20% and the richest 20% of the sample.

The household consumption expenditure data is not used in the current analysis since expenditure data were not collected in the first and fifth rounds of the survey, when the children were aged one and 15 years. This has important implications for the statistical modelling, namely that the developmental effect of expenditure could only be analysed from age five onward. Wealth data are available for each round of the survey, and thus it is possible to analyse the wealth effect on developmental trajectories from age one. For this reason, wealth is the more suitable predictor variable as information on the children’s wealth status is available prior to the administration of the developmental assessments.

3.4.3 Analysing development trajectories (the base model)

IBM SPSS 25 and MPlus Version 8 are used to fit the statistical models (IBM Corp., 2012; Muthén and Muthén, 2017). Developmental trajectories are first analysed without taking into account any antecedent variables, to examine: (i) the difference in the means across measurement occasions; (ii) the shape of the developmental trajectory, whether it is linear or more complex; and (iii) the variation in initial ability (i.e. intercept) and ability growth (i.e. slope) across children. Developmental patterns are assessed by comparing individual growth curves to the overall mean trends. Differences may remain stable over time (constant growth), they may narrow or grow wider.

The analysis seeks to fit a multilevel model (MLM) for change to estimate initial status and growth rate parameters, following the methodology and notation proposed by Singer and Willet (2003). The main advantage of multilevel growth curve modelling lies in the possibility of estimating developmental trajectories while controlling for factors that may influence both initial starting points and the growth rate. Repeated measures ANOVA can also be used to examine mean-level changes over time and how much of the variation in scores at each time point is attributable to contextual factors, i.e. wealth. This is done by estimating the between-subject main effect of wealth on test scores at each time point. However, since MLM treats time more flexibly as a continuous variable, MLM allows for the estimation of individual differences in growth over time, which is not achievable with ANOVA. Additional limitations of using repeated measures ANOVA and multivariate analysis of variance (MANOVA) to analyse repeated measures data are discussed in Heck et al. (2009: 167-168).

The MLM for change involves two components: a level-1 individual growth model that describes how each person changes over time and a level-2 inter-individual difference model that describes the between-child variance in the change (Raudenbush and Bryk, 1987; Singer and Willett, 2003).

Level-1 model. The level-1 *individual growth model* describes how the child changes over time (within-child variability). Equation 3.4 represents the linear growth curve for child i. For each child i, and at each time point t, cognitive ability $(PPVT_{ti})$ is modelled as a function of age in months (AGE_{ti}), and an error term ε_{ti}. Thus, the time of observation varies across individuals. The level-1 parameters, β_{0i} and β_{1i}, represent child i's initial ability and the period rate of linear change in ability, respectively. To observe progress from the children's initial starting points the level-1 variable "age" is centred on its first value which corresponds to 52 months, the age of the youngest child in the sample and the time point from which the cognitive developmental trajectory is "launched" (Singer and Willet, 2003: 182). The unobserved error variance is assumed to be heterogeneous across time points (i.e. age five, eight, 12 and 15), and the covariance terms are set to zero. This simple "diagonal" variance-covariance structure is deemed an appropriate starting point for repeated measures when there are few measurement occasions (Heck et al., 2014: 190).

$$PPVT_{ti} = \beta_{0i} + \beta_{1i}AGE_{ti} + \varepsilon_{ti} \qquad \text{...3.4a}$$

$$\varepsilon_{ti} \sim N\,(0,\sigma_{\varepsilon}^{2}) \qquad \text{...3.4b}$$

Level-2 unconditional model. The second component is a level-2 *inter-individual difference model* that describes how the intercept and the growth rate vary across individuals. The model is hierarchical such that level-1 parameters depend on level-2 parameters. Equation 3.5 is specified to enable the intercept and slope parameters to vary between individuals (Singer and Willet, 2003: 57-60). The fixed effect parameters γ_{00} and γ_{10} are the expected intercept and the expected slope (also termed the grand means), and u_{0i} and u_{1i} are the individual deviations from the means. The heterogeneous covariance structure describes the variance of the random effects, σ_0^2 and σ_1^2 (i.e. unequal across time points) and the covariance between the intercept and slope, σ_{01}/σ_{10}.

$$\beta_{0i} = \gamma_{00} + u_{0i} \qquad \text{...3.5a}$$

$$\beta_{1i} = \gamma_{10} + u_{1i} \qquad \text{...3.5b}$$

$$\begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \\ \sigma_{10} & \sigma_1^2 \end{bmatrix}\right) \qquad \text{...3.5c}$$

When Equations 3.4 and 3.5 are combined, the model may be written as,

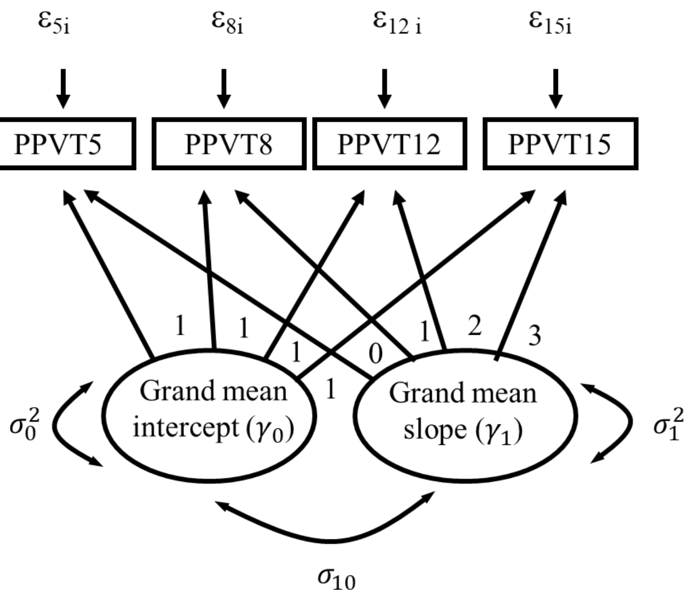
$$PPVT_{ti} = \gamma_{00} + \gamma_{10}AGE_{ti} + u_{0i} + u_{1i}AGE_{ti} + \varepsilon_{ti} \qquad \text{...3.6}$$

The model can be modified to investigate the shape of the trajectory by including a quadratic term, as shown in Equation 3.7. The quadratic effect is treated as fixed across individuals.

$$PPVT_{ti} = \gamma_{00} + \gamma_{10}AGE_{ti} + \gamma_{20}[AGE_{ti}]^2 + u_{0i} + u_{1i}AGE_{ti} + \varepsilon_{ti} \qquad \ldots 3.7$$

The linear multilevel model for change, or base model, can also be specified latent growth model within a structural equation framework, as shown in Figure 3.5.

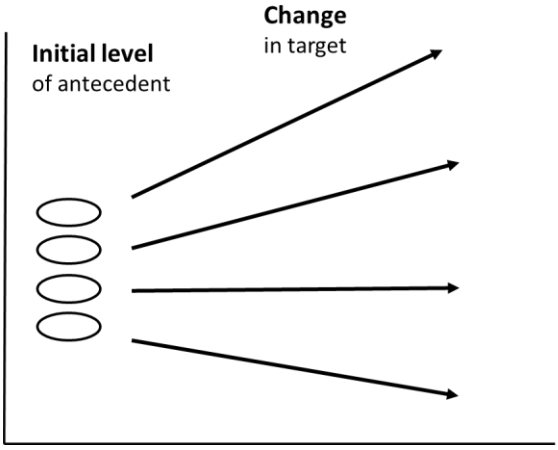
Figure 3.5 The multilevel model for change as a latent growth model



3.4.4 The effects of initial wealth on development (launch model)

The conventional 'Launch Model' of developmental change (Skinner et al., 1998), shown in Figure 3.7, is adapted to analyse how differences in wealth in early childhood affect differences in later development. The model is based on assumptions about the importance of early experiences for cognitive progress throughout childhood, the stability of context over time, and continuity in development (Skinner et al., 1998). It posits that the initial level of an antecedent launches the developmental trajectory. A higher initial level of the antecedent variable launches a higher growth in the target developmental trajectory than a lower initial level. Thus, individual differences in progress are modelled as a function of household wealth measured when the children were one year old. Since MLM partitions the variance according to its level, it is possible to ask how much variation in each of the fixed effects parameters (initial status and the growth rate) is explained by initial wealth and the extent of between-child variation in both initial status and growth.

Figure 3.6 Launch Model of development (Skinner et al., 1998)



The main antecedent variable, initial wealth, is modelled as a level-2 time-invariant predictor. The level-1 model remains unchanged (Equation 3.4). The level-2 conditional model is elaborated in Equations 3.8. Both the intercept and slope parameters are modelled conditionally as a function of initial wealth, W_i, the population means and individual deviations from those means. The random and fixed effects are assumed to be normally distributed. An unstructured variance-covariance matrix is assumed for the random effects to obtain estimates for the covariance between onset and growth. The parameters γ_{00} and γ_{10} are the expected level of initial ability and the rate of change per period, adjusted for W_i. The parameters γ_{01} and γ_{11} represent the mean difference in initial ability and growth, controlling for W_i. The random effects u_{0i} and u_{1i} are the effects of child i on expected initial ability and the expected rate of change, controlling for W_i. The variance terms are σ_0^2 (the intercept variance), σ_1^2 (slope variance for time) and the covariance between initial status and time (σ_{01}/σ_{10}).

$$\beta_{0i} = \gamma_{00} + \gamma_{01}W_i + u_{0i} \qquad \text{...3.8a}$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11}W_i + u_{1i} \qquad \text{...3.8b}$$

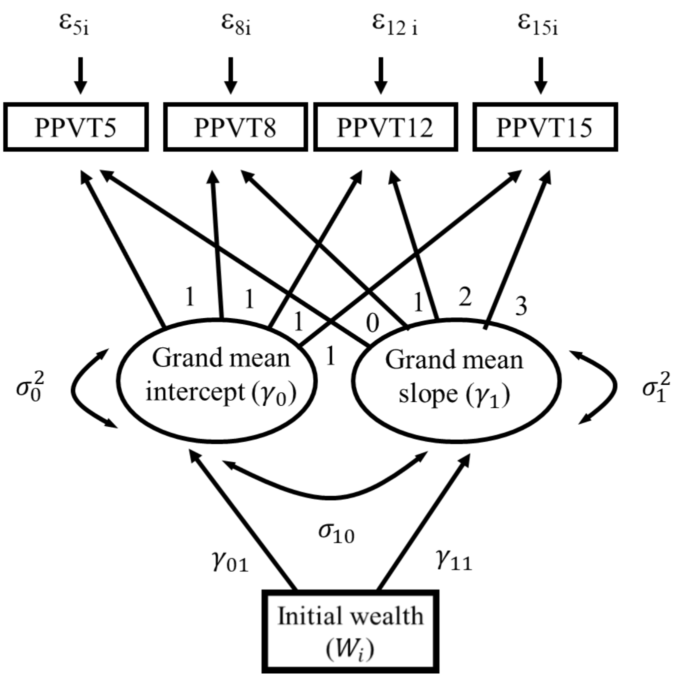
$$\begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \\ \sigma_{10} & \sigma_1^2 \end{bmatrix}\right) \qquad \text{...3.8c}$$

The full linear model can be written as,

$$PPVT_{ti} = \gamma_{00} + \gamma_{10}AGE_{ti} + \gamma_{01}W_i + \gamma_{11}W_iAGE_{ti} + u_{0i} + u_{1i}AGE_{ti} + \varepsilon_{ti} \qquad \ldots 3.9$$

Of additional interest is the magnitude and significance of the gap in initial status and growth between wealth quintiles. Growth curves are plotted for children in the highest and lowest quintiles. This will show the gap in ability at five years and whether the two trajectories are parallel (i.e. no slope variance) or if they become more or less divergent over time. A t-test will indicate the significance of the difference in initial status and growth rate estimates between the two groups. The launch model can also be specified as a latent growth model, depicted in Figure 3.7.

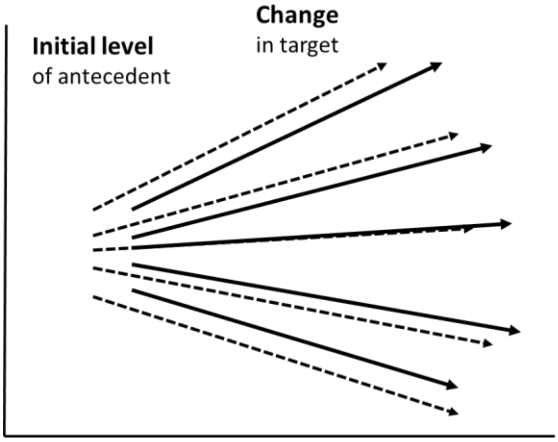
Figure 3.7 The launch model expressed as a latent growth model



3.4.5 Th effects of changes in wealth on development (change-to-change model)

The 'Change-to-Change' model of development (Skinner et al., 1998), shown in Figure 3.8, is adapted to explore associations between patterns of change in vocabulary ability and changes in living standards. The model assumes that increments or decrements in a target developmental trajectory would be influenced, not simply by the initial level of a predictor variable or its average over time, but by the pattern of change in that variable (Skinner et al., 1998). The empirical model posits that change in ability is a function of change in wealth. This model is applied only if the children's growth rates significantly vary, to determine if increases in wealth corresponds to an improvement in the developmental trajectory, and conversely, if decreasing wealth is associated with developmental stagnation or regression. Individual differences in progress are modelled as a function of wealth, expressed as a time-varying predictor measured at each time point.

Figure 3.8 Change-to-Change Model of development (Skinner et al., 1998)



Level-1 model with a time-varying covariate. Wealth, $Wealth_{ti}$, is modelled as a time-varying continuous variable (level-1 variable). The within-subject variability in wealth over time is reflected in the level-1 model, which is specified in Equation 3.10. Three parameters now characterise each child $i's$ trajectory, β_{0i}, β_{1i}, θ_i.

$$PPVT_{ti} = \beta_{0i} + \beta_{1i}AGE_{ti} + \theta_i Wealth_{ti} + \varepsilon_{ti} \qquad \text{…3.10a}$$

$$\varepsilon_{ti} \sim N\,(0,\sigma_\varepsilon^2) \qquad \text{…3.10b}$$

MLM and SEM place different assumptions on time-varying covariates (discussed in section 3.4.3). Consequently, the change-to-change model is specified slightly differently in the two approaches. Within an MLM framework, time-varying covariates vary over individuals and have random effect coefficients. Within a structural equation framework, the slopes for time-varying covariates vary over time points rather than individuals and have fixed effects coefficients. The fixed effect parameter for wealth is therefore specified as θ_t rather than θ_i. The level-1 model is written as follows in Equation 3.10b.

$$PPVT_{ti} = \beta_{0i} + \beta_{1i}AGE_{ti} + \theta_t Wealth_{ti} + \varepsilon_{ti} \qquad \text{…3.10c}$$

Meaningful centring of Level-1 predictors is crucial in MLM, particularly when the predictor of interest has no true zero point, i.e. variables such as age, height, intelligence, and wealth. Deciding where to centre is not straightforward (Singer and Willet, 2003; Enders and Tofighi, 2007). As noted by Enders and Tofighi (2007):

> … the centering of Level 2 (e.g. organizational level) variables is far less complex than the centering decisions required at Level 1, as it is only necessary to choose between the raw metric and centering at the grand mean; centering within cluster is not an option because each member of a given cluster shares the same value on the Level 2 predictor.
>
> Enders and Tofighi (2007: 121)

To observe the effect of the wealth trajectory on the growth trajectory, the variable "age", which is measured in months, is centred at the starting value (52 months or 4 years and 4 months). This value relates to the age of the youngest child in the sample. This is particularly important for comparing the hypothesised statistical models, launch and change-to-change. However, if the primary interest is in understanding effects and relationships at the end of the growth process, then time should be centred on its last value (Biesanz et al., 2004). In contrast, we are interested in understanding the effects of living standards on ability at onset in early childhood and then at each subsequent stage of childhood, i.e. the effect of wealth on cognitive outcomes in middle childhood, in comparison to its effects in early adolescence, and so on. The wealth variable is centred on its grand mean value. Centring age at the last time point will result in a model that describes ability at the end of the period, and the effect of wealth on the end point of the growth process.

Level-2 model conditional model. The level-2 conditional model is elaborated in Equation 3.11. Initial wealth, measured at age one year, is modelled as a time-invariant level-2 predictor. The purpose of retaining wealth at age one is to determine whether additional variance is explained by including changes in wealth over time. The level-2 model parameters can be interpreted as per the launch model.

$$\beta_{0i} = \gamma_{00} + \gamma_{01}W_i + u_{0i} \quad \text{...3.11a}$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11}W_i + u_{1i} \quad \text{...3.11b}$$

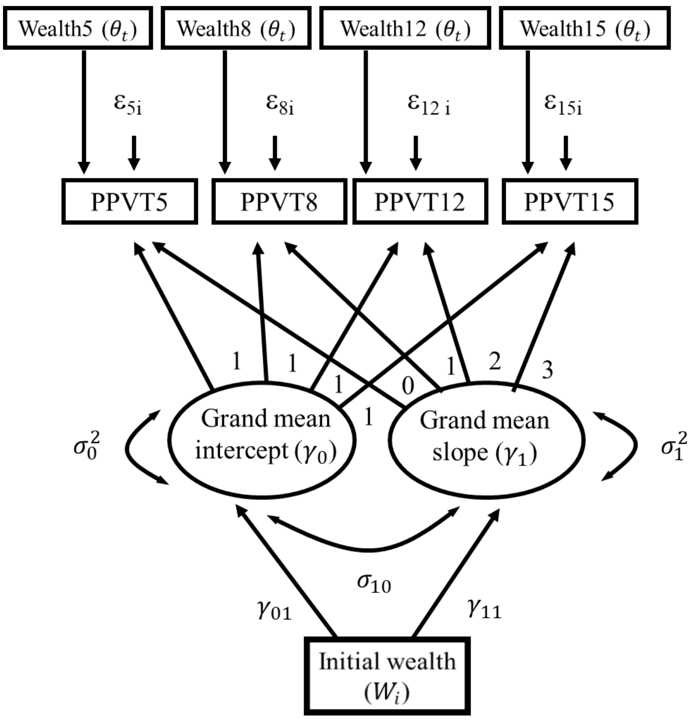
$$\begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \\ \sigma_{10} & \sigma_1^2 \end{bmatrix}\right) \quad \text{...3.11c}$$

The full model is specified in 3.12.[9]

$$PPVT_{ti} = \gamma_{00} + \gamma_{01}W_i + (\gamma_{10} + \gamma_{11}W_i)AGE_{ti} + \theta_t Wealth_{ti} + u_{0i} + u_{1i}AGE_{ti} + \varepsilon_{ti} \quad \text{...3.12}$$

This linear change-to-change model is specified as a latent growth model below.

Figure 3.9 The change-to-change model expressed as a latent growth model



---

[9] Note that the mathematical model conforms to SEM assumptions. The slope for the time-varying covariate $Wealth_{ti}$ is assumed to vary over time rather than individuals. Thus, θ_t is not involved in the random part of the model (i.e. Level 2).

3.5 Research design: mediated wealth effects on vocabulary development

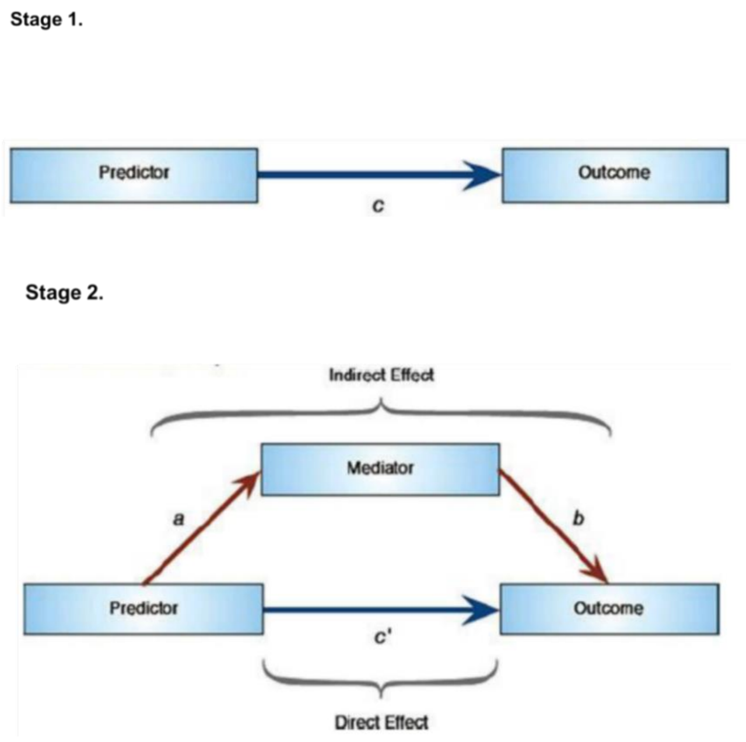
3.5.1 Conceptual framework for the mediation analysis

The final stage of the research is to identify the pathways through which poverty affects children's development. It is expected that multiple mechanisms partially explain individual differences in initial ability and the rate of growth, as well as the developmental gap between the richest and the poorest. If individual differences are consistent with the launch model of developmental change, the focus of the mediation study will be on characterising the poverty-growth relationship in earlier childhood using quantitative information on the children's wider environments at age one and at five years. Such findings would suggest that to promote future learning in Ethiopia, early interventions that address initial gaps are needed. If changes in wealth predict developmental trajectories more efficiently than the launch model, then it will be important to identify the pathways through which living standards influence development at different stages of childhood, and to examine the longer-term effects of early experiences on the poverty-growth relationship.

Multilevel mediation techniques are used due to the hierarchical structure of the data, i.e. time points nested in children, and building on the growth modelling strategy laid out in the previous chapter. Vocabulary development is the outcome of interest, wealth is the independent variable and the intervening contextual variables are the mediators. The previous section laid out the strategy for analysing the direct effect of wealth on development. This section describes how wealth is modelled to predict development both directly and indirectly through mediator variables.

Figure 3.10 shows the standard two-step conceptual model of the relationship between the predictor, the outcome, and the mediator, with arrows representing the relationship between the variables. If the strength of the relationship between the predictor and the outcome variable is reduced by including the mediator variable in the model, this would be a mediation effect. In single-level mediation, the mediation effect is also referred to as an indirect effect and is expressed as c – c' = a × b, where a × b is an interaction effect. The effect of the main predictor on the outcome variable is referred to as the *direct effect* (c'). Full mediation occurs when c' is zero. Partial mediation is observed when c' is less than c in that the effect of the main predictor is reduced when the mediator variable is included.

Figure 3.10 Diagram of a basic mediation model

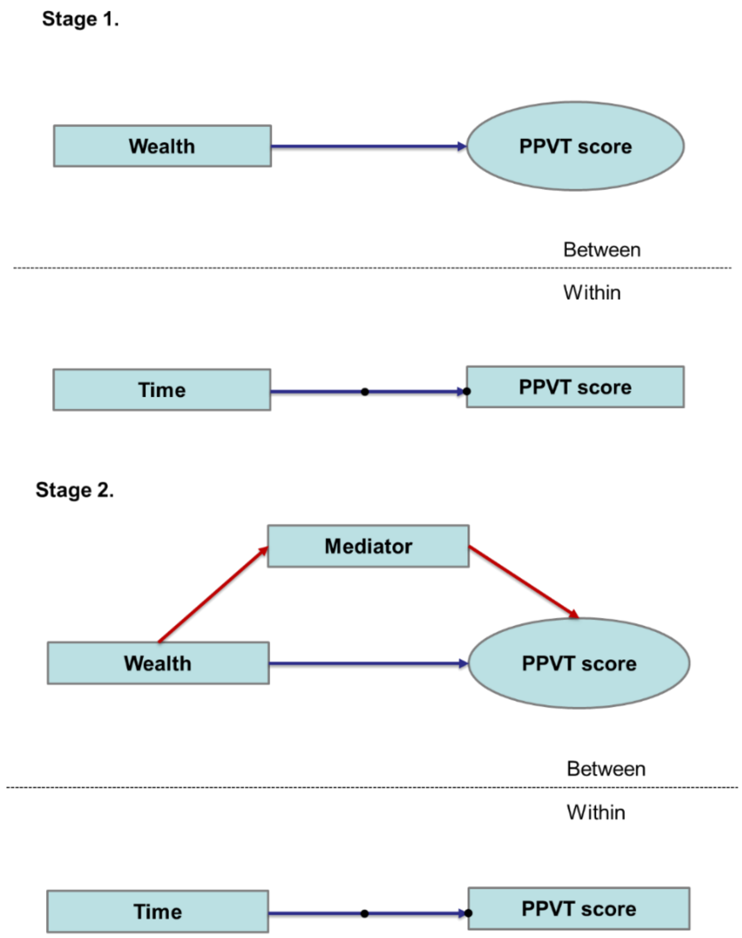


Source: Field, 2013, figure 10.11.

For the PPVT data, which is structured hierarchically with time points nested in children, two-level mediation models can be specified within and between individuals. The first level, the within-level mediation model, describes the effect of time on scaled scores. The second level, the between-level mediation model, describes the relationship between the main predictor (i.e. wealth), the mediator and the outcome (i.e. the scaled scores). The between-group indicators are depicted as ovals because “they are viewed as latent variables between groups and represent the underlying heterogeneity among individuals due to their randomly varying intercepts in the within-group portion of the model” (Heck and Thomas, 2015: 17). The solid dots on the within-group model represent the hypothesised randomly varying intercepts and slopes across individuals, which the analysis will seek to explain.

Figure 3.11 shows the conceptual model for within- and between-level mediation as a two-step process, comparing the strength of the relationship between the predictor and the outcome variable before and after including the mediator in the model. Stage 1 of this process is the focus of the second empirical study which looks at the effects of wealth on the children's developmental trajectories, the methodology for which is presented in the previous section. The set of equations corresponding with the mediation analysis are presented in the section 3.5.3.

Figure 3.11 Diagram of the multilevel mediation model depicted as a two-step process



3.5.2 Additional measures used for the study

Proximal influences in earlier childhood, which are also referred to as the "engines of development" (Mayes and Lewis, 2013: 19), are selected as the key indirect effects for the mediation study. Proximal factors include the home learning environment and interactions with caregivers (Hart and Risley, 1995; Roberts, Jurgens, and Burchinal, 2005; Caldwell and Bradley, 1984), health and nutrition in early childhood (Walker et al., 2007), and the quality of preschool experiences (Sammons et al., 2014; Sylva et al., 2014; and Hall et al.). Indicators related to these constructs have shown to vary significantly with living standards, which is discussed in the literature review section. Potential mediator variables were explored at each round of the Young Lives household survey. The survey data includes information on different aspects of childhood that are important for understanding children's development and learning, these include: nutrition and health, parental educational and occupational background, aspects of the home learning environment, preschool attendance, pregnancy and delivery, breastfeeding and infant health. One of the main advantages of the Young Lives survey is the breadth of the data set, which seeks to quantify childhood experiences and conditions across a broad range of themes. The survey does not provide in-depth information on every theme and is not designed to measure all constructs comprehensively.

Underscoring the primacy of early years' experiences for development and taking into account the data limitations, three key early childhood influences have been selected for the mediation study: children's health and nutrition, the home learning environment, and children's preschool experience. Data on pregnancy, delivery, and breastfeeding are not used in the analysis. The data structure includes two indicators of mothers' breastfeeding experience and one indicator of children's birthweight, based on mothers' self-reports when the children were one year old. To obtain this information, mothers were asked retrospectively if they had ever breastfed their children and for how long breastfeeding took place. Mothers were also asked to recall their child's birthweight. However, apart from being based on retrospective information collected from mothers, the variables are also significantly skewed and there is a large amount of missing data. Four indicators are used to proxy the mediating constructs, in addition to the outcome variable and the main predictor (Table 3.6)

Table 3.6 Mediation model variables

Level	Variable
Outcome	PPVT scaled score
Main predictor	Wealth index score
Between-child mediator	Height-for-age z-score BMI-for-age z-score Mothers' educational level Participation in preschool education

The Young Lives data set includes composite measures of stunting and wasting, recorded at each time point. The first is the height-for-age z-score, which is a measure of stunting. The data structure also includes the body mass index (BMI)-for-age z-score, which measures thinness or wasting. The BMI is defined as a person's weight in kilograms divided by the square of the person's height in meters. The z-score is the deviation of the value for an individual from the median value of the reference population, divided by the standard deviation for the reference population (WHO, 1995). Both measures are computed and scaled based on the WHO growth standards.

Stunting indicates slowing in skeletal growth, which may be reduced from birth and worsened by poor overall economic conditions, chronic or repeated infections, and inadequate nutrient intake (WHO, 1986). Low height-for age "reflects a process of failure to reach linear growth potential as a result of suboptimal health and/or nutritional conditions" (WHO, 1997: 46). The process of stunting slows down around three years of age. Thus, low height-for-age among older children reflects a state of having failed to grow (WHO, 1997). Wasting refers to a deficit in tissue and fat mass compared with the amount expected in a child of the same height or length and may result either from failure to gain weight or from actual weight loss (WHO, 1986). Low weight-for-height, which is usually very low even in poor countries, indicates a severe process of weight loss resulting from a recent severe event such as a severe food shortage or disease (WHO, 1997).

Taken together, weight-for-height and height-for-age account for more than 95% of the variance in weight-for-age, thus it is recommended to consider both when assessing a child's anthropometry (WHO, 1986). However, since height-for-age reflects cumulative exposure to social and economic deprivation, it is widely used as an indicator of a child's overall nutritional condition (WHO, 1986). Weight-for-height is important for describing a child's current health and nutritional situation, especially in an emergency context. Moreover, weight-for-height measures, such as BMI are particularly useful for understanding longitudinal patterns of children's growth and development (Furlong et al., 2016).

A home learning environment scale was not included in the design of the household survey. The data set does include a measure of caregiver's functional literacy, based on the caregiver's ability to read a letter or newspaper. However, this variable is only available for half of the sample and is skewed in that a small number of caregivers were able to perform the task. Mothers were also asked to report the highest educational grade they completed: none, grade one through 12, post-secondary or vocational, university, adult literacy, or religious education. For the present study, mothers' educational level was recoded as: no schooling, primary, secondary, post-secondary, adult literacy and other. The variable is available for most of the sample and is strongly and positively correlated wealth, particularly when the "adult literacy" and "other" categories are suppressed. For this reason, mothers' educational level is used as a proxy for the home learning environment instead of caregiver literacy.

When the Young Lives children were five years old, parents were asked if their child ever attended preschool. This is a simple "yes" or "no" question. Information on the duration of pre-schooling, hours per day in preschool or the quality of preschool was not collected at this sweep. Information on children's pre-schooling is available for the older cohort children in later rounds of the survey. However, the data collected are based on parent's retrospective accounts. When the older cohort children were 15 years old, their parents were asked to recall the years during which their child attended an early childhood facility, the type of programme (e.g. creche, nursery, or kindergarten), the age at which their child started preschool, and the type of preschool ownership (e.g. whether it was government funded or not). This information was not collected for the younger cohort sample.

All the mediator variables are associated with main predictor and the outcome, discussed in more detail in the empirical chapter. Mothers' educational level shows a high correlation with wealth ($r = 0.624$, p-value < 0.01). A simple t-test indicates that children who attended preschool are wealthier than children who did not ($md = 0.252$, $p < 0.01$). Height-for-age is positively correlated with wealth ($r = 0.228$, p-value < 0.01), as is BMI-for-age, although the correlation coefficient is small ($r = 0.058$, p-value < 0.01). Moreover, the mediator variables are significantly positively correlated with the scaled scores. Height-for-age is positively correlated with test scores ($r = 0.170$, p-value < 0.01). BMI-for-age is negatively correlated with test scores ($r = -0.170$, p-value < 0.01). For mothers' education, the correlation is positive ($r = 0.38$, $p < 0.01$). Children who have never been to preschool score 74 points lower on the PPVT ($p < 0.01$). Additionally, the results conform with the theoretical expectation that proficiency increases with the child's age ($r = 0.711$, p-value < 0.01). Wealth is weakly correlated with mothers' age ($r = 0.093$, p-value < 0.001). Simple t-tests show that the mean level difference in performance is not significant between boys and girls, but the difference is significant between geographic stratum (p-value < 0.01). Children in urban areas are wealthier than their rural counterparts ($md = 0.280$, p-value < 0.01).

3.5.3 Analysing the mediated effects of wealth on vocabulary development

There are different approaches to testing hypotheses about mediating pathways (Baron and Kenny, 1986; MacKinnon, et al., 2002; MacKinnon et al., 1995). However, as

Preacher et al. (2010) have pointed out, traditional methods are inappropriate with hierarchically structured data, “primarily because the assumption of independence of observations is violated when clustered data are used, leading to downwardly biased standard errors if ordinary regression is used” (Preacher et al.; 2010: 209). Different multilevel techniques have been proposed for testing mediation with nested data (for example, Krull and MacKinnon, 2001; MacKinnon, 2008). Preacher and colleagues (2010, 2011) propose multilevel structural equation modelling (MSEM) which harnesses the advantages of both multilevel and SEM approaches. Within an MSEM framework, higher level mediators can be included for different levels of analysis, measurement error can be explicitly modelled and the conflation between within- and between-subject effects is minimised (Preacher et al., 2010 and 2011). MSEM is based on factor analysis and path analysis techniques. It can be extended for growth curve modelling with the repeated measures data structured in wide format such that each row represents a respondent with their repeated responses in separate columns. Alternatively, a multilevel model (MLM) for change can be used to estimate growth curves. MLM for change is an application of multilevel multivariate regression (Mehta and Neale, 2005). The data are structured in long format and each row represents one time point so that each respondent has multiple data rows for each variable. Variables that remain constant over time will have the same value in each row. Whereas MLM approaches were specifically developed to deal with the methodological challenges associated with nested data structures, MSEM is an extension of SEM to longitudinal and hierarchical data. Despite the differences in their conceptual underpinnings, from a methodological point of view, “multilevel regression and SEM provide analytically identical model solutions in many multilevel applications (e.g. growth modelling) despite slightly different ways of defining and estimating these models” (Heck and Thomas, 2015: 20).

In the current analysis, the scaled score is modelled as the level-1 continuous outcome variable, which is measured at each measurement occasion. Representation of the main predictor variable will depend on which statistical model provides a better explanation of the intercept and slope variance. If the launch model of developmental change more efficiently describes the data, then wealth is modelled as a level-2, time invariant predictor, i.e. as an antecedent variable that launches the developmental trajectory. If the change-to-change model shows better data-model fit in that changes in wealth provide a better explanation of cognitive development, then wealth is modelled as a level-1, time-varying predictor.

3.5.4 Mediated launch model of developmental change

The level-1 *individual growth model* states that for each child i, and at each time point t, ability $(PPVT_{ti})$ is modelled as a function of age AGE_{ti}, and an error term ε_{ti}. Age is measured in months and is centred on its first value to observe progress from the children's initial starting points. The level-1 parameters, β_{0i} and β_{1i}, represent child i's initial ability and the rate of linear change in ability. The unobserved error variance is assumed to be heterogeneous for each time point and the covariance terms are set to zero.

$$PPVT_{ti} = \beta_{0i} + \beta_{1i}AGE_{ti} + \varepsilon_{ti} \qquad \text{…3.13a}$$

$$\varepsilon_{ti} \sim N\,(0,\sigma_{\varepsilon}^{2}) \qquad \text{…3.13b}$$

The level-2 *random intercept model* is specified in Equations 3.14 with wealth as the main predictor. To capture the direct effect of living standards on vocabulary trajectories, the intercept parameter is modelled as a function of initial wealth measured when the children were in their first year, denoted as W_i. The variable, M_i, is the mediator which is included to estimate its indirect effect on the scaled score. The mediators are modelled as time-invariant between-child predictors. In the case where the slope variance is small or insignificant, it is possible to fix the slope for the entire sample such that $\beta_{1i} = \gamma_{10} + \gamma_{11}W_i + \gamma_{12}M_i$. An unstructured variance-covariance matrix is assumed for the random effects to obtain an estimate of the covariance term.

$$\beta_{0i} = \gamma_{00} + \gamma_{01}W_i + \gamma_{02}M_i + u_{0i} \quad \text{…3.14a}$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11}W_i + \gamma_{12}M_i + u_{1i} \quad \text{…3.14b}$$

$$M_i = \gamma_{20} + \gamma_{21}W_i + u_{2i} \quad \text{…3.14c}$$

$$\begin{bmatrix} u_{0i} \\ u_{1i} \\ u_{2i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & & \\ \sigma_{10} & \sigma_1^2 & \\ \sigma_{20} & \sigma_{21} & \sigma_2^2 \end{bmatrix}\right) \quad \text{…3.14d}$$

The parameters γ_{00} and γ_{10} are the expected level of initial ability and the rate of change per period. The parameters γ_{01} and γ_{11} represents the mean difference in initial ability, controlling for W_i. Similarly, γ_{02} and γ_{12} represent the mean effect on the intercept, controlling for M_i. The parameter γ_{20} is the intercept of the mediator variable and γ_{21} is the mean effect of initial wealth on the mediator variable. The residuals are defined as u_{0i}, the between-person intercept residual (i.e. the random effect of person i on mean initial ability), u_{1i}, the between-person slope residual (i.e. the random effect of person i on mean growth). and u_{2i} is the mediator residual (i.e. the random effect of person i on the mediator). The intercept variance is denoted as σ_0^2, the slope variance is set to zero, the mediator variance is denoted as σ_1^2, σ_{10} is the covariance between the intercept and the mediator. The full model is specified in the following steps (equation 3.15 and 3.16).

$$PPVT_{ti} = \gamma_{00} + \gamma_{10}AGE_{ti} + \gamma_{01}W_i + \gamma_{11}W_iAGE_{ti} + \gamma_{02}M_i + \gamma_{12}M_iAGE_{ti} + u_{0i} + u_{1i} + \varepsilon_{ti} \quad \text{…3.15}$$

$$M_i = \gamma_{20} + \gamma_{21}W_i + u_{2i} \quad \text{…3.16}$$

The difference between the value of γ_{01} in Equation 3.15 and the estimate of γ_{01} in the original unmediated launch model (i.e. Equation 3.9) represents the indirect effect c – c' in Figure 3.9. The product γ_{02} x γ_{21} is the product a x b (Figure 3.9). In single-level mediation, c – c' is equivalent to a x b. In multilevel mediation, they are not exactly the same. However, the discrepancy is small in larger samples and not likely to be problematic (Krull and Mackinnon, 2001). Deciding which estimate to use may be based on practical considerations, for example, "in a model including more than one mediated effect, use of the a x b estimator provides a separate estimate of each of the mediated effects, while the c – c' estimator gives only an estimate of the total mediated effect" (Krull and Mackinnon, 2001: 260). A model where the antecedent and the mediator are both measured at level-2 and the outcome variable is measured at level-1 is referred to a 2-2-1 model (Zhang et al., 2009), which is the type of mediation model described here. Examples of the 2-2-1 mediation mechanism can be found in Chen et al. (2007), Pituch et al. (2006) and Krull and MacKinnon (2001). In 2-2-1 models, the main predictor and mediator variables are grand mean centred, which is equivalent to centring within context. In following this 2-step procedure, the specific questions to address would be: does the amount of variance explained increase when the mediator variable in included? How does the size and significance of the wealth effect on the intercept change with the inclusion of the mediator?

3.5.5 Mediated change-to-change model

Equation 3.17 specifies the multilevel mediation model for a within-level main predictor (level-1) and a within-level mediator to describe within-subject variability. The change-to-change model is hypothesised if the null and launch model estimates reveal significant slope variance to be explained. In this model, the mediator variable, M_{ti}, is added as a covariate in the level-1 model along with wealth, $Wealth_{ti}$. Age is measured in months and is centred at the starting value to observe progress from the children's initial starting points. The wealth and the mediator variables are centred on their grand mean values. The variances of the error terms are assumed to be heterogeneous across time points and the covariance terms are set to zero.

$$PPVT_{ti} = \beta_{0i} + \beta_{1i}AGE_{ti} + \beta_{2i}W_{ti} + \beta_{3i}M_{ti} + \varepsilon_{Pti} \quad \text{...3.17a}$$

$$M_{ti} = \beta_{4i} + \beta_{5i}W_{ti} + \varepsilon_{M_{ti}} \quad \text{...3.17b}$$

$$\varepsilon_{ti} \sim N\,(0,\sigma_{\varepsilon}^{2}) \quad \text{...3.17c}$$

At level 2, the intercepts and slopes are allowed to vary randomly between children around the group mean. An unstructured covariance relationship is assumed for the random effects (u_{ji}) to obtain estimates for the covariance terms. The random effects allow for individual differences in the intercept in relation to the grand mean. The intercept, adjusted for wealth, is denoted as γ_0, and γ_4 is the expected value of the mediator adjusted for wealth. The parameter γ_1 represents the expected rate of change per period, adjusted for wealth and the mediator. Similarly, γ_2 is the mean effect of wealth on ability controlling for the mediator, and γ_3 is the mean effect of the mediator on ability controlling for wealth. The final parameter, γ_5, is the mean effect of wealth on the mediator. The covariance between two parameters is zero if they are estimated as fixed (e.g. $\beta_{ji} = \gamma_j$) as opposed to random.

$$\beta_{0i} = \gamma_0 + u_{0i} \quad \text{…3.18a}$$
$$\beta_{1i} = \gamma_1 + u_{1i} \quad \text{…3.18b}$$
$$\beta_{2i} = \gamma_2 + u_{2i} \quad \text{…3.18c}$$
$$\beta_{3i} = \gamma_3 + u_{3i} \quad \text{…3.18d}$$
$$\beta_{4i} = \gamma_4 + u_{4i} \quad \text{…3.18e}$$
$$\beta_{5i} = \gamma_5 + u_{5i} \quad \text{…3.18f}$$
$$u_{ji} \sim N\,(\mu, \sigma^2) \quad \text{…3.17c}$$

3.6 Ethical considerations

The Young Lives project obtained ethics approval from the University of Oxford in 2006. Fieldworker manuals were developed to provide guidance on ethical research procedures, emphasising special considerations in working with children and families living in very poor conditions. The fieldworkers then underwent an extensive training on research ethics. Young Lives field researchers are also required to record any ethics questions or dilemmas in their field notes, which are then reported back to the team leaders. Further details are provided in *Morrow, Virginia (2009). The Ethics of Social Research with Children and Families in Young Lives: Practical Experiences*. Young Lives Working Paper No. 53, Department of International Development, University of Oxford: Oxford.

The process of seeking consent from children was guided by two key principles: (1) taking the time to make sure that the children understand what they are agreeing to; (2) and ensuring that their involvement is voluntary (Young Lives, 2009). In working with children, **informed consent** requires researchers to take the time to explain their research in a way that can be understood by children, including the role of the enumerator, why the data are being collected and how it will be used. Similarly, explaining the concepts of *confidentiality* and *anonymity* using appropriate words is important. The children need to know that data collectors will only share the information they provide with members of the research team, and not with their family members, teachers, or other members of the community. They should also know that their names will never be used to identify or describe them. Moreover, children and young people may find it difficult to refuse adults, particularly those who are perceived to be in positions of authority. **Voluntary consent** requires researchers to make it clear to children that there are no negative consequences for refusing to participate. Children should not be pressured in any way and should be aware that they can opt out whenever they wish. Similarly, there should be no suggestion that something might be gained by participating. Researchers were further trained on child protect and safety standards when doing research with children such as never spending time alone with children, never harming a child, never engaging in physical contact and never using inappropriate language (Young Lives, 2009: 13-14).

Although the D.Phil thesis is based on secondary data, ethics approval was obtained through the Departmental Research Ethics Committee (DREC). CUREC approval for the study was received on 24 April 2013. The data used for the study were obtained from the UK Data Service. Additional steps have also been taken to ensure proper use of the data as follows:

1. *Fit for purpose*: A research proposal was approved by Young Lives, confirming project relevance and appropriateness of the research aims and questions in relation to the original Young Lives study, ensuring alignment with the research

purposes intended by Young Lives. Efforts have been made to ensure that the final set of research questions reflect the strengths and limitations of the data.

2. *Data protection*: The anonymised datasets are matched using unique identifiers and stored in their original format.
3. *Transparency and rigour*: Limitations about using secondary data – conceptual, methodological, and practical are documented in the thesis.
4. *Data problems*. Errors identified in data sets regarding data coding/entry or concerns about data use, are communicated openly with the Young Lives project.

Chapter 4 Psychometric properties of the Peabody Picture Vocabulary Test

4.1 Introduction and rationale

This chapter addresses the first set of research questions about the performance of the Peabody Picture Vocabulary Test (PPVT) when used across the multilingual sample of children over time. The aim is to establish a valid and reliable measure of the children's vocabulary abilities in each language group for the purpose of measuring changes in ability over time. This component of the research contributes to the emerging literature on the need for a careful and critical approach to test design and score interpretation. Reliability and validity are analysed within a Rasch measurement framework. The more restrictive Rasch model is considered better suited for establishing the validity of test scores compared with alternative paradigms due to its emphasis on individual item functioning, rather than overall test performance.

Items that are less effective in assessing vocabulary are discussed to inform researcher understanding of the validity of construct when applied to different language groups. Items are flagged if they do not fit the Rasch model based on the statistical criteria or if they show high statistical dependency. A key limitation in the analysis is that in Round 4, when the younger cohort children were 12 years old, three-quarters of the items were eliminated from the survey due to poor data-model fit. The items were also discarded from the fifth and final round of data collection. Therefore, the results presented in this chapter pertain to the subset of 55 items available across rounds. The full set of 204 PPVT items were analysed for Rounds 2 and 3 (age five and eight years). The main findings of the analysis are summarised in Appendix 2. The Item Characteristic Curves for the reduced scale off 55 items are presented in Appendix 3.

The psychometric analysis is extended to investigate item stability over time to determine whether the scores from the repeated test can be reliably calibrated onto a single metric. The reason for this component of the research is based on the rationale that person-time dependencies may affect the measurement of change due to response dependency over time (Smith, 2005: Marais, 2009), or changes in item functioning over time (Wright, 1996). Wright (1996) suggested using separate Rasch analyses for each time point followed by a test equating procedure. Anchoring item calibrations on the most stable items is thought to produce more reliable person ability estimates and thus a more accurate description of person change (Wright, 1996). Alternatively, when repeated measures are used, the data can be stacked/pooled by treating persons at each time point as distinct (Marais, 2009). In the current analysis, person-time dependency is investigated by examining Differential Item Functioning (DIF) of each item by time point. The extent of dependency will inform the scaling procedure for placing the scores from the repeated test on the same metric: stacking/pooling vs. equating/anchoring. If dependency is minimal, the stacking procedure will be used to obtain the estimates. If dependency is widespread, the most stable items will be used as anchor items for linking the tests. In other words, the item difficulties will be fixed on the items that produce the most consistent difficulty estimates across time.

The RUMM 2030 statistical software is used to perform the test of fit analysis (Andrich, Sheridan and Luo, 2014). The results are organised and presented as follows:

1. Overall fit of the items to the model;
2. Item targeting;
3. Order and location of the items;
4. Differential item functioning by language;
5. Individual item fit;
6. Response dependency and multidimensionality;
7. Scale stability.

In the final section 4.4, the validity of the Rasch scores is discussed by (i) referring back to the psychometric analysis; (ii) addressing the problem of guessing; (iii) comparing the Rasch scores with the original raw scores as well as external measures of achievement available in the Young Lives data structure; and (iv) verifying the scores conform with theoretical expectations. In this final part of this section, the Rasch scores are also transformed onto a scale with a mean of 500 and a standard deviation of 100 in order to avoid negative achievement scores in the subsequent growth curve analyses.

4.2 Characteristics of the item responses

The data set involves a large proportion of missing data relating to the items that were not administered. Item responses are missing as a result of being too hard for the child or because the items were below the basal range. How to handle skipped items is a decision that depends on the purpose of the test. Given the structure of the PPVT, not reached items are rescored "0" because the reason for skipping the item is that the child is not likely to be successful on it given their performance on earlier items; items that fall below the basal range are rescored "1" (Leon et al., 2015; Embretson and Harshberger, 1999). Moreover, although the RUMM2030 software can handle missing data, "if missing data is present in a data set then the allocation of persons to the set of class intervals will vary from item-to-item. Under-representation in one or more class intervals within an item can lead to unreliable estimates of the Chi-square statistic" (Andrich et al., 2009: 17).

Table 4.1 shows that most of the children in the sample chose to respond in one of the three most widely spoken languages in Ethiopia: Amharic, Oromifa and Tigrinya. Around 300 children responded in "other" languages. Only responses in the three main languages are considered for the analysis. The "other" language group category is not meaningful for results interpretation. Decomposing the category is not possible at every sweep. When it is possible, the result would be several very small subsamples.

Table 4.1 Languages used during test administration

	Amharic		Oromifa		Tigrinya		Total
Age	n	%	n	%	n	%	n
Five	798	54	287	19	371	25	1456
Eight	754	51	347	23	377	25	1478
Twelve	769	52	333	23	377	25	1479
Fifteen	692	51	279	21	369	27	1340

A total of 1,321 children who completed four rounds of the PPVT at ages five, eight, 12 and 15 years also responded consistently in the same language across rounds. For several children, the testing language was coded as missing once. If consistent language data was available at other time points, the missing information is recoded to match. Dropped from the sample are children who responded in different languages across rounds, and children for whom language data was missing for two or more rounds. The resulting sample of 1,321 children consists of 692 children who responded in Amharic, 260 children who responded in Oromifa and 369 Tigrinya-speaking children.

4.3 Psychometric analysis

4.3.1 Overall fit

The mean and SD of the item and person fit residuals provide preliminary information about the overall fit. The residuals should have distributions that are close to normal (M≈0, SD≈1).

Round 2, age 5. Rasch analysis of the test data at age five showed some data-model misfit overall. The mean and SD of the item fit residuals show minor deviation from model expectations ($\bar{z}_i$ =-0.21; SD = 0.43). The mean and SD of the person residuals deviate substantially ($\bar{z}_n$ = -1.37; SD= 3.69), indicating misfit among respondents. However, the high Person Separation Index (r_β = 0.92) suggests that, overall, the instrument distinguishes well between high and low performers.

Round 3, age 8. The mean item fit residual deviates marginally from model expectations ($\bar{z}_i$ = -0.54; SD = 0.99) and there is an improvement in the mean of the person residuals ($\bar{z}_n$ = -0.90; SD = 2.73). The Person Separation Index (r_β = 0.95) is very high, showing abilities spreads well across the continuum of items. In terms of overall performance, the test works better for this age group.

Round 4, age 12. The test works less well for 12-year olds. Person misfit persists ($\bar{z}_n$ = -0.55; SD = 3.09). However, there is no major item misfit as evidenced by an item fit residual that is close to standard normal ($\bar{z}_i$ = -0.21; SD = 0.83). The Person Separation Index drops but the coefficient indicates that the test distinguishes between test-takers (r_β = 0.87).

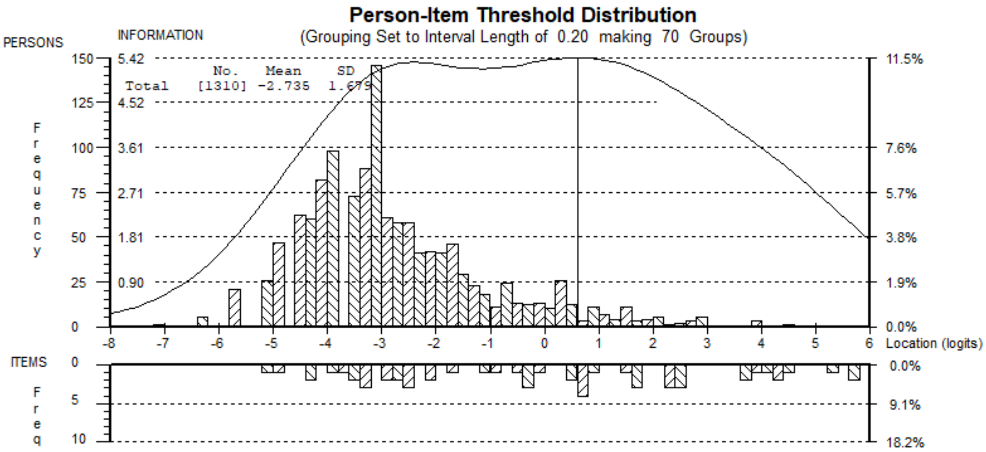
Round 5, age 15. The test works least well for this age group. After removing two items with no responses, overall fit statistics are like previous rounds ($\bar{z}_n$ = -0.57; SD = 2.73 and $\bar{z}_i$ = -0.15; SD = 0.53). However, person separation drops further (r_β = 0.85).

4.3.2 Item targeting

The person-item distribution plots the overall distribution of person abilities (upper histogram) and item difficulties (lower histogram) on the same continuum. This distribution shows whether there are sufficient items for the range of student proficiencies. Harder items and more able persons are located towards the right of the distribution, easier items/less able persons are towards the left. A test is well targeted if the average of item difficulties is about the same as the average of the students' abilities and the item difficulties are evenly spread across the ability distribution (OECD, 2012: 222).

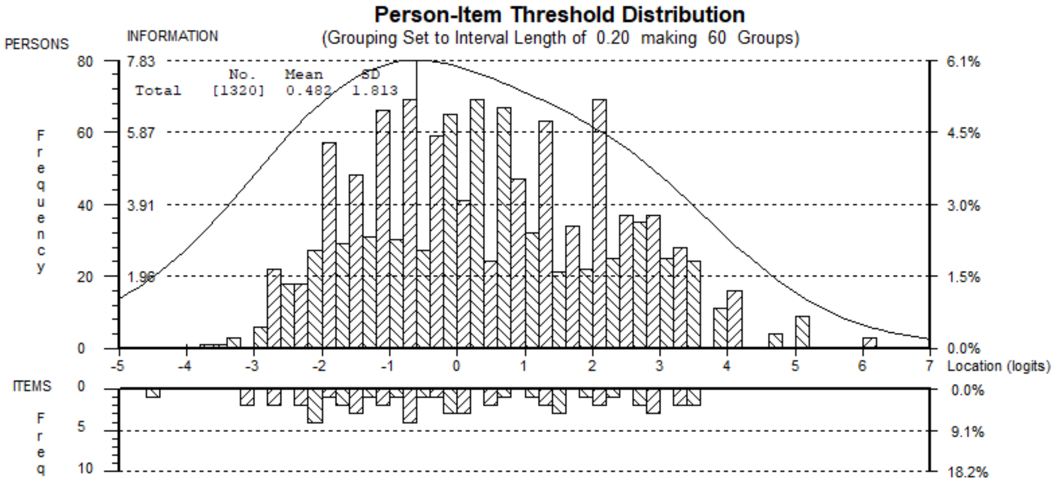
Round 2, age 5. Figures 4.1 - 4.4 shows a wide spread of items relative to persons, -7.0 to +5.0 logits. However, the majority of children are located at the lower end of the scale, indicating the test does not differentiate between students at the lower end of the ability distribution. The information function peaks close to 1.0 logit, which means the test is targeting children with a higher ability. The mean person location is far from the mean item location of zero ($\bar{\beta}_n$ = -2.74; SD = 1.70) and there are not enough easy items to measure ability among the least able children. The results suggest that the variance in the children's initial starting points will be slightly underestimated in the growth models.

Figure 4.1 Person-item distribution, age five



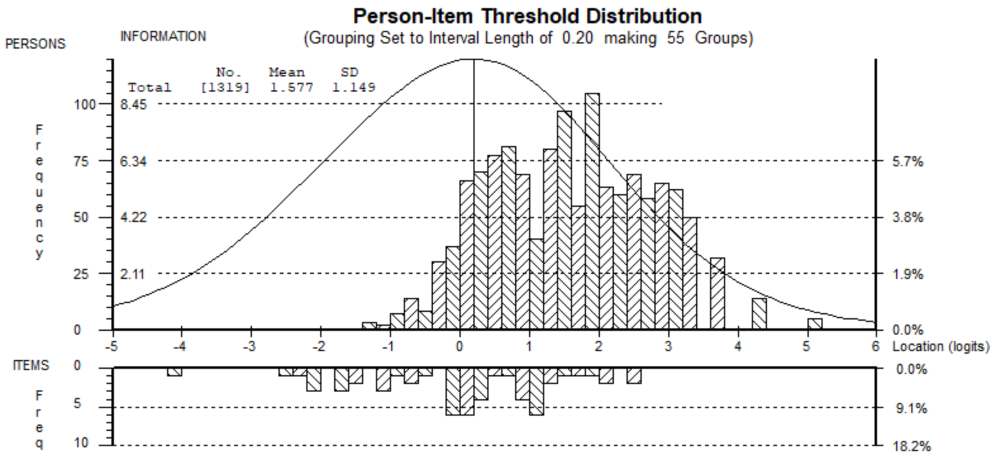
Round 3, age 8. The distribution shows a wide spread of persons on the continuum relative to the items, -4.0 to 6.0 logits. The information function peaks at around -0.6 logits, indicating that the test targets lower-to-moderate ability children in this sample. The distribution is slightly skewed to the right ($\bar{\beta}_i$ = 0.48; SD = 1.81). There is a small ceiling effect in that some children were able to answer the hardest items, suggesting that the test may not adequately measure high proficiency children in the next rounds.

Figure 4.2 Person-item distribution, age eight



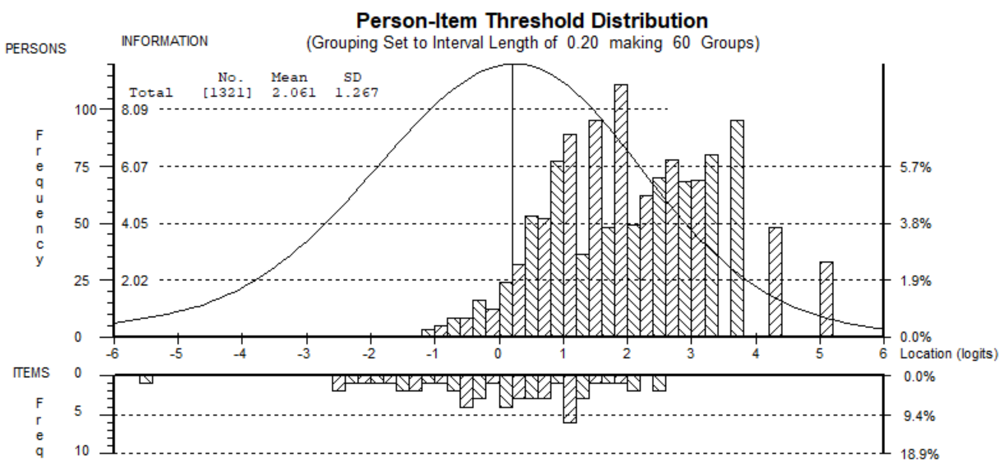
Round 4, age 12. The person distribution is skewed towards the right ($\bar{\beta}_i$ = 1.58; SD = 1.15), and the information function peaks at 0.2 logits. Thus, a large proportion of more able children are not adequately measured by the scale. The variance in person abilities is underestimated at higher ability levels with implications for growth modelling.

Figure 4.3 Person-item distribution, age 12



Round 5, age 15. By age 15, the distribution is substantially skewed towards the right, with a large proportion of children not adequately assessed by the scale. The mean person location is well below the mean item location ($\bar{\beta}_n$ = 2.06; SD = 1.27). The test is targeting a level of ability that is far below the range of abilities in the sample. The implication for growth modelling is that the variance in growth rate will be underestimated, and the shape of the growth curves may be flatter than expected.

Figure 4.4 Person-item distribution, age 15



4.3.3 Order and location of items

The item map places the difficulty estimate of each item on the same scale as the ability estimate of each child. This is a graphical representation of the rank ordering of items and individuals on the same scale. It highlights gaps along the continuum of items. It also illustrates whether the ordering of item difficulties is consistent with the ordering of the items on the test, providing further evidence of test quality. For a well-constructed test, easier items should be asked earlier in the test and harder items later on.

<u>Round 2, age 5</u>. Figure 4.5 shows that most of the items are working as expected. However, there are several gaps along the continuum indicating levels of ability that are not targeted by the items on the test. Additional easier and moderately easier items would have improved the differentiation between children's abilities at lower proficiency levels.

<u>Round 3, age 8</u>. Figure 4.6 shows that the item ranking at age eight is consistent with the previous ranking at five years. Most of the earlier items on the test are at the bottom of the map and the harder items are towards the top. There are a few gaps in the continuum in that some moderate-to-higher ability children are not adequately assessed. Previous evidence of a small ceiling effect at eight years is supported by a gap above 3.5 logits.

<u>Round 4, age 12</u>. There are gaps along the continuum at low and high ability levels as illustrated in Figure 4.7. Modelling change in children's abilities will be affected this limitation in that the change measure will be underestimated in terms of the slope and the curvature. Some of the easier items on the test appear harder than anticipated, given their location of the test (e.g. Item 2). While the location of the items is expected to change with repeated measures, the relative rank ordering should remain the same.

<u>Round 5, age 15</u>. The item map (Figure 4.8) illustrates a large ceiling effect at 15 years. There are no items targeting children's abilities above 2.4 logits. It is also evident that many items are far too easy for the children in the sample.

Figure 4.5 Item map, age five

```
------------------------------------------------------
LOCATION          PERSONS     ITEMS [locations]
------------------------------------------------------
  6.0                     |
                          |
                          | I0054  I0055
                          |
                          | I0051
  5.0                     |
                          |
                          |
                          | I0053
                          | I0052  I0049
  4.0                     | I0048
                          | I0045
                          | I0047  I0050
                          |
                          |
  3.0                     |
                        x |
                          |
                          | I0041  I0044  I0043
                          | I0039  I0042  I0046
  2.0                   x |
                        x |
                          | I0038  I0040  I0035
                        x | I0037
                        x |
  1.0                   x |
                        x | I0033
                          | I0036  I0029  I0031  I0030
                       xx | I0034  I0032
                      xxx |
  0.0                   x |
                       xx | I0026
                       xx | I0023  I0028  I0027
                       xx | I0022
                      xxx |
 -1.0                   x | I0020
                       xx | I0024
                      xxx |
                     xxxx |
                   xxxxxx | I0025
 -2.0               xxxxx |
                    xxxxx | I0018  I0019
                    xxxxx |
                  xxxxxxx | I0008  I0017  I0021
                  xxxxxxx | I0015  I0011
 -3.0            xxxxxxxx | I0014  I0013
        xxxxxxxxxxxxxxxxx |
              xxxxxxxxxxx | I0009  I0012  I0016
                xxxxxxxxx | I0002  I0010
                          | I0003
 -4.0        xxxxxxxxxxxx | I0004
               xxxxxxxxxx |
                 xxxxxxxx | I0006  I0007
                 xxxxxxxx |
                          |
 -5.0              xxxxxx | I0005
                      xxx | I0001
                          |
                          |
                      xxx |
 -6.0                     |
                          |
                        x |
                          |
                          |
 -7.0                     |
------------------------------------------------------
x = 8 Persons
```

Figure 4.6 Item map, age eight

```
-----------------------------------------------------------
LOCATION          PERSONS      ITEMS [locations]
-----------------------------------------------------------
  7.0                         |
                              |
                              |
                              |
                              |
  6.0                       x |
                              |
                              |
                              |
                              |
  5.0                      xx |
                              |
                            x |
                              |
                              |
  4.0                    xxxx |
                          xxx |
                              |
                       xxxxxx | I0055  I0053
                      xxxxxxx | I0051  I0049
  3.0                  xxxxxx |
                    xxxxxxxxx | I0048  I0054  I0045
                    xxxxxxxxx | I0047  I0052
                    xxxxxxxxx |
                       xxxxxx | I0039
  2.0       xxxxxxxxxxxxxxxxx | I0042  I0050
                       xxxxxx | I0046
                    xxxxxxxxx |
                        xxxxx | I0043  I0041  I0044
             xxxxxxxxxxxxxxxx | I0040  I0038
  1.0                xxxxxxxx | I0035
                 xxxxxxxxxxxx |
            xxxxxxxxxxxxxxxxx | I0037
                       xxxxxx | I0030  I0036
            xxxxxxxxxxxxxxxxx |
  0.0              xxxxxxxxxx | I0023  I0033  I0022
             xxxxxxxxxxxxxxxx | I0031  I0029  I0026
              xxxxxxxxxxxxxxx | I0020
                      xxxxxxx | I0019
            xxxxxxxxxxxxxxxxx | I0008  I0032  I0028  I0024
 -1.0                xxxxxxxx | I0017
            xxxxxxxxxxxxxxxxx | I0034  I0015
                     xxxxxxxx | I0018
                 xxxxxxxxxxxx | I0009  I0014  I0027
                      xxxxxxx | I0011  I0016
 -2.0          xxxxxxxxxxxxxx | I0003
                      xxxxxxx | I0021  I0025  I0012  I0002
                        xxxxx | I0004  I0013
                        xxxxx |
                       xxxxxx | I0010  I0006
 -3.0                      xx |
                              | I0005  I0007
                            x |
                              |
                              |
 -4.0                         |
                              |
                              |
                              | I0001
                              |
 -5.0                         |
-----------------------------------------------------------
x = 4 Persons
```

Figure 4.7 Item map, age 12

```
----------------------------------------------------------------------
LOCATION          PERSONS     ITEMS [locations]
----------------------------------------------------------------------
  6.0                        |
                             |
                             |
                             |
                             |
  5.0                      x |
                             |
                             |
                             |
                          xx |
  4.0                        |
                             |
                       xxxxx |
                             |
                    xxxxxxxx |
  3.0             xxxxxxxxxx |
                 xxxxxxxxxxx |
                  xxxxxxxxxx |
                xxxxxxxxxxxx | I0042  I0039
                  xxxxxxxxxx |
  2.0            xxxxxxxxxxx | I0047  I0051
          xxxxxxxxxxxxxxxxxx | I0045
                   xxxxxxxxx | I0055
            xxxxxxxxxxxxxxxx | I0048
               xxxxxxxxxxxxx | I0022  I0053
  1.0                xxxxxxx | I0029  I0037  I0038  I0049  I0035  I0043
                xxxxxxxxxxxx | I0008  I0036  I0054  I0023
              xxxxxxxxxxxxxx | I0046
               xxxxxxxxxxxxx | I0024
                xxxxxxxxxxxx | I0041  I0052  I0017  I0040
  0.0            xxxxxxxxxxx | I0026  I0020  I0031  I0050  I0002  I0014
                      xxxxxx | I0030  I0015  I0033  I0009  I0044  I0019
                       xxxxx |
                           x | I0018
                          xx | I0016  I0032
 -1.0                      x | I0006
                             | I0012  I0028  I0013
                           x |
                             | I0011  I0034
                             | I0027  I0003  I0004
 -2.0                        |
                             | I0007  I0021  I0025
                             | I0005
                             | I0010
                             |
 -3.0                        |
                             |
                             |
                             |
                             |
 -4.0                        |
                             | I0001
                             |
                             |
                             |
 -5.0                        |
----------------------------------------------------------------------
x = 6 Persons
```

Figure 4.8 Item map, age 15

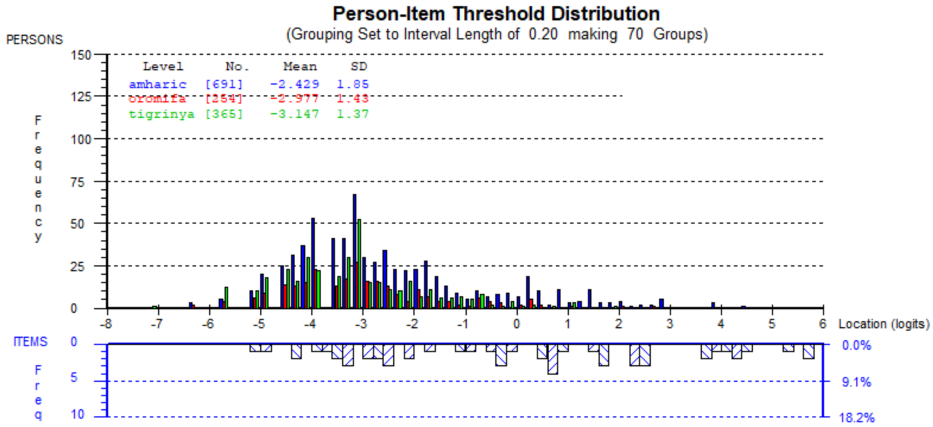
```
-----------------------------------------------------------------------
LOCATION          PERSONS     ITEMS [locations]
-----------------------------------------------------------------------
  6.0                      |
                           |
                           |
                           |
                           |
  5.0               xxxxxx |
                           |
                           |
                           |
                  xxxxxxxx |
  4.0                      |
                           |
          xxxxxxxxxxxxxxxx |
                           |
             xxxxxxxxxxxxx |
  3.0         xxxxxxxxxxxx |
               xxxxxxxxxxx |
             xxxxxxxxxxxxx |
              xxxxxxxxxxxx | I0039  I0042
                xxxxxxxxxx |
  2.0             xxxxxxxx | I0045  I0047
       xxxxxxxxxxxxxxxxxxx | I0055
                  xxxxxxxx | I0051
          xxxxxxxxxxxxxxxx | I0022
                    xxxxxx | I0023  I0029  I0053
  1.0      xxxxxxxxxxxxxxx | I0043  I0008  I0037  I0048  I0038  I0035
             xxxxxxxxxxxxx | I0049
                 xxxxxxxxx | I0046  I0036  I0054
                 xxxxxxxxx | I0024  I0041  I0031
                     xxxxx | I0052  I0017  I0040
  0.0                 xxxx | I0009  I0033  I0002  I0014
                        xx | I0019
                       xxx | I0015  I0050  I0020
                         x | I0044  I0026  I0030  I0018
                         x | I0016  I0013
 -1.0                    x | I0032
                         x | I0028
                           | I0034  I0012
                           | I0027  I0011
                           | I0004
 -2.0                      | I0007
                           | I0021
                           | I0005
                           | I0025  I0010
                           |
 -3.0                      |
                           |
                           |
                           |
                           |
 -4.0                      |
                           |
                           |
                           |
                           |
 -5.0                      |
                           |
                           |
                           | I0001
                           |
 -6.0                      |
-----------------------------------------------------------------------
x = 6 Persons
```

4.3.4 Differential Item Functioning

Differential Item Functioning (DIF) is a form of item bias which occurs when different response patterns emerge for different subgroups of respondents with the same level of proficiency. If DIF is identified in the data, the scale is not invariant/unbiased across groups and the items in the scale should not be treated as the same for different subgroups. DIF is sensitive to sample size in that large sample analysis could result in an erroneously significant DIF effect. Thus, the pattern of responses by language is examined using the Bonferroni-adjusted probability value which takes into account the problem that chi-square probabilities tend to be small with large samples.

<u>Round 2, age 5.</u> Figure 4.9 shows the person-item distribution by language group to illustrate the differences in the performance between the subgroups at age five. One-way ANOVA shows a statistically significant difference in the mean scores between language groups, with children being assessed higher on average in Amharic (F = 26.09(2); p <0.001). Mean location estimates are -2.43 (SD = 1.85, n= 691) in Amharic, -2.98 (SD = 1.43, n= 254) in Oromifa and -3.15 (SD = 1.37, n=365) in Tigrinya.

Figure 4.9 Person-item distribution by language, age five

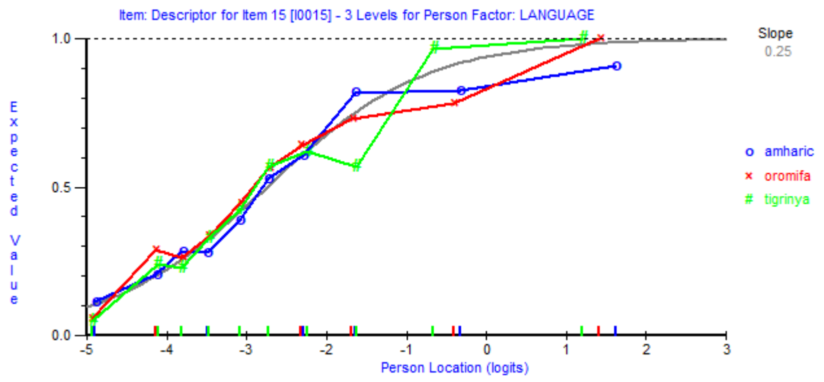


Uniform DIF for language was statistically significant for six items. Non-uniform DIF was significant for 10 items. Figure 4.10 presents an example of an item showing no significant DIF for language, an item showing uniform DIF and one with non-uniform DIF.

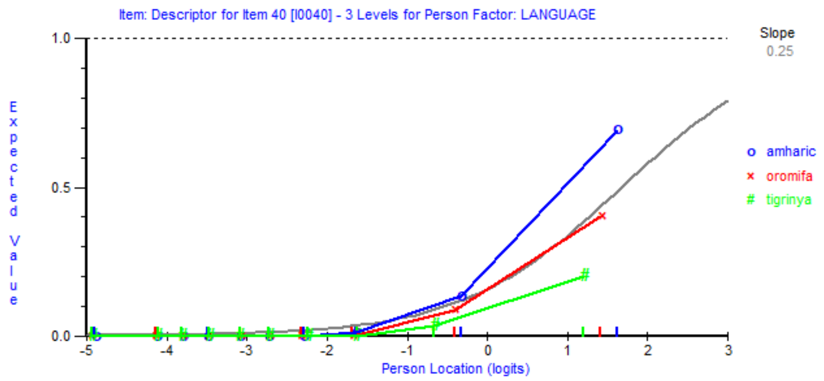
Since DIF affects a substantial proportion of the items, more than one third, a conservative approach to handling language bias is needed to strengthen the validity of the construct.

Figure 4.10 Item Characteristic Curve with and without DIF, age five

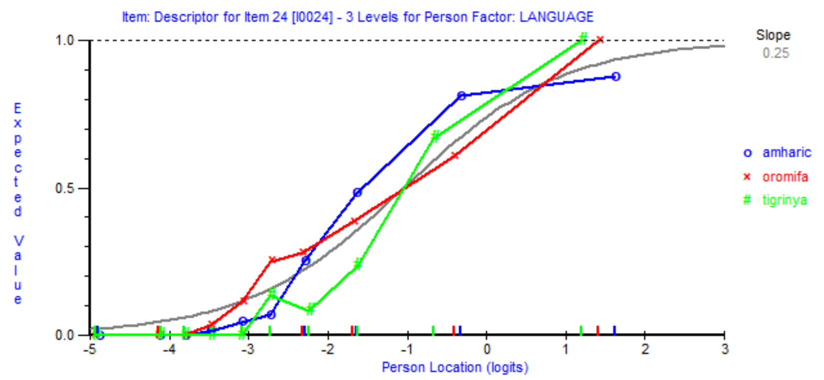
Item with no significant DIF



Item with uniform DIF

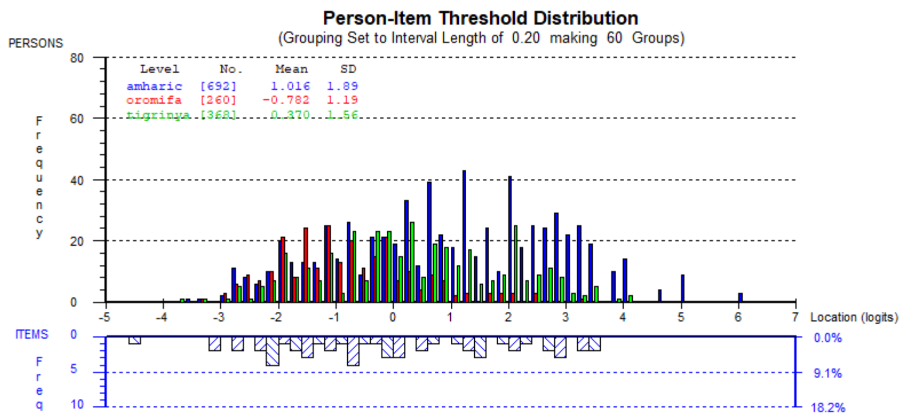


Item with non-uniform DIF



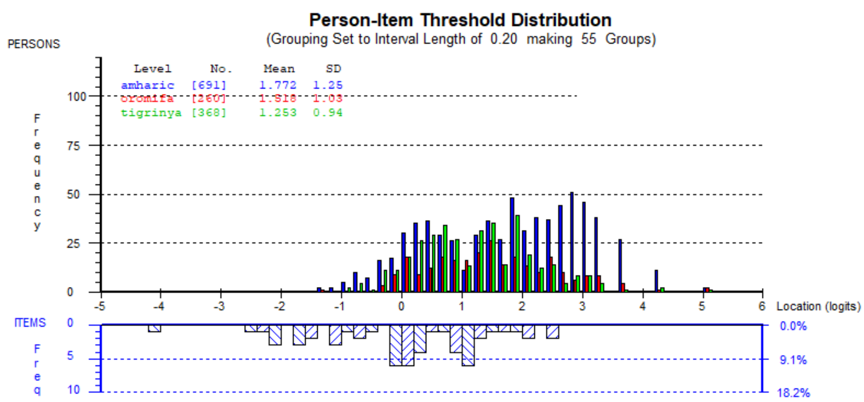
Round 3, age 8. Over half of the items in the scale showed DIF for language, with a large and significant difference in the scores between language group, as shown in Figure 4.11. Children are assessed higher on average in Amharic (F = 109.34(2); p <0.001). Mean location estimates for each subgroup are as follows: Amharic: 1.02 (SD = 1.89, n= 692); Oromifa: -0.78 (SD = 1.19, n= 260); and Tigrinya 0.37 (SD = 1.56, n= 368).

Figure 4.11 Person-item distribution by language, age 8



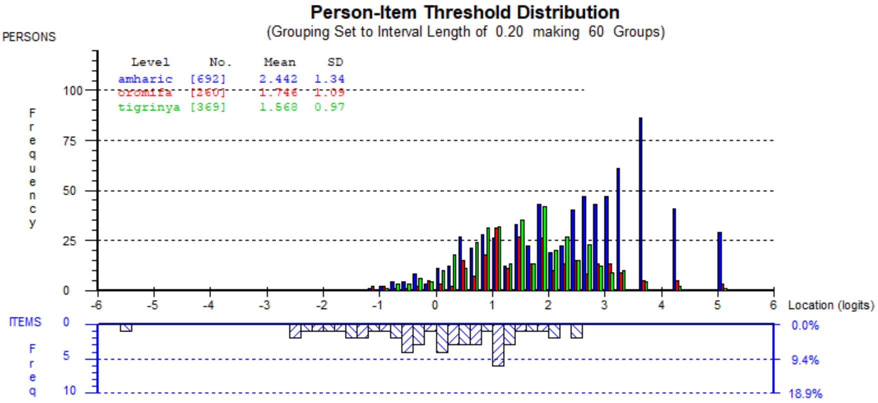
Round 4, age 12. Over half of the items showed statistically significant DIF for language. The vast majority of biased items showed uniform DIF. The results show a significant difference in scores between language groups as shown in Figure 4.12 (F = 25.88(2); p <0.001). Mean location estimates are 1.77 (SD = 1.25, n= 691) in Amharic, 1.52 (SD = 1.03, n = 260) in Oromifa and 1.25 (SD = 0.94, n = 368) in Tigrinya.

Figure 4.12 Person-item distribution by language, age 12



<u>Round 5, age 15</u>. Half of the items show DIF for language. In this final round, all the offending items show uniform DIF. Non-uniform DIF is no longer detectable. The findings, which are illustrated in Figure 4.13, also indicate a significant difference in scores between language groups (F = 74.79(2); p <0.001). Mean location estimates are 2.44 (SD = 1.34, n= 692) in Amharic, 1.75 (SD = 1.09, n = 260) in Oromifa and 1.57 (SD = 0.97, n = 369) in Tigrinya.

Figure 4.13 Person-item distribution by language, age 15



4.3.5 Individual item fit

To reduce the chance of making a Type I error, the Bonferroni correction is used to compute p-values which are interpreted alongside the fit residuals and the ICCs. The power to detect misfit using the chi-square statistic depends on the sample size. The larger the sample, the greater the power. In very large samples, the power is so great that all items misfit, even when the observed and expected values are very close (RUMM2030 Laboratory, 2009). In the current analysis, items are only flagged if the fit statistics are considerably out of range and if the observed values deviate substantially from the ICCs.

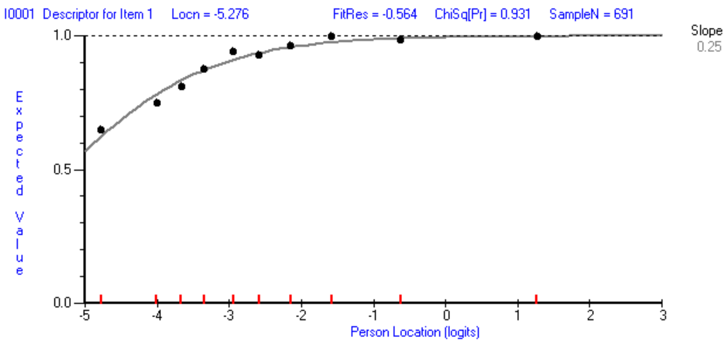
A distinction is made between items that overfit, i.e. high discriminating items and those that underfit, i.e. low discriminating items. Items showing low discrimination do not differentiate well between individual abilities. High discrimination may be associated with response dependency, i.e. redundant items. Positive discrimination is generally viewed as desirable unless it is so high that the item repeats information provided by other items on the test (Kelley et al, 2002). Masters also (1988) notes,

> High item discrimination can be a symptom of a special kind of measurement disturbance introduced by an item that gives persons of high ability a special advantage over and above their higher abilities. This type of disturbance, which can be interpreted as a form of item "bias," can be encouraged by methods that routinely interpret highly discriminating items as the "best" items on a test and may be compounded by procedures that weight items by their discrimination. The type of measurement disturbance described and illustrated in this paper occurs when an item is sensitive to individual differences on a second, undesired dimension that is positively correlated with the variable intended to be measured. Possible secondary influences of this type include opportunity to learn, opportunity to answer, and test-wiseness.
>
> (Masters, 1988: 15)

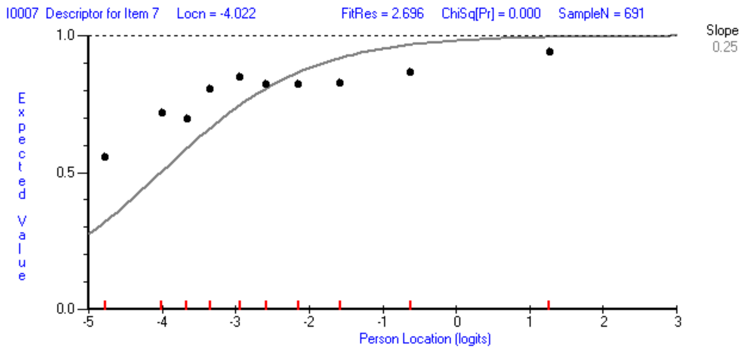
<u>Round 2, age 5</u>. In the Amharic version of the test, 10 items showed data-model misfit based on the fit residual value and the chi-square probability. Six items discriminate poorly between the children. These items are easier items. The remaining four items over-discriminate, as evidenced by larger negative fit residual values and steep ICCs. The overfitting items do not discriminate at lower-to-moderate ability levels but show higher than expected scores for higher proficiency children, as shown in Figure 4.14 (i.e. item showing high discrimination). These items may therefore be working more effectively for some children than for others.

Figure 4.14 Item Characteristic Curve with items showing good and bad fit, age five

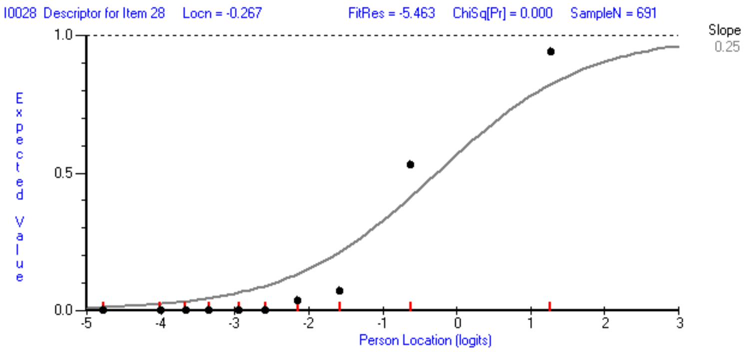
Item showing good data-model fit



Item showing low discrimination



Item showing high discrimination



For the Oromifa and Tigrinya subsamples, there are no deviating items. The harder items on the test were relatively difficult for all children and were left unanswered by most. The hardest nine items had to be removed from the Tigrinya version of the test as nearly all responses were incorrect or missing. These items could not differentiate between the children at age 5 and therefore did not contribute to the measure.

Round 3, age 8. For the Amharic subsample, three of the harder items show high discrimination. As in the previous round, these overfitting items do not discriminate at lower proficiencies but show much higher than expected discrimination at higher proficiencies. The Oromifa version of the test shows no major misfit. However, there are very few responses on the harder test items. For the Tigrinya subsample, the extreme items that could not be entered into the analysis now work well.

Round 4, age 12. Nine items misfit in the Amharic version of the test. Five harder items do not discriminate as well as expected but are needed to maintain the continuum of harder items. There is no misfit in the Oromifa test. For the Tigrinya subsample, Item 1 is removed from the fit analysis since all children answered it correctly.

Round 5, age 15. To analyse item fit among the Amharic subsample, Item 1, which was correctly answered by all children, was removed. Nine items show deviation from model expectations. Half of the misfitting items are low discriminating and half are high discriminating. All the misfitting items target moderate-to-higher abilities. In Oromifa and Tigrinya, none of the items violate model assumptions.

4.3.6 Violations of local independence

The assumption of local independence can be violated in two ways, through response dependency and multidimensionality. An analysis of the item residual correlations reveals items that are redundant, or pairs of responses that are dependent. A principal components analysis provides preliminary evidence of whether a single dimension or multiple dimensions are present in the data.

Round 2, age 5. In Amharic, one pair of the hardest items have high residual correlations. Similarly, in Tigrinya and Oromifa, the hardest items are highly correlated with each other. This finding may be due to few children having attempted these items, and even fewer children answering them correctly. In all three language versions of the test, a principal components analysis of the residuals showed that the factor loadings on PC1 loaded in opposite directions (e.g. positive and negative loadings), suggesting two possible dimensions. The two subsets of items correspond with a set of harder items (i.e. positive factor loadings) and a set of easier ones (i.e. negative loadings). Items with the largest factor loadings are those previously identified as highly discriminating and very difficult. The findings suggest that a second minor dimension is possibly being measured, one that gives higher ability children an advantage on the harder items.

Round 3, age 8. In Amharic, a pair of harder items are highly correlated. Two pairs of very hard items are highly correlated in the Oromifa version of the test. Response dependency does not appear to be a problem in the Tigrinya test. There is evidence of multidimensionality based on an analysis of the pattern of residual factor loadings. The high factor loadings concern the hardest items on the test.

Round 4, age 12. Response dependence is not a problem across the subsamples. No pairs of items show high residual correlations in any of the language versions of the test. As in previous rounds, the factor loadings on PC1 suggest two subsets of items.

Round 5, age 15. Response dependency does not appear to be an issue among any of the language subgroups. As in the previous round, the factor loadings on PC1 load in opposite directions. However, unlike earlier rounds of the test, the coefficients are small.

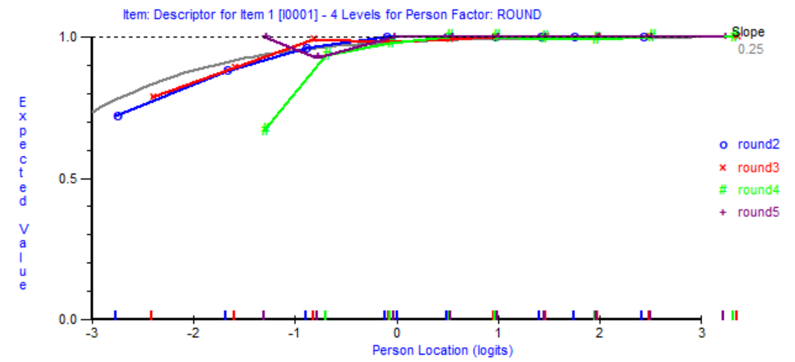
4.3.7 Scale stability

This section investigates the stability of item functioning over time. The purpose of this component of the Rasch analysis is to identify a reliable way of placing the person ability estimates onto a single metric for growth trajectory analysis. This step of the analysis is based on the rationale that the measurement of change may be affected by person-time dependencies (Smith, 2005: Marais, 2009), or changes in item functioning over time (Wright, 1996). This is a general concern when using repeated measures. If dependency is minimal, the scores from the repeated test can be pooled to produce a panel data set. The Rasch model can then be estimated to obtain person ability estimates. If dependency is substantial, the most stable items can be used as anchor items to equate the tests. Person-time dependency is investigated by conducting a Differential Item Functioning (DIF) analysis of each item by time point. The data are stacked to examine DIF of the same item by time point (Marais, 2009; Wright, 1996; also see Barley and Jones, 2006). Large DIF effects by time indicate scale instability.

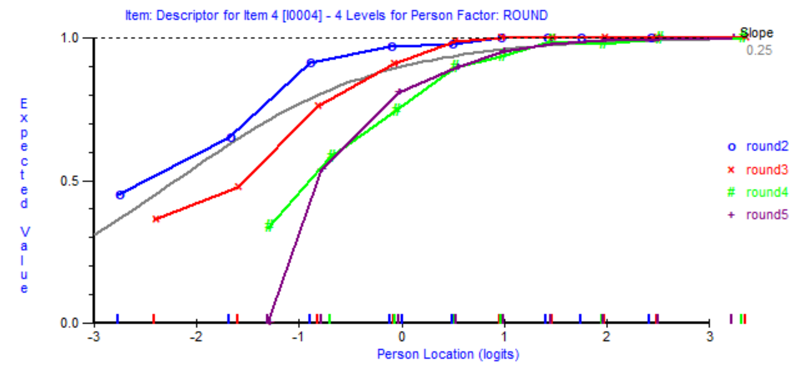
The DIF effect for time is large for all but two of the easiest items on the test. Figure 4.15 presents an example of an item showing no significant DIF for time, an item showing uniform DIF and one with non-uniform DIF. The scale of the DIF effect warrants the use of anchor items to equate the tests across rounds.

Figure 4.15 Item Characteristic Curve with and without DIF for time, pooled sample

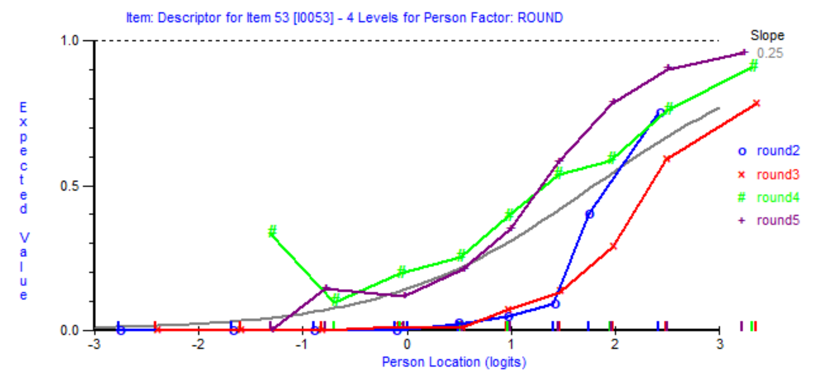
Item with no significant DIF for time



Item with uniform DIF for time

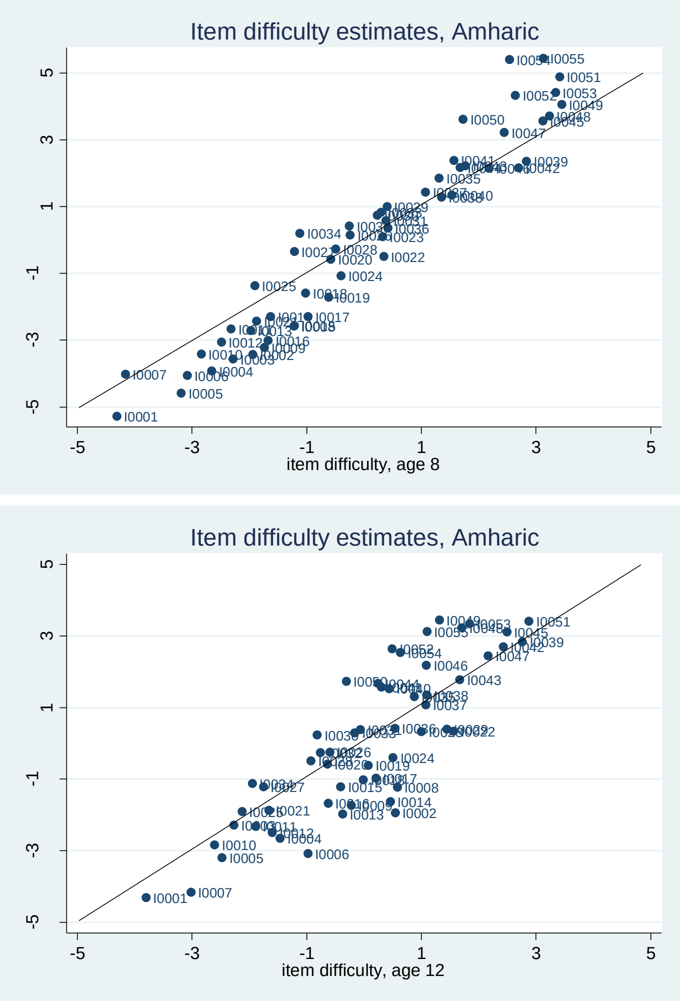


Item with non-uniform DIF for time



Cross plotting the items difficulty estimates obtained at each time point provides a visual indication of the stability of scores. The figures below show that the items evenly scatter along the identity line from -5.0 to +5.0 logits, most notably in Amharic. However, there are differences in the difficulty estimates over time such that several items lie far from the line. The estimates not only scatter most evenly in Amharic, they are also closest to the identity line. Item drift appears to be particularly problematic in the Tigrinya version of the test. The cross plots suggest that the items function most consistently in Amharic. In Oromifa, the items spread progressively less well along the identity line over time. In Tigrinya, item difficulties are dispersed haphazardly. For all subsamples, consistency is strongest in the final two rounds of the survey, between ages 12 and 15 years.

Figure 4.16 Comparison of item difficulty estimates across time, Amharic



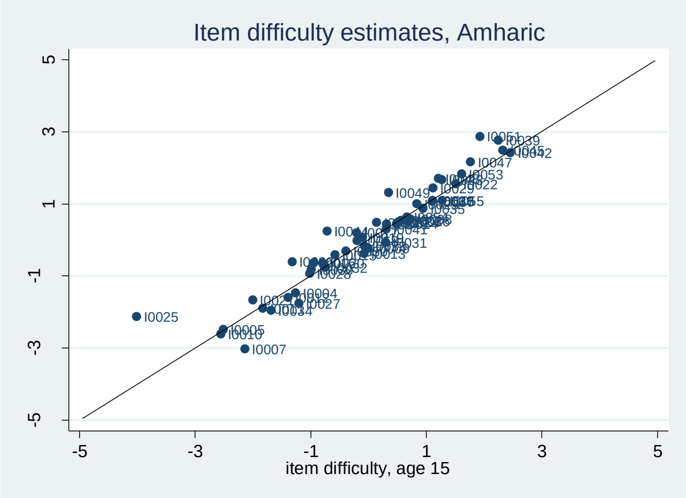
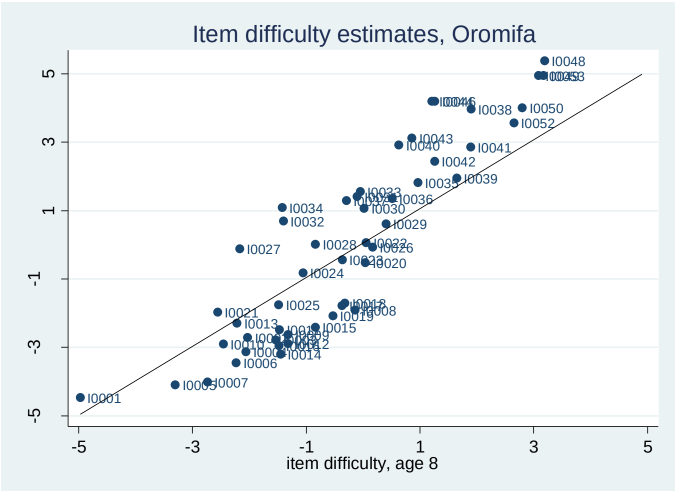
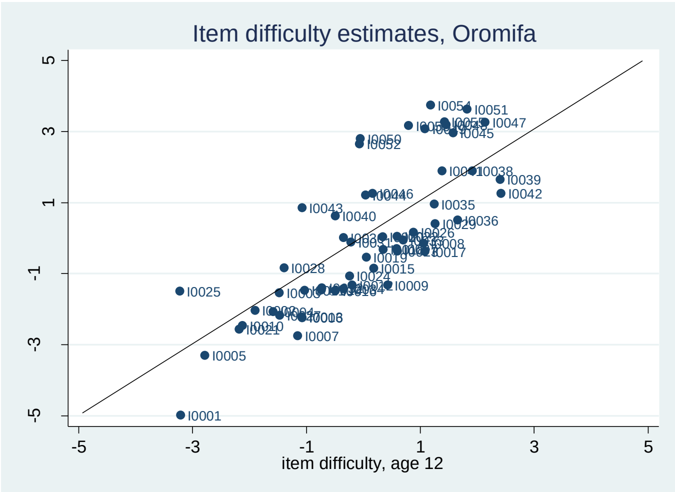


Figure 4.17 Comparison of item difficulty estimates across time, Oromifa





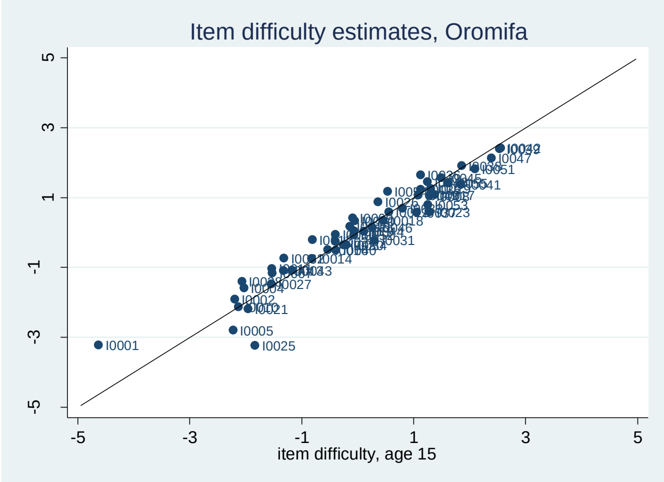
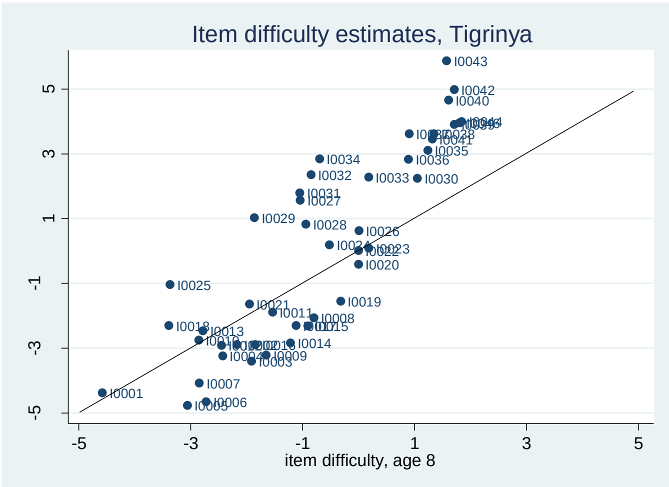
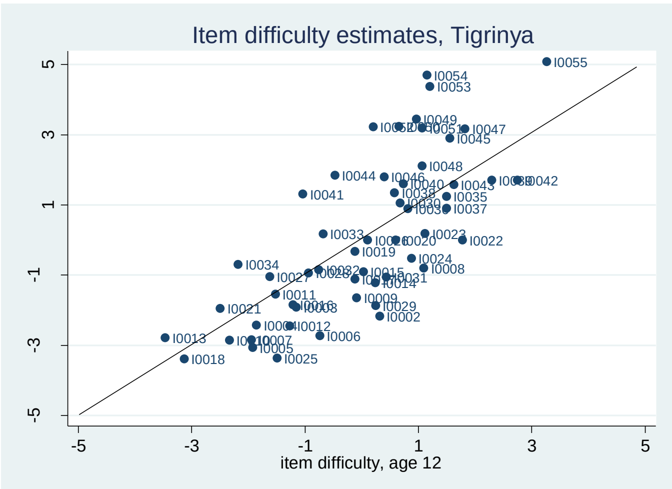
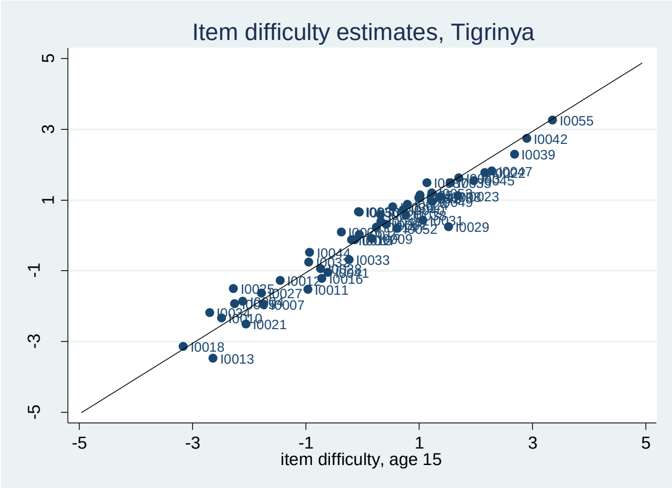


Figure 4.18 Comparison of item difficulty estimates across time, Tigrinya







The validity of the scale is strengthened by the high correlation between difficulty estimates across time, as outlined in the following tables. The largest coefficients are observed for the Amharic speaking subsample. The correlations follow a simplex pattern. Correlations are stronger with more proximate measures (i.e. item difficulty estimates at age five are more highly correlated with estimates obtained at age eight than age 12). The highest correlation occurs between the age 12 and age 15 estimates. This is explained by similarities in children's performance between the two rounds of testing, i.e. the ceiling effect.

Table 4.2 Correlations between item difficulty estimates, Amharic

	Item diff. age 5	Item diff. age 8	Item diff. age 12	Item diff. age 15
Item diff. age 5	1.00			
Item diff. age 8	0.97*	1.00		
Item diff. age 12	0.72*	0.85*	1.00	
Item diff. age 15	0.62*	0.76*	0.95*	1.00

*Statistically significant at the 5% level.

Table 4.3 Correlations between item difficulty estimates, Oromifa

	Item diff. age 5	Item diff. age 8	Item diff. age 12	Item diff. age 15
Item diff. age 5	1.00			
Item diff. age 8	0.90*	1.00		
Item diff. age 12	0.59*	0.79*	1.00	
Item diff. age 15	0.60*	0.79*	0.96*	1.00

*Statistically significant at the 5% level.

Table 4.4 Correlations between item difficulty estimates, Tigrinya

	Item diff. age 5	Item diff. age 8	Item diff. age 12	Item diff. age 15
Item diff. age 5	1.00			
Item diff. age 8	0.87*	1.00		
Item diff. age 12	0.55*	0.75*	1.00	
Item diff. age 15	0.55*	0.73*	0.97*	1.00

*Statistically significant at the 5% level.

Overall, the results suggest that to improve the estimation of developmental trajectories the separate test analyses should be followed by a test equating procedure. Item calibrations are anchored on the most stable items in the scale, those which are located closest to the identity line excluding items at the extreme ends of the distribution. Common item anchoring is thought to produce more reliable person ability estimates and a more accurate description of person change (Wright, 1996).

Tables 4.5 – 4.7 list the items that are used for test equating. The items are presented in terms of the fixed item difficulty estimates and point differences. The item location is fixed on the Rasch estimates obtained at the first round of testing (age five). Prior to identifying suitable anchors, extreme items were removed. Item estimates are most stable in Amharic resulting in more anchor items. The hardest items are the least stable in all three languages, and particularly so in Tigrinya.

Table 4.5 Change in item difficulty estimates, Amharic PPVT

Item no.	Fixed location/difficulty	Difference R2 to R3	Difference R3 to R4	Difference R4 to R5
I0010	-3.412	-0.574	-0.232	-0.047
I0011	-2.663	-0.344	-0.429	-0.061
I0020	-0.574	0.006	0.061	0.155
I0021	-2.431	-0.559	-0.210	0.343
I0028	-0.267	0.227	0.431	0.097
I0031	0.584	0.200	0.445	-0.356
I0033	0.826	0.530	0.459	-0.095
I0035	1.853	0.543	0.434	-0.067
I0036	0.358	-0.060	-0.119	-0.150
I0037	1.437	0.362	-0.001	-0.027
I0038	1.288	-0.063	0.254	-0.008
I0039	2.365	-0.471	0.074	0.518
I0042	2.166	-0.530	0.269	-0.027
I0043	2.228	0.456	0.101	0.408

Table 4.6 Change in item difficulty estimates, Oromifa PPVT

Item no.	Fixed location/difficulty	Difference R2 to R3	Difference R3 to R4	Difference R4 to R5
I0010	-2.900	-0.442	-0.335	0.013
I0020	-0.517	-0.557	-0.298	0.395
I0021	-1.963	0.598	-0.376	-0.221
I0022	0.064	0.012	-0.539	0.039
I0026	-0.073	-0.238	-0.715	0.519
I0029	0.612	0.209	-0.856	-0.041
I0039	1.953	0.307	-0.753	-0.135

Table 4.7 Change in item difficulty estimates, Tigrinya PPVT

Item no.	Fixed location/difficulty	Difference R2 to R3	Difference R3 to R4	Difference R4 to R5
I0004	-3.232	-0.804	-0.570	0.254
I0010	-2.745	0.111	-0.520	0.156
I0011	-1.883	-0.346	-0.013	-0.560
I0013	-2.465	0.317	0.686	-0.829
I0020	-0.412	-0.412	-0.596	0.276
I0021	-1.631	0.319	0.549	-0.441
I0023	0.096	-0.088	-0.932	-0.569
I0026	0.626	0.620	-0.096	0.478

4.4 Scale validity and the transformation of scores

In this final section, the validity of the PPVT scale is argued based on five criteria: (i) the psychometric properties of the test data; (ii) the distribution of the Rasch scores; (iii) the relationship between the Rasch scores and other measures of achievement available in the data structure; (iv) the relationship between the Rasch scores and wider contextual data; and (v) evidence of the effect of guessing on person ability estimates.

4.4.1 Summary of the psychometric properties

Using the Rasch measurement model, the PPVT data are screened for potential anomalies in the data.

<u>Item targeting and fit</u>. The presentation of children’s abilities is limited by the range of items available in the scale. When the children are aged five, the test does not precisely measure vocabulary at very low ability levels. The finding implies that children’s initial starting points will be slightly underestimated in the growth models. By age 15, there is a large ceiling effect for all subgroups, and most notably in Amharic. The implication for growth modelling is that individual differences in growth will be underestimated. Item

misfit is identified in the Amharic subsample only. Most of the misfitting items are harder items that overdiscriminate. These overfitting items do not differentiate well between the children at lower ability levels and discriminate much more than expected at higher ability levels. This may be a symptom of multidimensionality, which is further discussed below. The implication of the small slope variance is that it is not possible to fully investigate the relationship between changes in development and changes in living standards, or the mediating mechanisms in later childhood.

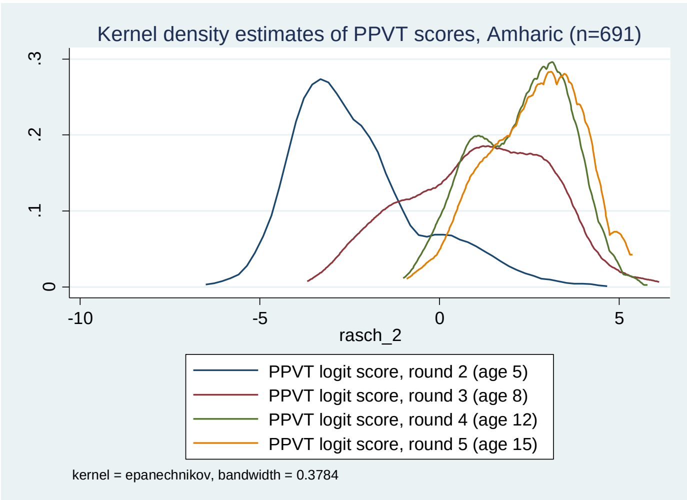
Differential item functioning. The DIF effect for language is significant for most of the items on the test. Therefore, the scale cannot be considered unbiased across language groups. There are different ways of treating DIF depending on whether it is uniform or non-uniform, as discussed on the methodology chapter. However, such methods are not intended to treat widespread DIF, which is a construct validity issue. The extent of the DIF effect in the present study provides considerable justification for separate language group analyses. Therefore, the trajectory analysis will seek to compare associations between vocabulary growth and contextual factors for each language group separately.

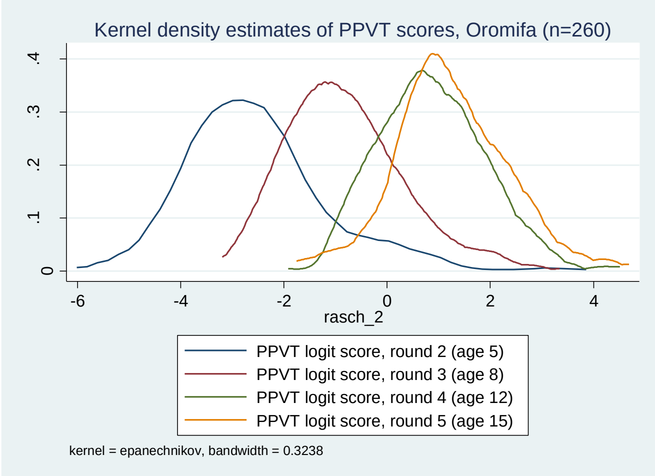
Multidimensionality. The results suggest the possibility of a multidimensional scale, or two or more factors that are being measured simultaneously. Some of the harder items on the test may be measuring some other factor that is correlated with the PPVT, in turn giving higher ability children some advantage. The children seem to be demonstrating their knowledge on this other construct in addition to vocabulary. For the present study, at each round of data collection subset comparisons were made to investigate the magnitude of multidimensionality. The procedure involves fitting two subsets of items to the Rasch model: one with factors loadings that load positively on the first principal component and another with negative factor loadings. The person parameters for the two subsets are estimated and their difference is compared based on a paired t-test. Paired t-tests indicated that person ability estimates obtained from the two subsets of items significantly differ in all three languages ($p < 0.05$). Deciding how to handle multidimensionality is “a relative issue and depends on the purpose of the scale or assessment” (Marais and Andrich, 2008a: 2012). Unidimensionality is regarded as a continuum, and a researcher must determine at which point multidimensionality threatens the interpretation of results given the aims of the research (Smith, 2002). For the present study, multidimensionality is acknowledged as a potential measurement disturbance and explored in the next chapter in the context of urban-rural differences.

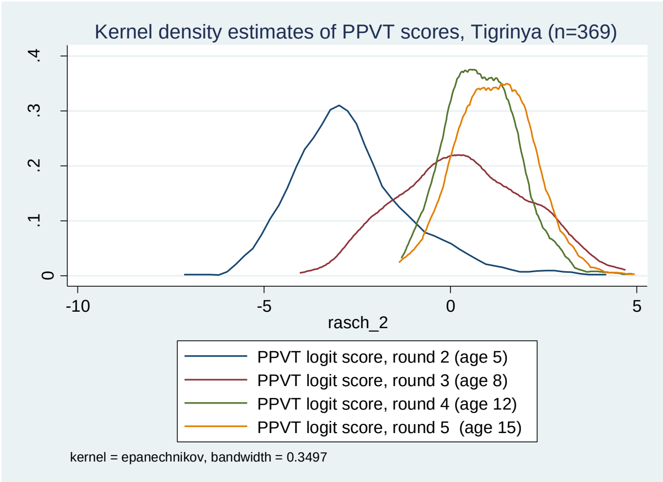
4.4.2 Distribution of the Rasch scores

The final set of scores showed progress over time. The figures below show the change in the distribution of the Rasch scores for each of the three language groups. The distribution shows the proportion of children achieving at different levels. Each distribution represents a different time point. The findings from the assessments show evidence of progress over the period as the spread of scores shifts rightwards, with a statistically significant growth in the mean score at the 1% level for all language subgroups. Progress between age 12 and 15 is small.

Figure 4.19 Change in the distribution of vocabulary scores, by language group







The Rasch score distributions highlight the ceiling effect. In Oromifa, the transformed scores produce a distribution that is close to normal, reflecting greater homogeneity among the sample. Among the Tigrinya language group, the results show greater sample heterogeneity at eight years, indicating a wider range of abilities than suggested by the raw scores. As further support for the robustness of the Rasch scores, the results showed that the mean raw score and the mean Rasch score are significantly positively correlated at the 5% level (Amharic: ρ = 0.85; Oromifa: ρ = 0.93; Tigrinya: ρ = 0.79).

4.4.3 Vocabulary ability by other measures of achievement

Table 4.8 shows that at each round of the survey, the Rasch scores are significantly positively correlated with all measures of achievement available in the Young Lives data structure: reasoning, literacy and mathematics. At age five, the scaled scores correlate particularly well with the Cognitive Developmental Assessment that was administered at age five. The results support the construct validity of the PPVT instrument when used within language subgroups. The PPVT scores at age 12 are very highly correlated with the age 15 scores, a result of the ceiling effect. Table 4.9 provides a summary of the correlations between the scaled Rasch scores and the scaled achievement test scores for each language group separately. Overall, the correlations are lowest in the Tigrinya language version of the test.

Table 4.8 Correlations between measures of achievement, full sample

	PPVT5	PPVT8	PPVT12	PPVT15	CDA5	Lit8	Ma8	Lit12	Ma12	Lit15	Ma15
PPVT5	1.00										
PPVT8	0.45	1.00									
PPVT12	0.44	0.65	1.00								
PPVT15	0.40	0.61	0.81	1.00							
CDA 5	0.53	0.36	0.36	0.34	1.00						
Lit 8	0.32	0.57	0.51	0.48	0.26	1.00					
Math 8	0.44	0.66	0.61	0.59	0.35	0.70	1.00				
Lit 12	0.30	0.49	0.56	0.53	0.28	0.54	0.62	1.00			
Math 12	0.36	0.48	0.59	0.57	0.31	0.54	0.67	0.69	1.00		
Lit 15	0.32	0.48	0.63	0.63	0.30	0.48	0.57	0.65	0.66	1.00	
Math 15	0.36	0.44	0.55	0.56	0.31	0.46	0.58	0.58	0.67	0.67	1.00

*All correlation coefficients are statistically significant at the 5% level.

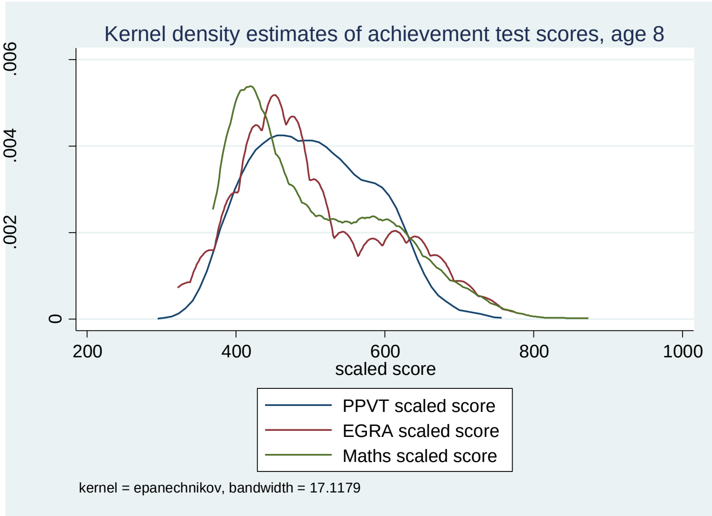
Table 4.9 Correlations between the Rasch scores and other measures, by language

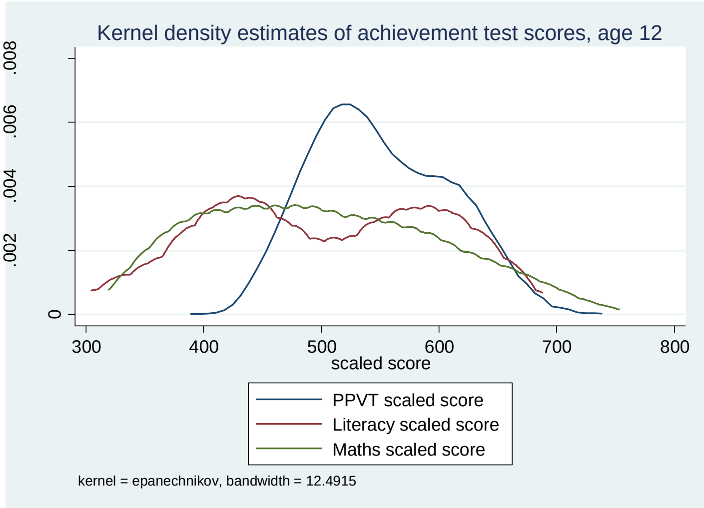
Outcome variable	Assessment instrument	Amharic	Oromifa	Tigrinya
Reasoning (5 years)	CDA	0.64	0.41	0.33
Reading (8 years)	EGRA	0.34	0.49	0.44
Maths (8 years)	30 arithmetic items	0.47	0.22	0.25
Reading (12 years)	24 comprehension items	0.42	0.34	0.28
Maths (12 years)	28 mathematics items	0.44	0.33	0.31
English literacy (15 years)	28 comprehension items	0.42	0.41	0.38
Maths (age 15)	28 mathematics items	0.40	0.42	0.35

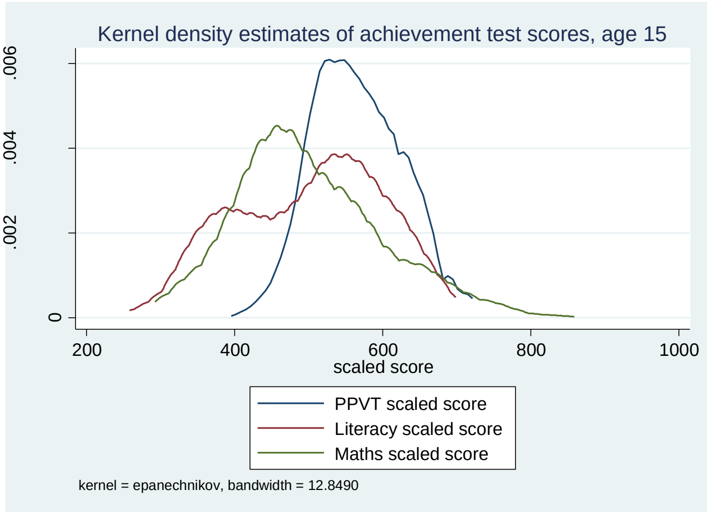
*All correlation coefficients are statistically significant at the 5% level.

In the figures below, the Kernel distribution of the achievement test scores are shown for age eight, 12 and 15. Each distribution represents a different test: PPVT, maths or literacy. At age five, literacy and maths tests were not administered. All test scores have been transformed onto the same scale (mean = 500, sd = 100). The results reported relate to the full sample of children. The distributions, which show substantial overlap particularly at age 8, provide additional support for the validity of the Rasch estimates.

Figure 4.20 Distribution of achievement test scores, whole sample



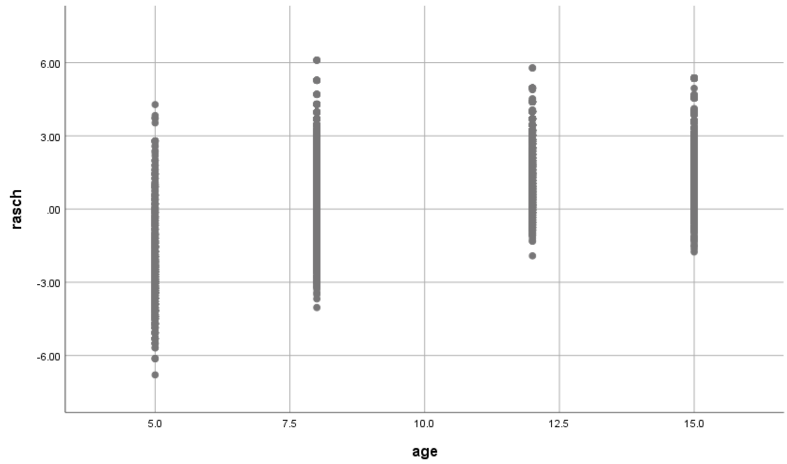




4.4.4 Vocabulary ability by contextual factors

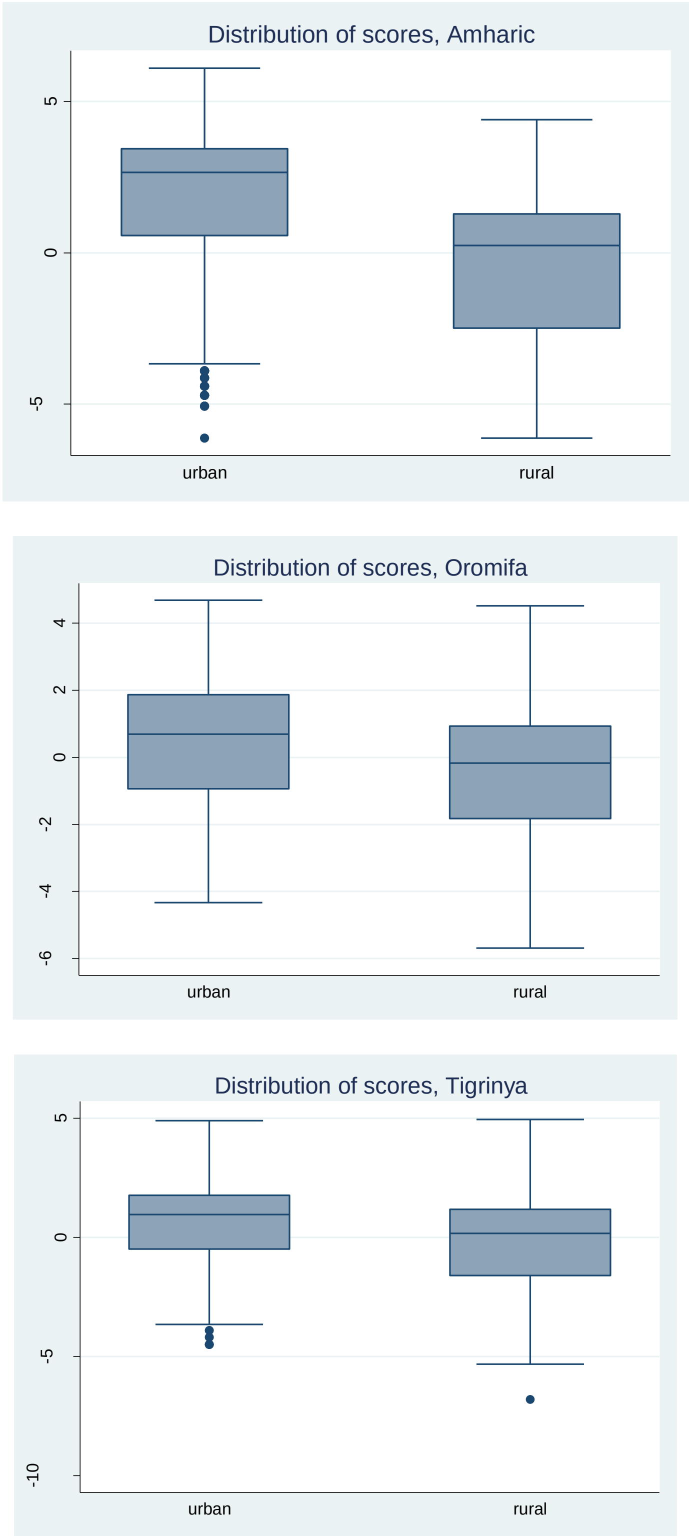
The final set of scores correlated with wealth, nutrition, mothers' educational level and preschool attendance as expected. Figure 4.21 shows the scores by age, which indicate a wide spread of scores at each age level. A simple linear regression indicates a positive linear association between the test scores and age (b_{ppvt}= 0.42, p-value < 0.001). The results confirm the theoretical expectation that proficiency increases with age. The difference in the mean scores between 12 and 15-year-olds is statistically significant at the 1% level, with 15-year-olds scoring 0.36 points higher on average. The finding suggests that despite the large ceiling effect, the PPVT is still picking up changes in learning between the later rounds.

Figure 4.21 Range of PPVT achievement, scores by age



The mean score difference between boys and girls is significant (F = 4.2(1); p <0.001). However, the difference is very small with girls outperforming boys by only 0.03 points overall. The mean difference in performance between children living in urban and rural areas is very large and significant (F = 31.5(1); p <0.001), with urban children scoring 1.83 points higher on average. Figure 4.22 depicts the difference in the range of scores, with the blue line indicating the median score for each subgroup.

Figure 4.22 The distribution of scores among urban and rural subsamples



The difference is most pronounced among the Amharic subsample. This descriptive finding taken together with the results of the psychometric analysis suggest that children's experiences vastly differ between urban and rural areas in terms of schooling and exposure. This difference may also be driving the problem of multidimensionality in the data, as discussed in the previous section.

4.4.5 The effect of guessing on parameter estimates

The Rasch model does not account for guessing. Within the Rasch paradigm, guessing is not considered as something to control for but rather a measurement disturbance, i.e. an anomaly in the data that should be addressed to improve the construct validity of the instrument. Random guessing is more likely to occur when a person has a low ability level in relation to the item's difficulty level such that more correct responses are generated than would have been the case if guessing had not been a factor (Andrich et al., 2012). Guessing is suspected when the proportion of persons who answered the item correctly is greater than the ICC prediction at lower levels of ability.

Guessing does not appear to be a problem for the harder items on the test. This is expected since items that were too difficult for a child were not administered and therefore could not attract guessing. However, in earlier rounds of the survey, guessing may have been a factor among low proficiency children in responses to several easier-to-moderately difficult items that are also low discriminating items. For these items, among the class intervals with low mean locations, the proportion of persons who answered the items correctly is higher than the theoretical prediction. Following the procedure suggested by Andrich et al. (2012), a pseudo-guessing parameter coefficient of 0.20 is obtained. The finding indicates that even the least able children have a 20% chance of a correct response overall. An alternative way of addressing the disturbance is to use the 3-paramerter logistic model to obtain the final ability estimates. The 3PL model estimates can be generated by the Stata14 OpenIRT module. OpenIRT includes Maximum Likelihood estimates (MLE), Bayesian expected posterior estimates (EAP), and plausible values (PV). Plausible values are thought to produce more accurate estimates of ability and dispersion (Das and Zajonc, 2010). However, problems with the 3PL model have been discussed widely in the literature, and has been presented in the literature review section, namely: (i) the 3PL model includes an item parameter for guessing even though guessing is a characteristic of person and not items and (ii) the model assumes that all persons employ random guessing despite the likelihood that guessing will only occur when an item is too difficult for a person.

4.4.6 Transformation of scores

The Rasch scores, which are expressed in logits, were obtained for a total of 1,310 children, 594 girls and 716 boys. To avoid negative achievement scores, the logit scores for all children were transformed onto a scale with a mean of 500 and a standard deviation of 100, as follows:

$$\theta_n^* = [(\theta_n - m)/sd] * 100 + 500 \qquad \text{...4.1}$$

where θ_n is the ability for child n in logits and θ_n^* the transformed ability for child n, m is the weighted mean score of all children and sd is the weighted standard deviation of all children. The transformed scores, which are referred to as scaled scores, are precise measures of where each child sits along the ability scale. The values of the mean and standard deviation of the logit and scaled scores are presented in Table 4.10.

Table 4.10 Mean and standard deviation of logit and transformed scores

	5 years	8 years	12 years	15 years
Mean logit score	-2.456	0.515	1.622	1.985
Standard deviation logit score	1.667	1.896	1.383	1.423
Mean scaled score	378.49	503.91	550.67	565.99
Standard deviation scaled score	70.35	80.04	58.40	60.09

4.5 Summary

Based on the psychometric properties of the test, the Rasch logit scores provide a measure of ability that is an improvement over the raw score totals. Issues are identified in the data that would have been missed by analysing developmental trajectories using the raw scores alone, i.e. ceiling effect, gaps in item targeting, multidimensionality, and score instability across time. The scores correlate well with other measures of achievement available in the data structure, which reinforces the construct validity of the measure. The scores also conform with theoretical expectations, which is further discussed in the next chapter. The Rasch scores also shed light on the extent of heterogeneity among the linguistic subsamples and the differences in the functioning of the test across groups. The Amharic version of the test could have been improved by including additional harder items to address the ceiling effect and by addressing urban-rural biases that have been suggested. The construct validity of the Oromifa and Tigrinya language versions of the test could have been be improved by including a wider range of relevant easier and harder items. Among the Tigrinya subsample, the extent of item instability indicates changes in the meaning underlying the items over time, possible translation issues, changes in administration procedures, or other testing errors.

The findings of the analysis suggest that growth trajectories, which are analysed in the next chapter, should be estimated and explained for each language group separately. The current chapter also highlights some of the challenges of interpreting assessment results in multicultural contexts and across time, and the need for further research in this area. In particular, the findings presented in this chapter highlight the importance of understanding substantive anomalies in the data such as item misfit, scale instability and multidimensionality and the effect of such anomalies on results interpretation.

Chapter 5 Associations between wealth and vocabulary development

5.1 Introduction and rationale

This chapter addresses the second set of research questions about the effects of poverty on children's vocabulary developmental trajectories. The aim of this component of the research is to better understand patterns of change in the relationship between poverty and children's vocabulary ability over time. The transformed Rasch estimates (M = 500; SD = 100), which are referred to as scaled scores from this point onward, are modelled to establish the mean trends in vocabulary development for each language subgroup to determine if there is any variation to explain in the children's initial starting points and the growth rates. The chapter goes on to examine the children's socioeconomic contexts, focusing first on the extent to which living standards in early childhood, as measured by household wealth, explains the variance in vocabulary scores. It then asks whether changes in living standards further help to explain changes in development.

A multilevel model for change approach is used to analyse individual differences over time. Two statistical models are estimated to examine the time effects of wealth on children's progress. The first model controls for the children's initial wealth status and the second model accounts for changes in wealth over time. In the first model, wealth in early childhood is assumed to "launch" the developmental trajectory. The choice of model is based on the premise that ability is largely formed in early childhood and that interventions designed to alleviate poverty effects are needed early in life to prevent developmental disadvantage and stagnation in later childhood. Any observed differences in outcomes between the highest and lowest wealth quintiles are discussed, namely whether the initial developmental gap widens, narrows, or remains constant throughout middle childhood and adolescence. The second model explores the extent to which changes in wealth over time further help to explain patterns of development. In this model, change in ability is modelled as a function of change in wealth. Results from the analysis will shed light on patterns of change in the poverty-growth relationship. Do improvements in household wealth during childhood improve the developmental trajectory, or conversely, does a decrease in wealth predict developmental stagnation?

The size and significance of the variance in the children's initial starting points and their growth rates will clarify which statistical model is best suited for explaining the relationship between cognitive development and wealth. A small or insignificant variance in the growth rate implies a launch model of developmental change and that vocabulary gaps observed in childhood are established in the early years. Developmental gaps would then be predicted by individual differences in children's early experiences and socioeconomic contexts. Alternatively, a sizeable and significant difference in individual growth rates suggests that the second statistical model, referred to as the change-to-change model, may be more effective in explaining the relationship.

The findings will also inform the selection of mediating variables and the choice of statistical mediation model, which is the third and final empirical study. If individual differences are consistent with the launch model of developmental change, the mediation analysis will focus on exploring early interventions to address disparities in vocabulary outcomes in early childhood. However, if the developmental trajectory is more efficiently predicted by changes in living standards over the course of childhood, the mediation study will investigate potential interventions at later stages of childhood as well.

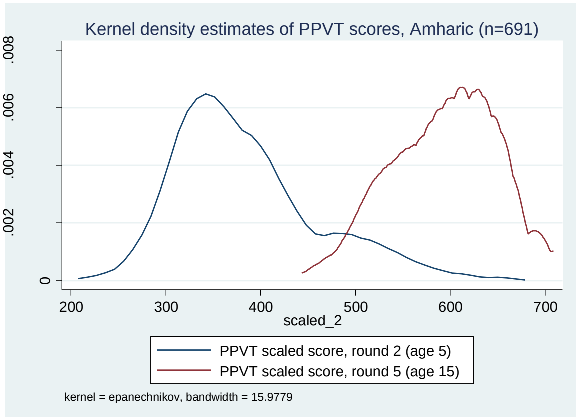
5.2 Descriptive statistics

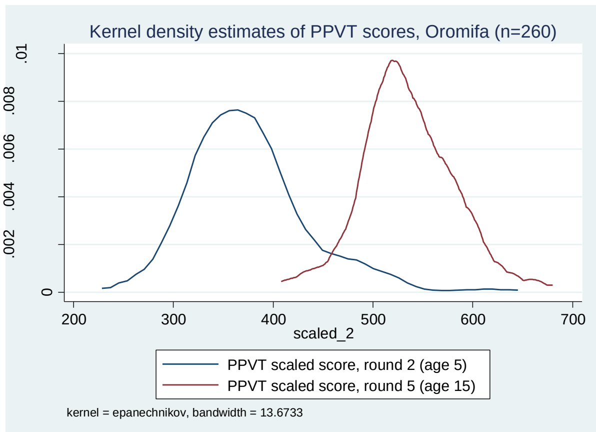
The study sample consists of a total of 1,321 children, 54.5% of whom are boys and 45.5% are girls. This total excludes children for whom Rasch scores could not be computed because their response patterns deviated considerably from the Rasch model. Children who took the test in different languages across rounds are also excluded from the analysis.

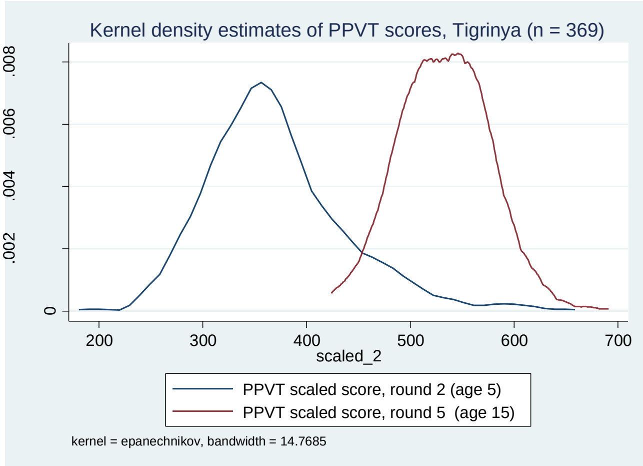
5.2.1 Changes in vocabulary ability between age five and fifteen years

Figure 5.1 shows the distribution of the PPVT scaled scores for each of the language subgroups separately, providing evidence of progression over the period. The spread of scores shifts rightwards, with a statistically significant growth in the mean score at the 1% level. There is more heterogeneity among the Amharic-speaking subsample, as evidenced by a wider spread of scores. This may be explained in part by extreme difference in the lives of children residing in the capital city, Addis Ababa, compared with their peers in rural and smaller urban areas. By the fifth and final round of the survey, a large proportion of children are demonstrating knowledge at the highest range of scores, a consequence of the ceiling effect reported in the psychometric study. Low variance at higher ability levels is most problematic within the Amharic speaking subsample.

Figure 5.1 Change in Kernel distribution of scaled vocabulary scores, by language group

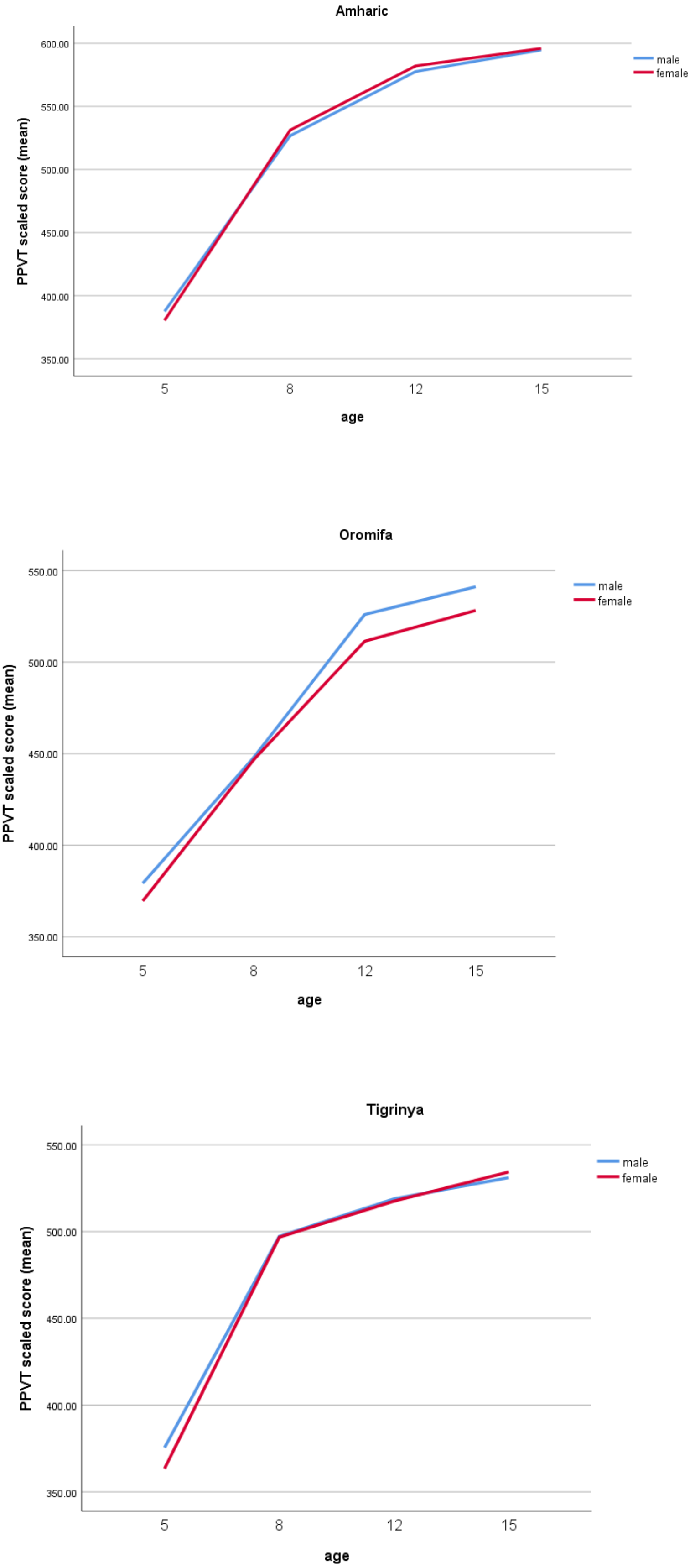






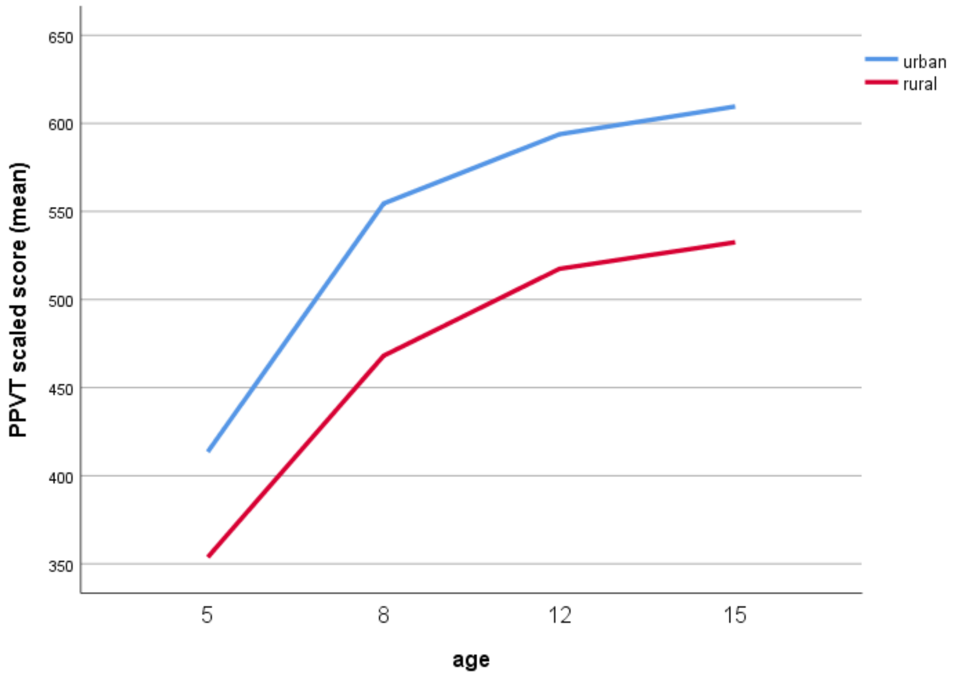
It was reported in the previous chapter that the gender difference in vocabulary outcomes were not significant overall; however, breaking down the analysis by round and language category shows small discrepancies between boys and girls (Figure 5.2). It is important to note that Differential Item Functioning for gender was not flagged as a problem in the psychometric analysis (section 4.3.4), suggesting that any observed gender differences more likely reflect real differences as opposed artefacts of test bias or the scaling procedure. On average, at age five, boys outperform girls on the Tigrinya language version of the test by 12 points ($p < 0.10$). This gap does not persist throughout childhood as girls catch up by age 15. Among the Oromifa-speaking subsample, five-year-old boys and girls perform at the same level on average. By age 15, boys achieve 13 points higher on average ($p < 0.05$). Mean level gender differences are not apparent among the Amharic-speaking subsample.

Figure 5.2 Mean PPVT scaled score at each time point, by gender



Children living in rural areas experience a cumulative disadvantage as inequalities in cognitive outcomes between urban and rural areas widen with time (Figure 5.3). This descriptive finding reinforces previous interpretations about the advantages of living in the capital city, and the possibility of a multidimensional scale. Among the Amharic-speaking subsample, which includes the children residing in Addis Ababa, at age five only 10% of rural children scored above the mean in comparison with more than 60% of urban children. By age 15, 10% of rural children are achieving above the mean in comparison with 77% of their urban peers. In other words, mean level urban-rural disparities in outcomes widen over time. The results are extreme compared with the other subgroups. In the Oromifa and Tigrinya language versions of the test, the gap is smaller and constant throughout the survey period, with around 55% of urban and 40% of rural children scoring above the mean at the beginning and the end of the testing process.

Figure 5.3 Change in the mean PPVT scaled score by geographic stratum, full sample



The urban advantage in children's outcomes is widely documented in the literature. Children living in urban areas are afforded more opportunities and services than their rural counterparts (UNICEF, 2018). Most of the Amharic-speaking urban children reside in the capital city. Their experiences differ vastly from the experiences of children living in rural areas. It is worth noting that the sentinel sites selected in Addis Ababa consist of: (i) an overcrowded area in the city centre, (ii) an industrial area in the southern part of the city, and (iii) a slum area. When the sampling occurred, all survey households located in Addis Ababa were poorer than households in the national Demographic Health Survey (DHS) sample (Outes-Leon and Sanchez, 2008). This was an intentional outcome resulting from oversampling the poor. The implication is that among the sample of children living in the Amhara region, poor children living in Addis Ababa have better cognitive outcomes than poor children living in rural areas. It would be useful to examine outcomes by sentinel site to observe which urban sites, if any, perform less well than rural areas. Findings from a recent UNICEF report explained how urban inequalities result in many urban children missing out and suffering more serve deprivations than rural children (UNICEF, 2018).

5.2.2 Poverty characteristics among the sample of children

Wealth information was collected at all time points, when the children were ages one, five, eight, twelve and fifteen years (Table 5.1). The wealth variable is normally distributed overall: approximately symmetric and slightly more peaked than normal (mean: 0.36; std. deviation: 0.20; skewness: 0.23; kurtosis: -0.70). Descriptive statistics for the wealth index score are presented in Table 5.1 for the full sample. Overall, the average wealth index score increases over time, from 0.32 to 0.44 over the period. The expansion in wealth is largely driven by improvements in access to services and ownership of consumer durables.

Table 5.1 Descriptive statistics for wealth across five data collection time points

Wealth index	Age 1	Age 5	Age 8	Age 12	Age 15
Mean	0.251	0.323	0.370	0.413	0.436
Std.Dev	0.188	0.195	0.193	0.180	0.175
Skewness	0.624	0.419	0.280	0.172	0.186
Kurtosis	-0.641	-0.669	-0.702	-0.607	-0.653

Wealth quintiles are constructed to examine mean level differences in cognitive outcomes between children living in the richest households and children in the poorest households (Table 5.2). The top quintile consists of the wealthiest 20% of the population, and the poorest 20% are in the bottom quintile. Children living in the top wealth quintile are six times richer than children in the bottom quintile.

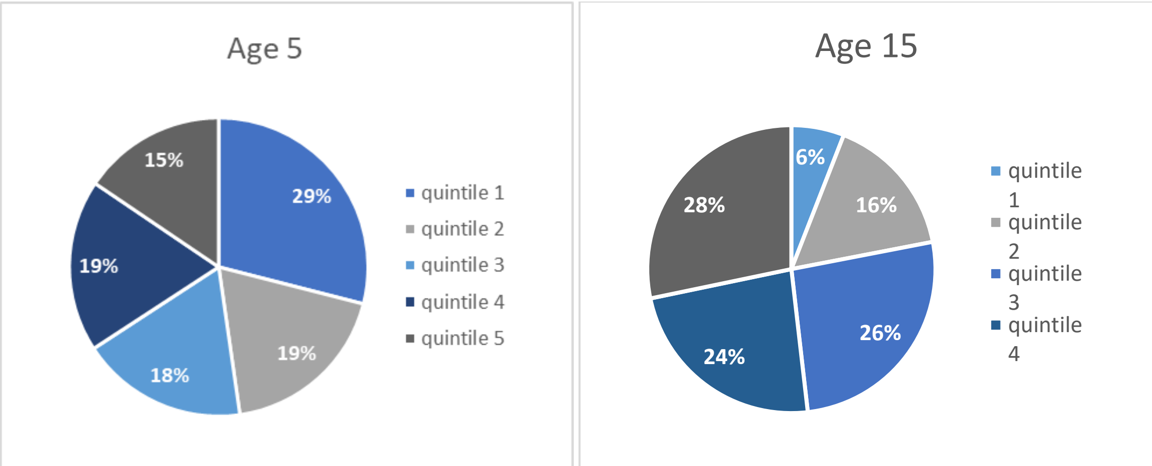
Table 5.2 Socioeconomic characteristics, full sample

	Age 5		Age 8		Age 12		Age 15	
Variable	Mean (SD)	N	Mean (SD)	N	Mean (SD)	N	Mean (SD)	N
Wealth index score	0.32 (0.19)	1305	0.37 (0.19)	1319	0.41 (0.18)	1316	0.44 (0.17)	1316
- Consumer durables	0.21 (0.19)	1310	0.27 (0.22)	1320	0.30 (0.21)	1318	0.32 (0.19)	1318
- Access to services	0.43 (0.32)	1306	0.49 (0.29)	1320	0.55 (0.27)	1317	0.58 (0.29)	1318
- Housing quality	0.32 (0.20)	1309	0.36 (0.19)	1319	0.39 (0.19)	1317	0.41 (0.19)	1316
Wealth, bottom quintile	0.10 (0.06)	377	0.11 (0.05)	243	0.12 (0.05)	118	0.13 (0.05)	78
Wealth, 2nd quintile	0.23 (0.03)	244	0.23 (0.03)	248	0.24 (0.03)	241	0.24 (0.03)	211
Wealth, 3rd quintile	0.35 (0.03)	235	0.35 (0.03)	285	0.35 (0.03)	307	0.35 (0.03)	345
Wealth, 4th quintile	0.47 (0.04)	243	0.47 (0.04)	267	0.47 (0.04)	312	0.48 (0.04)	310
Wealth, top quintile	0.64 (0.08)	202	0.65 (0.08)	275	0.65 (0.08)	337	0.66 (0.08)	372

Figure 5.4 presents the proportion of children in each wealth quintile. The data show that the surveyed children become less poor over time. The proportion of children living in the poorest two quintiles is more than halved over the period, from 48% to 22%. By age 15, only 6% of the sampled children are living in the bottom quintile. Correspondingly, the proportion of children living in the richest two quintiles grows from 34% to 52%. In other

words, more than half of the children are in the top two wealth quintiles by the age of 15. It is important to note that this result does not imply that half of the children are wealthy by the end of the survey period. As discussed in the methodology chapter, one of the main priorities of the Young Lives study is to provide in-depth data about children living in poverty. This resulted sampling procedure designed to over-sample poor areas. Thus, children in the top quintiles can be thought of as being less poor than children in the lower quintiles.

Figure 5.4 Proportion of children in each wealth quintile, full sample



The wealth index scores at each data collection time point are highly correlated as shown in the table below. The large and significant correlation coefficients conform with expectations. As previous research has shown, wealth is not expected to fluctuate as much as income or consumption expenditure (O'Donnell et al., 2008). Household wealth represents a more "permanent status" than other living standards measures (Rutstein and Johnson, 2004: 4).

Table 5.3 Correlations between wealth index scores across time

	Age 1	Age 5	Age 8	Age 12	Age 15
Age 1	1.00				
Age 5	0.82***	1.00			
Age 8	0.78***	0.87***	1.00		
Age 12	0.71***	0.78***	0.82***	1.00	
Age 15	0.69***	0.75***	0.78***	0.82***	1.00

*Correlation coefficient is statistically significant *p=0.1 **p=0.05 ***p=0.01

The wealth index score is also positively and significantly correlated with the scaled PPVT scores at each time point, as shown in Table 5.4. The correlation coefficient is largest for initial wealth in that the scaled scores at each time point are most strongly correlated with wealth at one year of age.

Table 5.4 Correlations between wealth index scores and scaled test scores

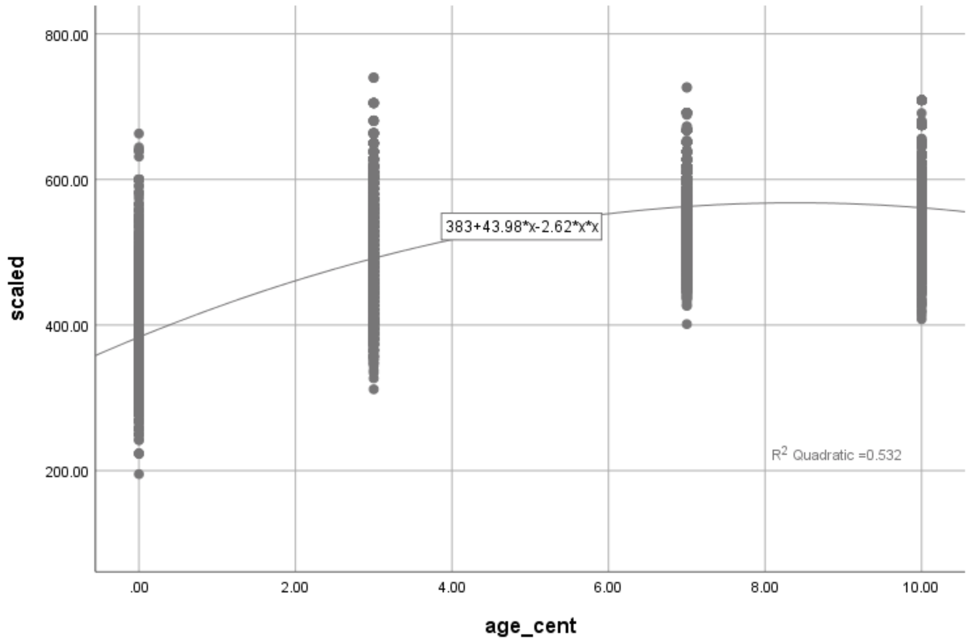
	Scaled score 5	Scaled score 8	Scaled score 12	Scaled score 15
Wealth-1	0.47***	0.61***	0.66***	0.65***
Wealth-5	0.46***	0.52***	0.64***	0.64***
Wealth-8	0.46***	0.51***	0.64***	0.63***
Wealth-12	0.45***	0.52***	0.63***	0.61***
Wealth-15	0.43***	0.53***	0.63***	0.63***

*Correlation coefficient is statistically significant *p=0.1 **p=0.05 ***p=0.01

5.2.3 Trends in vocabulary development between five and fifteen years

Figure 5.5 plots each child's scaled score at age five, eight, 12 and 15 years, with a quadratic fit line to explore the mean starting point, the slope, and the overall shape of the distribution. The x-axis represents the children's age, centred at five years, when the PPVT was first administered. The y-axis represents the scaled scores. From baseline, group mean cognitive ability increases over time, but at a diminishing rate after eight years, indicated by the negative quadratic term. In this simple quadratic regression framework, mean initial ability is estimated at 383 points, with an overall growth rate per period of 44 points. Ability growth slows over time at a rate of 2.62 points per period.

Figure 5.5 Dot density distribution of scaled scores, full sample



The figure suggests that there is a substantial variability in individual abilities at onset, when the children are five years old. There is less variation in scores at 12 and 15 years. Table 5.5 below provides a descriptive summary of the scores at each time point.

Table 5.5 Descriptive statistics PPVT scaled scores, full sample (n = 1,310)

	Mean	SD	Min	Max
Scaled score 5	378.49	70.35	195.21	662.96
Scaled score 8	503.91	80.04	311.80	739.81
Scaled score 12	550.67	58.40	401.24	726.38
Scaled score 15	565.99	60.09	408.08	708.69

5.3 Estimating the effects of wealth on children's vocabulary development

In this section, individual developmental trajectories are estimated and the significance of the variance in both the intercept and slope is examined. The multilevel model for change is specified as a latent growth model, i.e. within a structural equation framework, to obtain slope estimates for the time-varying covariates. In a structural equation approach, the slopes for the time-varying covariates (i.e. modelled at level-1) vary over time points in contrast with the multilevel approach in which they vary across individuals. The analyses are conducted in MPlus Version 8 (Muthén and Muthén, 2017), using the Maximum Likelihood (ML) estimator for continuous data.

The variables used in the analysis are summarized in Table 5.6. The “childid” variable is an identifier that comprises two letters to identify the country followed by six digits. The first two digits, which range from one to 20 relate to the sentinel site in which the child lives. The remaining digits identify the child within the sentinel site. The variable “age” indicates the measurement occasion and is represented by each child’s age in years. It is centred at onset in the statistical models (five years). Similarly, “agemon52” is represented by each child’s age in months and is centred at 52 months (4 years and 4 months), which is the age of the youngest child in the sample. The dependent variable “scaled” is the PPVT scaled score. The launch model covariate, initial wealth “wi_1”, describes the child’s wealth status at onset. In the change-to-change model, wealth is grand mean centred (“wi”).

Table 5.6 Contextual variables used in the analysis

Variable	Level	Description	Values	Scale
childid	Level-2	Child identifier	Integer	Nominal
age5	Level-1	Measurement occasion represented by each child’s age in years, baseline-centred in the launch model	“-4” = 1 year; “0” = 5 years; “3” = 8 years; “4” = 12 years; “10” = 15 years	Ordinal
agemon52	Level-1	Measurement occasion represented by each child’s age in months, baseline-centred in the launch model	-3.38 months to 11.67 months	Ordinal
scaled	Level-1	Dependent variable representing each child’s scaled score on the PPVT	195.21 to 739.81	Interval
wi_ 1	Level-2	Descriptive variable representing the initial wealth index score value at age 5, grand mean centred	0.00 to 0.74	Interval
wi	Level-1	Descriptive variable representing the wealth index score value at each measurement occasion, grand mean centred	0.00 to 0.92 (grand mean = 0.35)	Interval

5.3.1 The base model

To analyse the effects of poverty on development, the first step is to estimate the base model with only time as a fixed effect, and no other covariates. The purpose of this preliminary step is to determine whether the intercepts and slopes significantly differ across individuals and to define the shape of the growth trajectory, i.e. linear or quadratic (Heck and Thomas, 2009; Heck et al., 2009). The results of the base model will then inform the next stages of the analysis.

To illustrate individual differences in initial status and growth, developmental trajectories are presented in Figure 5.6 for a subsample of the children. Predicted values are obtained by estimating the quadratic model with a random intercept and a random slope effect to illustrate the shape of the trajectories. The graph demonstrates the large variance in the children's initial starting points and stability in the growth path. The figure depicts the wide spread of scores at onset, the small amount of variation in the growth rate, and decreasing marginal growth, all of which are consistent with developmental theory. That is, the graph reinforces the idea that *ability* stabilises earlier in childhood. However, the small variation in the growth rate may also be due, in part, to statistical measurement error resulting from the large ceiling effect observed at age 12 and 15. Such measurement error would lead to an underestimate of proficiency, particularly at higher ability levels, in turn leading to an underestimate progress over time among higher ability children.

Figure 5.6 Base model quadratic growth curves, mixed language subsample (n=500)

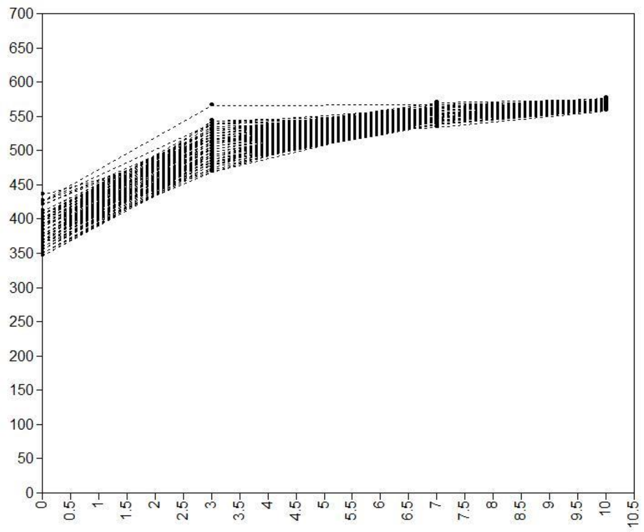


Table 5.7 provides a summary of the base model estimates for the full sample data. Both quadratic and linear models were estimated. In the first round of statistical modelling, only the intercept was allowed to vary, with the slope and quadratic terms fixed across individuals. The second round of models specified the intercept and slope as random effects, while holding the shape of the trajectory constant. These models provided the best description of the data according to the goodness of fit indices, including the deviance statistic (-2 x log likelihood). The smaller the value the better the fit. In the final round of data modelling, the intercept, slope, and quadratic terms are all modelled as random effects. Due to the negligible amount of variance in the quadratic effect relative to the large intercept variance, the linear model is selected for the analysis.

All fixed effects parameters are significant (Table 5.7). When the time score is zero, the scaled score is 330 points (γ_{00} = 330.28, $p < 0.01$). The value of the coefficient corresponds to the intercept of the average growth curve. The average growth rate is 7 points per month, although the growth slope factor is only weakly significant γ_{10} = 6.97, $p < 0.1$). The time of observation varies across individuals. At each wave, the age range is approximately 12 months. Due to the large variability in the time when children were assessed, age in months is included as a time-varying covariate. There are four beta coefficients representing the regressions of the scaled score on the child's age (i.e. $PPVT_5$ on $agemon_5, PPVT_8$ on $agemon_8, \ldots$), where age in months is centred on its starting value of 52 months. Recalling from the methodology chapter, the linear base model specification is as follows:

$$PPVT_{ti} = \beta_{0i} + \beta_{1i}AGE_{ti} + \varepsilon_{ti}$$

$$\beta_{0i} = \gamma_{00} + u_{0i}$$

$$\beta_{1i} = \gamma_{10} + u_{1i}$$

When combined, the model can be written as:

$$PPVT_{ti} = \gamma_{00} + \gamma_{10}AGE_{ti} + u_{0i} + u_{1i}AGE_{ti} + \varepsilon_{ti}$$

Including age in months as a time-varying covariate results in the modification shown below. The slopes for "age in months" vary over time points rather than individuals. This fixed effect parameter is specified as β_t.

$$PPVT_{ti} = \gamma_{00} + \gamma_{10}AGE_{ti} + \beta_t AGE_{ti} + u_{0i} + u_{1i}AGE_{ti} + \varepsilon_{ti}$$

The findings of the model show that at each time point, the effect of age in months on the scaled score is positive. The mean effect of age is decreasing with age, consistent with the theory of development. The declining slope also reflects the low variance in the scaled score at ages 12 and 15 (i.e. the ceiling effect).

Table 5.7 Base model estimates, full sample (n=1,310)

	Estimate	S.E
Intercept (γ_{00})	330.28***	4.61
Slope (γ_{10})	6.97*	4.47
Age 5 (β_1)	57.73***	5.60
Age 8 (β_2)	40.24***	3.71
Age 12 (β_3)	22.07***	3.90
Age 15 (β_4)	15.47***	3.98
Level-1 residual variances		
Age 5 (ε_{5i})	2847.72***	185.34
Age 8 (ε_{8i})	3015.58***	129.13
Age 12 (ε_{12i})	700.14***	53.70
Age 15 (ε_{15i})	477.92***	75.31
Level-2 variance components		
Intercept variance (σ_0^2)	2311.36***	163.98
Intercept*slope covariance (σ_{10})	-27.67*	14.66
Slope variance (σ_1^2)	13.91***	2.16
R-square		
ppvt_5	0.483***	0.024
ppvt_8	0.450***	0.018
ppvt_12	0.792***	0.015
ppvt_15	0.869***	0.019
Goodness of fit indices		
-2LL	57758.45	
RMSEA†	0.067	
CFI‡	0.962	
TLI‡	0.951	

*Correlation coefficient is statistically significant *p=0.1 **p=0.05 ***p=0.01. †The RMSEA ranges from 0 to 1. A value of 0.06 or less is indicative of good model fit; up to 0.08 is acceptable. ‡The CFI and TLI range from 0 to 1. A value of 0.90 or higher indicates acceptable model fit.

The variance components are also reported in Table 5.7. The estimates are used to determine how much variance in ability is present at each level. The level-1 variance ("repeated measures") summarises the variance to be explained at each wave, which is distinct at each time point and decreasing with time. The level-1 residual variance is largest at eight years of age and smallest at age 15. This is consistent with the findings from the psychometric analysis which showed that the widest spread of scores occurred at age eight, which is also the time point at which the PPVT performed most consistently with measurement expectations. The small residual variance at age 12 and 15 is consistent with developmental theory and the ceiling effect. At level 2, the findings show significant variation in the intercepts to be explained across individuals (σ_0^2 = 2311.36, p<0.01), and a relatively smaller but significant variance in the slope (σ_1^2 = 13.91, p<0.01). The negative sign on the covariance parameter indicates that a higher level of cognitive ability at onset implies slower growth. However, the covariance is only weakly significant (σ_{01} = -27.67, p<0.10). The results are represented as a latent growth model in Figure 5.7.

Figure 5.7 Latent growth model, full sample (n=1,310)

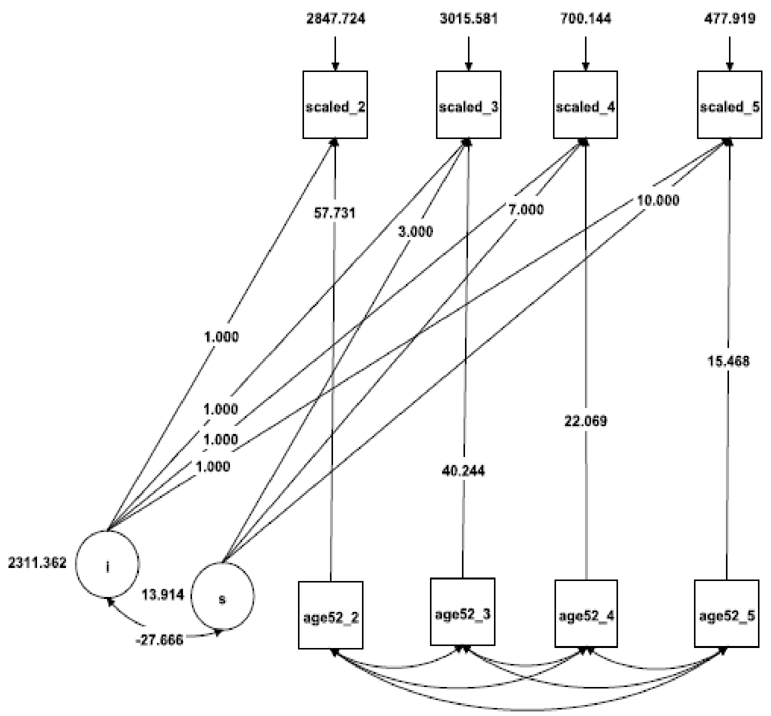


Table 5.8 presents the findings by language group. The R-square value is largest in Amharic. In Amharic and Oromifa, the results are consistent with the full sample estimates. However, by age 15 the effect of age on the scaled score is not significant, indicative of the low variance in scores at this age. In Tigrinya, the results deviate from expectations. The model does not fit the data as well, evidenced by poorer values of the goodness-of-fit indices. The growth slope factor is not significant.

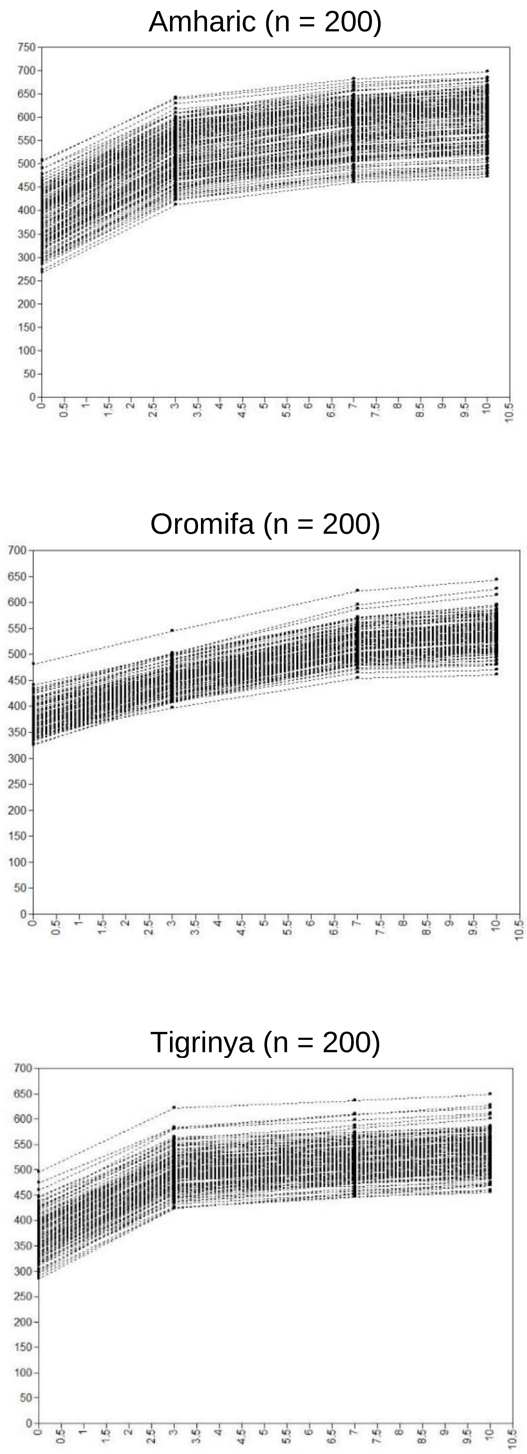
Table 5.8 Base model estimates, by language group

	Amharic (n=687)		Oromifa (n=254)		Tigrinya (n=365)	
	Estimate	S.E	Estimate	S.E	Estimate	S.E
Intercept (γ_{00})	336.47***	7.01	333.03***	9.76	315.20***	7.26
Slope (γ_{10})	19.71***	5.65	10.02*	7.02	-1.30	6.14
Age 5 (β_1)	56.15***	8.37	49.48***	11.66	68.53***	9.22
Age 8 (β_2)	34.91***	5.25	22.37***	6.14	49.30***	5.38
Age 12 (β_3)	13.47***	5.40	15.09**	6.45	27.38***	5.68
Age 15 (β_4)	5.74	5.46	9.58	6.60	21.48***	5.78
Level-1 residual variances						
Age 5 (ε_{5i})	2825.36***	248.05	2255.24***	294.26	3036.31***	366.29
Age 8 (ε_{8i})	2907.36***	184.14	1398.79***	154.86	2904.79***	208.66
Age 12 (ε_{12i})	566.44***	63.89	943.14***	124.82	400.38***	67.59
Age 15 (ε_{15i})	596.40***	88.67	785.11***	167.41	605.05***	119.59
Level-2 variance components						
Intercept variance (σ_0^2)	3126.90***	233.22	901.46***	248.11	1354.78***	243.27
Intercept*slope covariance (σ_{10})	-71.29***	21.77	-16.58	25.69	-37.51*	23.16
Slope variance (σ_1^2)	9.42***	2.77	7.30**	3.39	7.17**	3.38
R-square						
ppvt_5	0.549***	0.028	0.331***	0.066	0.390***	0.047
ppvt_8	0.500***	0.023	0.398***	0.050	0.351***	0.034
ppvt_12	0.822***	0.018	0.527***	0.042	0.760***	0.035
ppvt_15	0.816***	0.026	0.625***	0.070	0.694***	0.694
Goodness of fit						
RMSEA†	0.057		0.054		0.104	
CFI‡	0.975		0.952		0.879	
TLI‡	0.967		0.938		0.843	

*Correlation coefficient is statistically significant *p=0.1 **p=0.05 ***p=0.01. †The RMSEA ranges from 0 to 1. A value of 0.06 or less is indicative of good model fit; up to 0.08 is acceptable. ‡The CFI and TLI range from 0 to 1. A value of 0.90 or higher indicates acceptable model fit.

Differences are also evident in the plot of observed values for each subsample (Figure 5.8). In Amharic, the growth curves are consistent with expectations. In Oromifa, they appear linear. In Tigrinya, the growth curves are flatter and show some degree of developmental stagnation or poor progress over time. However, the results may also point to scale validity issues among this subgroup. In the previous chapter, section 4.3.7 discussed the stability of the scale. Item difficulty estimates were cross plotted across time for each language subgroup. The Tigrinya items did not scatter as evenly across the identity line and were dispersed haphazardly, indicating item drift. The results suggested that the items did not function in the same way across rounds. Moreover, the age five scores were only moderately correlated with the scores at age 12 and 15, compared with a high correlation between scores in the other two languages.

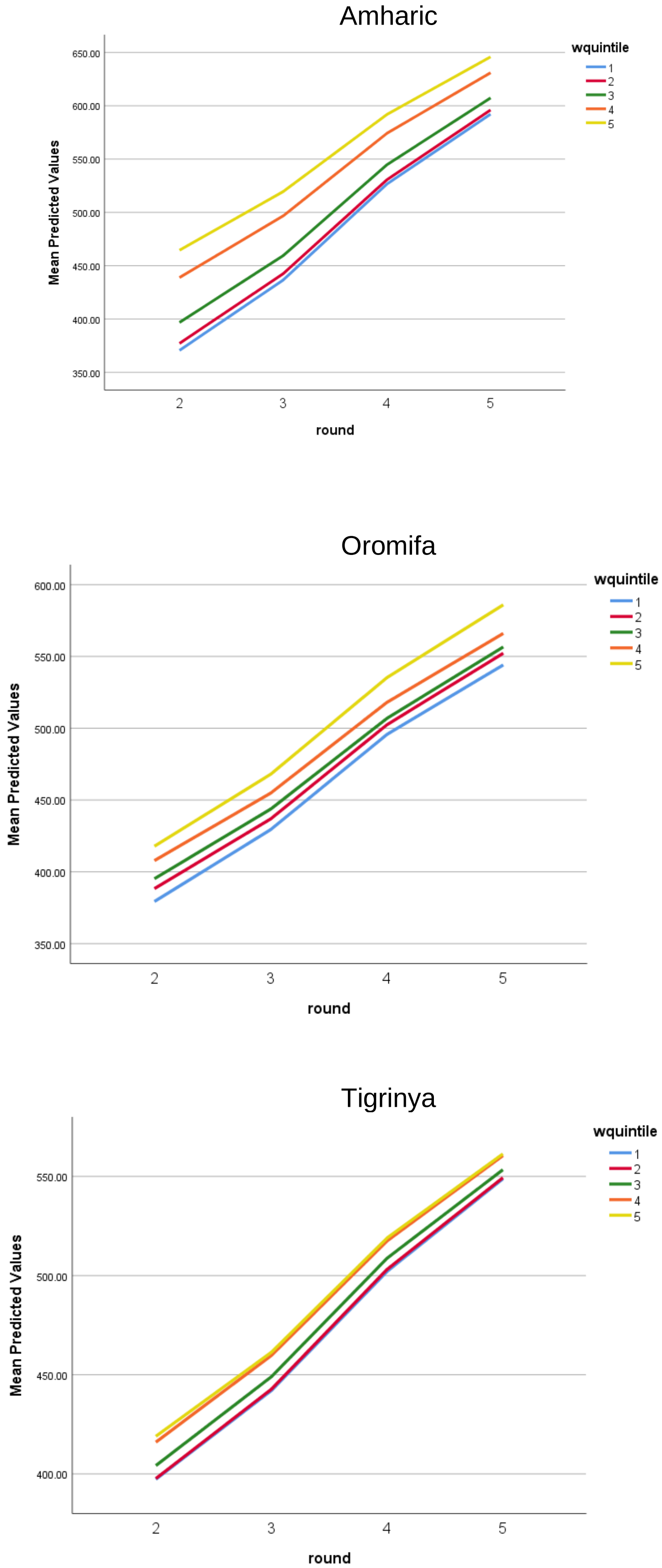
Figure 5.8 Linear growth trajectories, based on estimated individual values



The variance components are also shown in Table 5.8. The level-1 residual variance is decreasing with time. However, in Amharic the variance is largest at age eight, the time point at which the widest spread of scores is observed (i.e. good targeting). The relatively small residual variance at age 12 and 15 is consistent with developmental theory but may also reflect the ceiling effect evident at these final two rounds of testing. There is a large amount of intercept variance to be explained in all languages, and particularly in Amharic. The slope variance is considerably smaller but significant for all subgroups at the five percent level. In both Tigrinya and Amharic, the covariance term between initial status and the growth rate is significant and negative, but not in Oromifa in that a higher level of cognitive ability at onset does not imply slower progression for this subgroup.

Figure 5.9 shows the mean growth trajectories by wealth quintile for the full sample and for each language group separately to examine the significance and the magnitude of the gap between the two quintiles. Children in the top quintile (5th) are assessed higher on average than children in the bottom (1st). At each time point, group t-tests indicated a significant difference in the estimated mean scores between the two quintiles At age five, there is a greater disparity between quintiles in Amharic, and the curves become less divergent with time for this language group, by 0.47 SD or 40 points ($p < 0.01$). In Oromifa, the trajectories widen marginally, by approximately 4 points ($p < 0.05$), by 0.05 SD over the survey period. In Tigrinya, the gap narrows by 9 points or 0.15 SD ($p < 0.01$).

Figure 5.9 Average growth curves by quintile (random intercept and random slope)



5.3.2 The launch model of developmental change

The previous results showed a large and significant amount of variance to be explained in the intercept, and a small but significant variance in the growth rate between individuals. The next step is to examine what proportion of the variance in the random intercept and random slope is explained by initial wealth. The launch model of developmental change is adapted to analyse how differences in wealth in early childhood affect differences in development over the course of childhood. Household wealth, measured at one year of age, is modelled as the antecedent variable that "launches" the developmental trajectory. It is assumed that a higher wealth status at age one launches a higher vocabulary score at five years of age and a higher level of the developmental trajectory than a lower initial score. Wealth is therefore modelled as a level-2 time-invariant predictor. Since MLM partitions the variance according to its level, it is possible to ask two questions. First, how much of the variance in each of the fixed effects parameters (i.e. initial status and the growth rate) is explained by initial wealth? Secondly, how much of the between-child variance in initial status and growth is explained by initial wealth?

The results of the launch model are shown in Table 5.9 for the full sample. Drawing on the findings of the previous analysis, the parameter estimates were obtained by estimating the linear random intercept and random slope model. The intercept can be interpreted as the grand mean scaled score at onset, controlling for wealth (γ_{00}= 355, $p < 0.01$). The average linear growth rate is 8 points per month, controlling for wealth (γ_{10}= 8.24, $p < 0.01$). Wealth predicts the initial level of cognitive ability at age five (γ_{01} = 204.62, $p < 0.01$). When initial wealth takes a value of zero, the ability intercept is 204 points. For a unit increase in the initial wealth index score (W_i), the intercept increases by 204.62 (i.e. 204.62 x W_i). Initial wealth does not predict the growth rate (γ_{11} is not significant).

Table 5.9 Launch model estimates, full sample (n=1,292)

	Estimate	S.E
Intercept ($\boldsymbol{\gamma_{00}}$)	355.34***	4.48
Slope ($\boldsymbol{\gamma_{10}}$)	8.24**	3.42
Intercept on W1 ($\boldsymbol{\gamma_{01}}$)	204.62***	7.99
Slope on W1 ($\boldsymbol{\gamma_{11}}$)	0.197	0.97
Age 5 ($\boldsymbol{\beta_1}$)	54.16***	5.14
Age 8 ($\boldsymbol{\beta_2}$)	38.43***	2.86
Age 12 ($\boldsymbol{\beta_3}$)	20.56***	3.08
Age 15 ($\boldsymbol{\beta_4}$)	14.02***	3.17
Level-1 residual variances		
Age 5 ($\boldsymbol{\varepsilon_{5i}}$)	2888.44***	173.24
Age 8 ($\boldsymbol{\varepsilon_{8i}}$)	2905.44***	121.08
Age 12 ($\boldsymbol{\varepsilon_{12i}}$)	740.23***	47.55
Age 15 ($\boldsymbol{\varepsilon_{15i}}$)	435.14***	65.50
Level-2 variance components		
Intercept variance ($\boldsymbol{\sigma_0^2}$)	849.64***	108.37
Intercept*slope covariance ($\boldsymbol{\sigma_{10}}$)	-32.29**	12.62
Slope variance ($\boldsymbol{\sigma_1^2}$)	14.58***	2.01
R-square		
ppvt_5	0.478***	0.023
ppvt_8	0.462***	0.018
ppvt_12	0.782***	0.013
ppvt_15	0.880***	0.017
intercept	0.635***	0.036
slope	0.000	0.001
Goodness of fit		
-2LL	54338.87	
RMSEA†	0.068	
CFI‡	0.967	
TLI‡	0.955	

*Correlation coefficient is statistically significant *p=0.1 **p=0.05 ***p=0.01. †The RMSEA ranges from 0 to 1. A value of 0.06 or less is indicative of good model fit; up to 0.08 is acceptable. ‡The CFI and TLI range from 0 to 1. A value of 0.90 or higher indicates acceptable model fit.

The level-1 variance is decreasing with time and vastly different at each sweep. Initial wealth explains a large proportion of the intercept variance at 63% ($\sigma^2_{0,\mathrm{base}}$ = 2,311.36, p < 0.01, $\sigma^2_{0,\mathrm{launch}}$ = 849.64, p < 0.01). The size and significance of the slope variance compares to the base model results (σ_1^2= 14.58, p < 0.01). Thus, initial wealth status does not explain growth in vocabulary ability. The covariance between initial status and growth is also comparable to the base model (σ_{10} = -32.29, p < 0.05).

Table 5.10 reports the estimates for each language group separately. Across groups, wealth at age one predicts the initial level of cognitive ability at five years, with significantly higher onset performance in the Amharic language version of the test for the same level of wealth (γ_{01} = 228.29, p < 0.01). The result is consistent with the test having performed closest to expectations in Amharic. Moreover, in Amharic, for every unit increase in wealth, the growth rate decreases by 4.62 points. (p < 0.01). Initial wealth also explains 54% of the covariance between initial status and growth in Amharic ($\sigma_{10,\text{launch}}$ = -32.85, p<0.01; $\sigma_{10,\text{base}}$ = -71.29, p<0.01). Therefore, wealth status in early childhood helps to explain the convergence in the growth trajectories between top and bottom wealth quintiles.

In Amharic, wealth at age one year explains 74% of the variance in the intercept ($\sigma^2_{0,\text{launch}}$= 801.27, p < 0.01; $\sigma^2_{0,\text{base}}$ = 3126.90, p < 0.01). For the Oromifa subsample, initial wealth explains only 11% of the variance ($\sigma^2_{0,\text{launch}}$= 802.03, p < 0.01; $\sigma^2_{0,\text{base}}$ = 901.46, p < 0.01), and 15% of the intercept variance in Tigrinya ($\sigma^2_{0,\text{launch}}$= 1150.43, p < 0.01; $\sigma^2_{0,\text{base}}$ = 1354.78, p < 0.01). Therefore, while wealth in early childhood accounts for a large proportion of the ability gap at onset in Amharic, it does not fully explain individual differences.

Table 5.10 Launch model estimates, by language group

	Amharic (n=681)		Oromifa (n=250)		Tigrinya (n=361)	
	Estimate	S.E	Estimate	S.E	Estimate	S.E
Intercept (γ_{00})	349.24***	6.30	356.54***	9.73	336.43***	7.38
Slope (γ_{10})	20.16***	4.28	8.56	6.89	1.12	5.86
Intercept on W1 (γ_{01})	228.29***	8.87	105.65***	28.51	119.17***	27.64
Slope on W1 (γ_{11})	-4.62***	1.09	3.83	3.65	-0.21	2.84
Age 5 (β_1)	53.05***	7.11	50.65***	11.51	62.96***	9.05
Age 8 (β_2)	33.58***	3.64	24.43***	5.96	46.15***	5.05
Age 12 (β_3)	12.49***	3.86	17.36***	6.30	24.62***	5.37
Age 15 (β_4)	4.84	3.97	11.84*	6.46	18.82***	5.48
Level-1 residual variances						
Age 5 (ε_{5i})	2764.15***	222.89	2239.56***	284.77	2942.84***	359.22
Age 8 (ε_{8i})	2789.60***	170.38	1394.73***	151.38	2927.54***	210.11
Age 12 (ε_{12i})	640.65***	58.03	1015.32***	124.49	406.43***	66.63
Age 15 (ε_{15i})	520.37***	79.09	682.62***	160.68	589.87***	119.01
Level-2 variance components						
Intercept variance (σ_0^2)	801.27***	142.59	802.03***	226.96	1150.43***	211.04
Intercept*slope covariance (σ_{10})	-32.85**	16.95	-26.82	24.07	-38.24*	22.64
Slope variance (σ_1^2)	10.22***	2.52	8.58***	3.44	7.77**	3.34
R-square						
ppvt_5	0.554***	0.027	0.342***	0.066	0.385***	0.047
ppvt_8	0.515***	0.023	0.402***	0.049	0.345***	0.034
ppvt_12	0.802***	0.016	0.514***	0.038	0.757***	0.034
ppvt_15	0.837***	0.024	0.671***	0.072	0.700***	0.052
intercept	0.749***	0.037	0.146**	0.072	0.129**	0.054
slope	0.087**	0.043	0.021	0.041	0.000	0.002
Goodness of fit						
RMSEA†	0.061		0.055		0.091	
CFI‡	0.977		0.951		0.902	
TLI‡	0.969		0.933		0.866	

*Correlation coefficient is statistically significant *p=0.1 **p=0.05 ***p=0.01. †The RMSEA ranges from 0 to 1. A value of 0.06 or less is indicative of good model fit; up to 0.08 is acceptable. ‡The CFI and TLI range from 0 to 1. A value of 0.90 or higher indicates acceptable model fit.

5.3.3 The change-to-change model of development

The findings from section 5.3.1 and 5.3.2 showed that for all language groups there is substantial intercept variance to be explained and a relatively small amount of slope variance. Wealth, measured when the children were aged one, explains 74% of the intercept variance in Amharic, 11% in Oromifa and 15% in Tigrinya. The Amharic results are consistent with developmental theory and empirical evidence, which posits that a child's standard of living in the earliest years of life is an important predictor of cognitive outcomes in early childhood. In Oromifa and Tigrinya, the results suggest that other influences are at play. Initial wealth predicts the growth rate in Amharic only. However, it does not explain variability in the growth rate. In Amharic, initial wealth explains 54% of the covariance between initial status and growth, i.e. the widening gap between the poorest and richest children.

As individual differences are observed for all growth factors, i.e. the intercept, the growth rate, and the relationship between the two, the final step is to examine whether changes in wealth over time further help to explain differences in developmental patterns. The change-to-change model of development is adapted to analyse associations between the pattern of change in vocabulary ability and changes in living standards. The model assumes that the developmental trajectory is influenced not simply by the initial level of a predictor variable but also but the pattern of change in that variable. Of interest is whether an increase in wealth corresponds to an improvement in the developmental trajectory, i.e. narrowing of the gap between wealth quintiles, and conversely, whether a decrease in wealth is associated with developmental stagnation or regression.

Building on the findings of the launch model, individual differences in progress are modelled as a function of both initial wealth, a level-2 time-invariant predictor, and wealth measured at each time point, a level-1 time-varying predictor. In other words, the model examines the effect of wealth at age one on onset ability and the growth rate, as well as the effect of wealth status at each subsequent stage of life on the scaled score at each time point. Wealth is grand mean centred (centred on its expected value).

The results of the change-to-change model are shown in Table 5.11 for the full sample. Drawing on the findings of the previous analysis, the parameter estimates were obtained by estimating the linear random intercept and random slope model. The intercept and slope are regressed on initial wealth, measured at age one. The scaled score at each time point is regressed on the child's age and wealth at each time point. The expected value of the PPVT scaled score adjusted for wealth is 335 points ($p < 0.01$). The average linear growth rate is 9.5 points per period, controlling for wealth ($p < 0.01$).

The findings are consistent with the launch model of development. Wealth at age one predicts the ability intercept at age five (γ_{01} = 171.27, $p < 0.01$). For a unit increase in the initial wealth index score (W_i), the intercept increases by 171.27 (i.e. 171.27 x W_i). It does not predict the growth rate (γ_{11} is not significant). The fixed effect of a continuous variable modelled at level-1 can be interpreted as unit change in the outcome variable associated with a unit change in the predictor variable. The fixed effect of wealth modelled at level-1 represents change in the PPVT scaled score associated with a unit change in wealth. The effect of wealth on the scaled score is smallest at five years (θ_t = 25 points, $p < 0.01$) and largest at eight years (θ_t = 67 points, $p < 0.01$).

The inclusion of wealth as a time varying predictor does not improve model fit, evidenced by poorer goodness of fit indices in comparison to the launch model. Moreover, including wealth as a time varying covariate only explains an additional 3% of the intercept variance. It does not explain the variance in the growth rate or the relationship between the intercept and growth. Taken together, the findings suggest that the data are better described by the launch model of developmental change.

Table 5.11 Change-to-change model estimates, full sample (n=1,279)

	Estimate	S.E
Intercept (γ_{00})	353.46***	4.47
Slope (γ_{10})	9.41***	3.30
Intercept on W1 (γ_{01})	171.27***	12.07
Slope on W1 (γ_{11})	-0.40	1.45
Age 5 (β_1)	53.20***	5.16
Age 8 (β_2)	36.75***	2.76
Age 12 (β_3)	18.82***	2.97
Age 15 (β_4)	12.25***	8.28
Wealth 5 (θ_1)	25.52***	12.89
Wealth 8 (θ_2)	67.02***	10.53
Wealth 12 (θ_3)	56.62***	7.55
Wealth 15 (θ_4)	62.23***	8.28
Level-1 residual variances		
Age 5 (ε_{5i})	2871.16***	171.03
Age 8 (ε_{8i})	2920.15***	125.28
Age 12 (ε_{12i})	751.61***	47.44
Age 15 (ε_{15i})	445.10***	64.43
Level-2 variance components		
Intercept variance (σ_0^2)	797.56***	108.93
Intercept*slope covariance (σ_{10})	-38.72***	12.64
Slope variance (σ_1^2)	14.47***	1.99
R-square		
ppvt_5	0.456***	0.025
ppvt_8	0.482***	0.020
ppvt_12	0.773***	0.014
ppvt_15	0.873***	0.018
intercept	0.562***	0.046
slope	0.000	0.003
Goodness of fit		
-2LL linear	53812.66	
RMSEA†	0.083	
CFI‡	0.927	
TLI‡	0.901	

*Correlation coefficient is statistically significant *p=0.1 **p=0.05 ***p=0.01. †The RMSEA ranges from 0 to 1. A value of 0.06 or less is indicative of good model fit; up to 0.08 is acceptable. ‡The CFI and TLI range from 0 to 1. A value of 0.90 or higher indicates acceptable model fit.

Table 5.12 shows the results by language group. In all three models, the goodness of fit is reduced by adding wealth as a time varying covariate, apart from Tigrinya for which data-model fit is the same. The initial wealth effect on onset ability remains positive and significant but is reduced in all three models. The negative effect of initial wealth on the slope is no longer significant in Amharic once wealth is included as a predictor at each time point. The fixed effect for wealth is large and significant at all time points in Amharic ($p < 0.01$). In Tigrinya, the including of wealth as a level-1 predictor halves the initial wealth effect on the intercept, due to the strong influence of the children's wealth status at age five (θ_t = 105 points, $p < 0.01$).

Table 5.12 Change-to-change model estimates, by language group

	Amharic (n=670)		Oromifa (n=250)		Tigrinya (n=359)	
	Estimate	S.E	Estimate	S.E	Estimate	S.E
Intercept (γ_{00})	338.36***	6.23	335.23***	9.92	331.60***	8.29
Slope (γ_{10})	19.86***	4.11	9.55	6.55	0.67	5.67
Intercept on W1 (γ_{01})	163.45***	16.53	96.82***	33.06	53.06*	27.64
Slope on W1 (γ_{11})	-1.48	1.95	0.863	4.18	-2.07	3.49
Age 5 (β_1)	52.19***	7.17	49.23***	11.46	62.49***	8.98
Age 8 (β_2)	32.35***	3.48	22.25***	5.64	43.75***	4.80
Age 12 (β_3)	12.28***	3.70	15.22***	5.96	23.43***	5.13
Age 15 (β_4)	4.66	3.81	9.68	6.11	17.95***	5.24
Wealth 5 (θ_1)	55.15***	18.15	22.02	26.06	104.87***	27.79
Wealth 8 (θ_2)	107.48***	14.22	21.22	22.88	27.08	21.07
Wealth 12 (θ_3)	58.14***	9.64	46.43**	21.55	57.32***	11.25
Wealth 15 (θ_4)	57.90***	10.79	94.53***	22.77	62.14***	13.52
Level-1 residual variances						
Age 5 (ε_{5i})	2762.46***	218.40	2202.36***	284.77	2743.45***	339.08
Age 8 (ε_{8i})	2626.22***	168.68	1401.21***	151.38	2927.54***	220.45
Age 12 (ε_{12i})	649.16***	56.88	1032.10***	124.49	406.43***	67.30
Age 15 (ε_{15i})	530.44***	78.08	711.12***	160.68	589.87***	116.92
Level-2 variance components						
Intercept variance (σ_0^2)	717.47***	141.98	852.08***	226.96	1132.23***	197.88
Intercept*slope covariance (σ_{10})	-36.60**	16.71	-43.18*	24.07	-52.72**	22.09
Slope variance (σ_1^2)	10.34***	2.49	8.93***	3.44	9.72**	3.28
R-square						
ppvt_5	0.523***	0.029	0.360***	0.068	0.426***	0.046
ppvt_8	0.586***	0.026	0.386***	0.051	0.306***	0.033
ppvt_12	0.791***	0.017	0.473***	0.038	0.742***	0.035
ppvt_15	0.832***	0.024	0.646***	0.072	0.710***	0.052
intercept	0.629***	0.060	0.119*	0.074	0.029	0.032
slope	0.010	0.025	0.001	0.010	0.005	0.018
Goodness of fit						
RMSEA†	0.065		0.080		0.091	
CFI‡	0.960		0.854		0.857	
TLI‡	0.946		0.802		0.807	

*Correlation coefficient is statistically significant *p=0.1 **p=0.05 ***p=0.01. †The RMSEA ranges from 0 to 1. A value of 0.06 or less is indicative of good model fit; up to 0.08 is acceptable. ‡The CFI and TLI range from 0 to 1. A value of 0.90 or higher indicates acceptable model fit.

Changes in wealth over time does not help to explain individual differences in developmental trajectories. Including wealth as a time varying covariate explains an additional 10% of the intercept variance in Amharic, only 2% in Tigrinya and it does not explain any additional intercept variance in Oromifa. This is expected as wealth measured in later childhood is not a true antecedent variable in relation to the intercept. However, wealth does not explain the variance in the slope either. The slope variance remains significant and the amount of variance to be explained is unchanged.

5.4 Summary

This section examined the time effects of poverty on children's vocabulary development using the PPVT scores that were scaled according to the Rasch methodology and two distinct measures of wealth: the children's initial wealth status measured at age one year wealth measured at each administration of the PPVT (i.e. age five, eight, 12 and 15). Two theoretical models were considered: (i) in the first approach, wealth status during the first year of life is modelled as "launching" the vocabulary trajectory and (ii) in the second model changes in wealth status are thought to influence the trajectories. Prior to investigating the models statistically, a base model was estimated to observe the amount of variance in the intercepts, slopes, and the shape of the growth curves. The results of the base model confirmed large variability in the intercepts, a small but significant variance in the slopes among the Amharic subsample (but not in Tigrinya and Oromifa) and showed evidence of a linear growth function as opposed a quadratic specification.

The launch model addresses the following key questions: (i) how much of the variance in each of the fixed effects parameters, i.e. initial status and the growth rate, is explained by initial wealth, and (ii) how much of the between-child variance in initial status and growth is explained by initial wealth? Analysis of the full sample data revealed that initial wealth predicted the onset vocabulary score but not the growth rate over time. However, discrepancies emerge when examining the data by language subgroup. Across subgroups, initial wealth has a large and significant effect on the onset score. Initial wealth predicts the slope only in Amharic such that an increase in wealth at one year is associated with a decrease in the slope for this subgroup, thereby narrowing the gap between the bottom and top quintiles. In Oromifa and Tigrinya, the slope factor was not significant (more homogeneous samples) and initial wealth therefore had no effect on the slope. The launch model explains 74% of the variability in the intercepts in Amharic, 11% in Oromifa and 15% in Tigrinya, suggesting that wealth at time one is not the most efficient predictor of vocabulary across the whole sample of children.

The change-to-change model addresses the following key question: do changes in wealth over time further help to explain individual differences in development? Although intuitive, the main limitation with the model is the small slope variance to be explained. The fixed effect of wealth is significant across subsamples. Moreover, the mean effect of wealth at each time point on the scaled score is positive and significant, with idiosyncrasies across groups. However, changes in wealth over time does not explain individual differences in the growth factors. Low variance among the key study variable restricts the possibility of answering these questions fully.

Taken together, the findings suggests that the data are better described by the launch model of developmental change, which conforms with developmental theory and empirical evidence, i.e. the earliest years of life are crucial for the development of cognitive processes, and a child's standard of living in the earliest years of life is an important predictor of children's cognitive outcomes. However, the small variation in the growth rate may also be due, in part, to statistical measurement error resulting from the large ceiling effect observed at age 12 and 15, and hence, low variance in the slope factor.

Chapter 6 Pathways through which wealth influences vocabulary development

6.1 Introduction and rationale

This chapter addresses the third and final research question. It investigates the extent to which the relationship between wealth and children's vocabulary development is explained by the children's nutritional status, their home environment and preschool exposure. The variables are presented in Table 6.1 and discussed in more detail in the methodology chapter (section 3.4.2). The purpose of this final stage of analysis is to identify the pathways through which poverty affects children's vocabulary development with a view to identify interventions to address the wealth gap at different stages of childhood. Given the empirical evidence discussed in the literature review, it is expected that the effect of wealth on PPVT growth is partially mediated by each of the between-child mediating variables.

Table 6.1 Mediation model variables

Level	Variable
Outcome	PPVT scaled score
Main predictor	Wealth index score
Between-child health and nutrition mediator	Height-for-age z-score/ BMI-for-age z-score
Between-child home environment mediator	Mothers' educational level
Between-child preschool mediator	Participation in preschool education

In the previous chapter, two statistical models were estimated to examine the magnitude and significance of the variance in onset ability and the growth rate explained by each model. The next step is to identify the mechanisms through which wealth influences onset ability and the growth rate. A well-fitting launch model of development implies that the wealth gap is predicted by individual differences in the children's early experiences and contexts. The focus of the mediation study would then be on characterising the poverty-vocabulary relationship in earlier childhood using quantitative information on the children's wider environments at one and five years of age. Such findings would suggest that to promote children's development and future learning in Ethiopia, early interventions that address initial gaps in cognitive ability are needed. Alternatively, if the data are more efficiently explained by a change-to-change model of development, the mediation study should consider the pathways through which living standards influence development at later stages of childhood as well.

The results of the previous analysis supported a launch model of developmental change. The findings showed substantial individual differences in the children's initial starting points. Individual differences in the growth rate were found to be relatively small. Initial wealth, measured at one year of age, explained a large proportion of the variance in onset ability (Table 6.2). Due to the small slope variance, neither the launch nor the change-to-change models can explain individual differences in the growth rate. The current chapter therefore seeks to explain this onset variation across individuals by investigating the mechanisms through which initial wealth influences vocabulary outcomes. Drawing on the theoretical research and previous empirical findings (Bradley et al., 2001; Brooks-Gunn et al., 2000; Magnuson and Duncan, 2001; McLeod, 1990; Guo and Harris, 2000), factors known to influence vocabulary ability and vary with wealth are modelled as mediator variables.

Table 6.2 Goodness of fit and R-square for each model (full sample)

Measure	Base model		Launch model		Change model	
	Estimate	S.E	Estimate	S.E	Estimate	S.E
R-square						
ppvt_5	0.483***	0.024	0.478***	0.023	0.456***	0.025
ppvt_8	0.450***	0.018	0.462***	0.018	0.482***	0.020
ppvt_12	0.792***	0.015	0.782***	0.013	0.773***	0.014
ppvt_15	0.869***	0.019	0.880***	0.017	0.873***	0.018
intercept	N/A		0.635***	0.036	0.562***	0.046
slope	N/A		0.000	0.001	0.000	0.003
Goodness of fit						
-2LL	57758.45		54338.87		53812.66	
RMSEA†	0.067		0.068		0.083	
CFI‡	0.962		0.967		0.927	
TLI‡	0.951		0.955		0.901	

*Correlation coefficient is statistically significant *p=0.1 **p=0.05 ***p=0.01. †The RMSEA ranges from 0 to 1. A value of 0.06 or less is indicative of good model fit; up to 0.08 is acceptable. ‡The CFI and TLI range from 0 to 1. A value of 0.90 or higher indicates acceptable model fit.

6.2 Descriptive statistics

The contextual variables used in the analysis are summarized in Table 6.3.

Environmental factors that are important for children's developmental outcomes include biological characteristics, experiences during pregnancy and infancy, home influences such as parental sensitivity and responsiveness in early childhood; child-caregiver relationships throughout childhood; as well as access to cognitive stimulation and other developmentally enriching activities. However, the current analysis is limited to the information available in the Young Lives data set. The data structure does include some information on pregnancy, delivery, breastfeeding practices, and infant health. However, this information is narrow, based on caregiver reports and retrospectively collected, i.e. recorded when the children were one year old.

Table 6.3 Contextual variables used in the analysis

Variable	Description	Values	Measurement
Height-for-age score	Variable measuring the degree of stunting	-5.36 to 2.64 (4 z-score units from the observed mean)	Interval
BMI-for-age score	Variable measuring the degree of wasting or thinness	-5.20 to 2.80 (4 z-score units from the observed mean)	Interval
Mothers' education	Variable representing the number of years of formal schooling	0.00: no formal schooling 1.00 to 12.00: grades 1 to 12 13.00: post-secondary 14.00: university	Ordinal
Household size	Variable representing the number of individuals living in the child's household	0.00 to 15.00	Ordinal
Preschool attendance	Variable representing a child's preschool participation	0 if no; 1 if yes	Nominal
Gender	Child's gender	1 if male; 2 if female	Nominal
Geographic stratum	Child's urban or rural residence	1 if urban; 2 if rural	Nominal

*Normally, it is recommended to assess height-for-age in children less than two years of age, and to transition to BMI for older children and adults (WHO, 1996).

Table 6.4 presents the mean and standard deviation of the variables used in the analysis. The importance of each underlying construct is discussed in the literature review section. The measures are discussed in the methodology (section 3.4.2 and 3.5.2). At the start of the survey, 6.4% of one-year olds were reported to be stunted based on the height-for-age z-score and 5.6% showed evidence of wasting/thinness according to the BMI-for-age z-score, using the WHO threshold of -2 SD. By age 15, the prevalence of stunting drops to less than 1% and thinness increases to 16%. Around 60% of mothers have no formal education when their children are infants. The proportion decreases to 36% by the end of the survey period. Overall, mothers complete three years of schooling. Around one third of the children in the sample attended some form of preschool (29%).

Table 6.4 Descriptive statistics for key study variables

	Age 1		Age 5		Age 8		Age 12		Age 15	
Variable	Mean (SD)	N	Mean (SD)	N	Mean (SD)	N	Mean (SD)	N	Mean (SD)	N
PPVT scaled	-	-	378.49 (70.35)	1310	503.91 (80.04)	1320	550.67 (58.40)	1319	565.99 (60.09)	1321
Wealth index	0.25 (0.19)	1296	0.32 (0.19)	1305	0.37 (0.19)	1319	0.41 (0.18)	1316	0.44 (0.17)	1316
Height-for-age	-1.59 (1.92)	1272	-1.38 (1.10)	1310	-1.10 (1.13)	1319	-1.42 (0.98)	1317	-1.31 (1.10)	1319
BMI-for-age	-0.52 (1.37)	1202	-0.67 (1.12)	1310	-1.33 (1.04)	1319	-1.83 (1.02)	1317	-1.61 (1.17)	1317
Mothers' education	2.37 (3.67)	1272	2.80 (3.90)	1167	2.97 (3.95)	1109	3.29 (4.08)	1038	3.45 (4.21)	1021
Household size	5.73 (2.14)	1310	6.02 (2.08)	1310	6.14 (1.20)	1320	5.82 (1.93)	1318	5.65 (1.92)	1318
Male	55%	1310	55%	1310	55%	1320	55%	1319	55%	1321
Went to preschool	-	-	29%	1310	29%	1310	29%	1310	29%	1310
Residence is urban	41%	1310	41%	1310	41%	1320	44%	1318	43%	1317

6.2.1 Nutritional factors

The data set includes two composite measures of health and nutrition, both of which are computed and scaled based on the WHO standard. The first measure, the height-for-age z-score (HAZ), provides an indication of stunting, i.e. low skeletal growth. Stunting is normally associated with poor overall economic conditions, chronic or repeated infections, and inadequate nutrient intake (WHO, 1986). The second measure, the body mass index

(BMI)-for-age z-score, is an indicator of wasting which refers to a deficit in tissue and fat mass. Wasting is associated with disease or negative economic shocks affecting the family food supply. As stunting is thought to be determined by 24-36 months, at which time it stabilises (Victoria et al., 2008; WHO, 1986), height-for-age is commonly used to assess children under two years of age. BMI is normally used for older children and adults (WHO, 1986). Stunting has been identified as one of the most proximal risk factors encountered in the first five years of life in low-income countries, with long-term implications for human development (Walker et al., 2007; Grantham-McGregor et al., 1999a; Grantham-McGregor et al., 1999b; Walker et al., 2005).

The distributions of the z-scores are equivalent due to the common scaling procedure. The low correlation between variables illustrates that the measures perform differently (Table 6.5). Wasting and stunting often show no association in populations, and present different patterns at different ages and in different populations (WHO, 1986). It has been shown that weight-for-height and height-for-age account for 95% of the variation in weight-for-age (WHO, 1986), and the use of both indices is recommended. Moreover, as BMI-for-age reflects an individual's current health status, it "may enable improved monitoring of longitudinal growth patterns in young children" (Furlong et al., 2016: 7).

Table 6.5 Height-for-age and BMI-for-age correlations across time, full sample

	HAZ 1	HAZ 5	HAZ 8	HAZ 12	HAZ 15	BMI 1	BMI 5	BMI 8	BMI 12	BMI 15
HAZ 1	1									
HAZ 5	.426**	1								
HAZ 8	.440**	.655**	1							
HAZ 12	.408**	.634**	.749**	1						
HAZ 15	.373**	.543**	.635**	.724**	1					
BMI 1	-.062*	.220**	.151**	.135**	.135**	1				
BMI 5	.011	-.277**	.017	.048	-.003	.134**	1			
BMI 8	.032	.086**	-.137**	.096**	.055*	.234**	.433**	1		
BMI 12	.137**	.165**	.157**	.173**	.140**	.284**	.452**	.582**	1	
BMI 15	.152**	.182**	.214**	.267**	.278**	.234**	.305**	.435**	.679**	1

*Correlation coefficient is statistically significant **p=0.01 *p=0.05

According to the World Health Organization, cut-offs are needed to identify outliers resulting from errors in data entry errors or measurement (WHO, 1995). The WHO recommends alternative approaches to handling outliers, depending on the observed mean z-score. For observed mean scores of less than -1.5, observations are removed if the z-score lies outside of the range of four z-score units from observed mean score, with a maximum score of 3.0 (CDC, 2013). When the observed mean z-score is above -1.5, the cut-offs are defined as <-5.0 and >+3.0 (CDC, 2013). For BMI, the z-score cut-offs follow the weight-for-height thresholds of <-4.0 and >+5.0 (CDC, 2013).

Table 6.6 presents nutrition-related child outcomes by wealth quintile. On average, the difference in the HAZ between the top and bottom wealth quintiles is 0.80 SD. The difference is largest when the children are aged one year, at which time children in the top quintile are one standard deviation taller, on average. HAZ increases until age eight and then declines from eight years onward (0.69 SD at 15 years). In contrast, the mean difference in BMI-for-age increases from 0.67 SD when the children are aged one year to 0.83 SD by age 15.

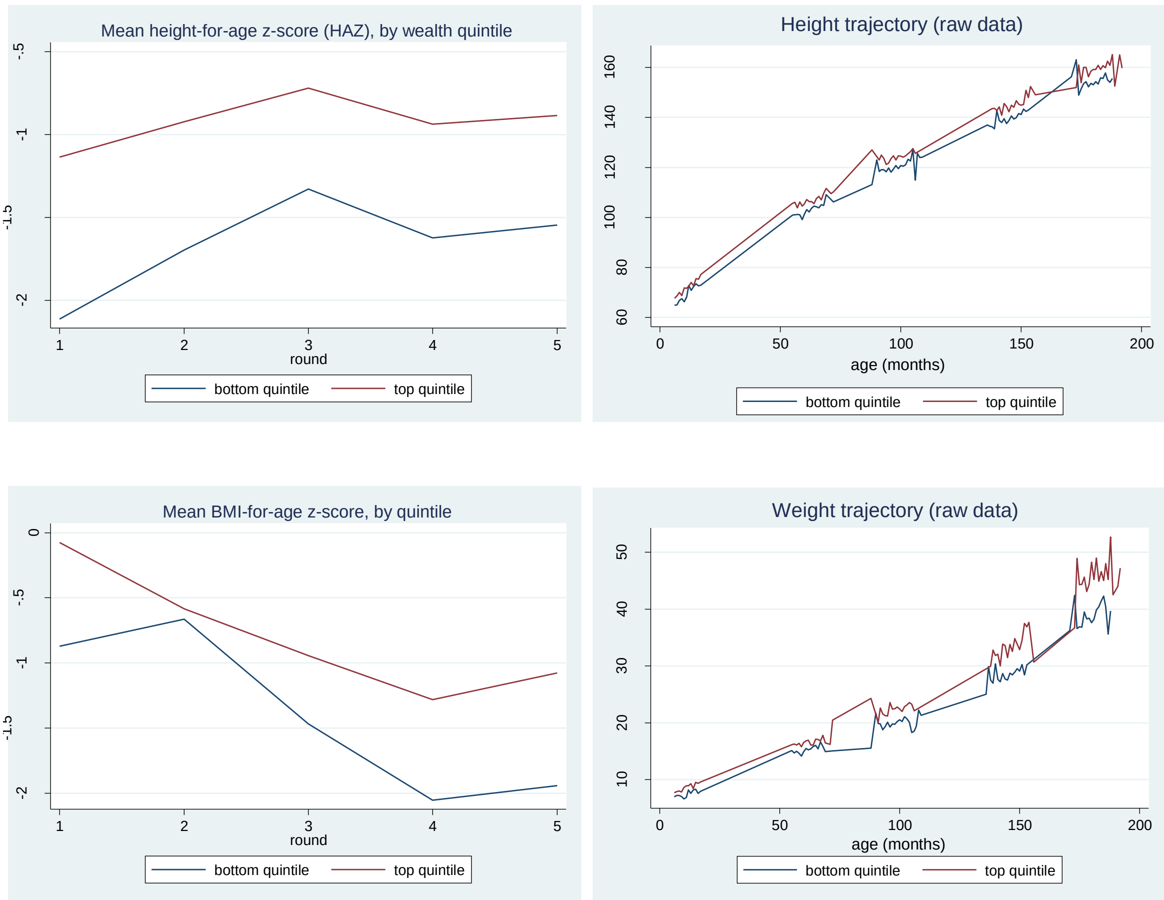
Table 6.6 Child outcomes by wealth disparity, full sample

	Bottom quintile			Top quintile		
Variables	mean	sd	n	mean	sd	n
Wealth index age 1	0.09	0.06	595	0.62	0.06	124
Wealth index age 5	0.11	0.05	377	0.64	0.08	202
Wealth index age 8	0.11	0.05	243	0.65	0.08	275
Wealth index age 12	0.12	0.05	118	0.65	0.08	337
Wealth index age 15	0.13	0.05	78	0.66	0.08	372
PPVT scaled score age 1	-	-	-	-	-	-
PPVT scaled score age 5	349.39	53.26	377	437.80	81.85	202
PPVT scaled score age 8	461.04	62.34	243	574.11	68.18	275
PPVT scaled score age 12	500.17	35.31	118	601.60	48.15	337
PPVT scaled score age 15	515.72	44.61	78	612.98	47.00	372
Height-for-age z-score age 1	-1.97	1.88	578	-0.98	1.58	121
Height-for-age z-score age 5	-1.62	1.07	377	-0.93	0.90	202
Height-for-age z-score age 8	-1.42	1.01	243	-0.62	1.13	274
Height-for-age z-score age 12	-1.65	0.78	118	-1.06	1.02	336

Height-for-age z-score age 15	-1.67	1.06	78	-0.98	1.01	370
BMI-for-age z-score age 1	-0.86	1.31	517	0.19	1.24	121
BMI-for-age z-score age 5	-0.77	1.07	377	-0.57	1.03	202
BMI-for-age z-score age 8	-1.55	0.90	243	-1.00	1.17	274
BMI-for-age z-score age 12	-2.20	0.96	118	-1.43	1.09	336
BMI-for-age z-score age 15	-2.10	0.97	78	-1.25	1.20	370

On average, the disparity in the PPVT scaled score widens by 10 points over the period, from 88 points between quintiles at age five to 98 points by age 15. The positive correlation between the height-for-age z-score and the PPVT scaled score is small but significant overall (r = 0.18, p <0.01) BMI-for-age is negatively correlated with the cognitive scores (r = -0.17, p <0.01). On average, the thinner the child the lower the cognitive score. The mean difference in the nutritional trajectories between the quintiles is depicted in Figure 6.1. The difference is significant at each time point at the 1% level.

Figure 6.1 Mean differences in nutritional trajectories by wealth quintile, full sample



6.2.2 Mothers' education and preschool

Mothers' education is measured in the Young Lives survey to characterise the socioeconomic conditions of the home. For lack of better alternative, in the current study it is used to proxy maternal literacy, and in turn, the quality of the home environment. There is broad evidence that maternal literacy influences the development of language skills among children (Bradley and Caldwell, 1984). Children's learning is further shaped by a range of other factors in the home, some of which are discussed in Chapter 2, including: parental warmth and affection towards children, communication and interactions, the availability of stimulating learning materials in the home, and parent's engagement in play activities. Information on these characteristics are not available in the Young Lives survey.

Mothers have completed an average of three years of formal schooling, although the variance is large (m = 2.94, sd = 3.97). Table 6.7 shows mothers' educational participation by wealth quintile. The vast majority of mothers in the poorest households have no formal education, 78% compared with 10% of mothers in the top quintile. Only 3% of mothers in the bottom quintile have attended secondary school, compared with 30% of mothers in the top quintile. None of the mothers living in the poorest households have attended post-secondary education, whereas 8% of the richest mothers have. Mothers' education is positively and strongly correlated with the wealth index score ($r = 0.62$, $p < 0.01$) and moderately correlated with the PPVT scaled score ($r = 0.38$, $p < 0.01$).

Table 6.7 Mothers' educational attainment by wealth disparity, full sample

Mothers' educational attainment	Bottom quintile		Top quintile	
	n	%	n	%
No formal schooling	831	78.2	134	10.4
Primary Grade 1-4	169	15.9	155	12.0
Primary Grade 5-8	59	5.6	421	32.5
Secondary, Grade 9-12	3	0.3	382	29.5
Vocational	0	0.0	77	6.0

University	0	0.0	26	2.0
N	1062	79%	1195	92%
Missing cases	278	100%	99	100%

Research has shown that preschool education has multiple benefits including improved academic performance and improved behavioural outcomes (Sylva et al., 2014; Sammons et al., 2014). Moreover, children whose parents have low educational outcomes experience the most positive effects on cognitive outcomes (Sylva et al., 2014). For children experiencing family risk (e.g. poverty) or individual risk (e.g. low birth weight) in early childhood, quality pre-schooling moderates the effects of experienced risk on cognitive outcomes (Hall et al., 2009). The data for the present study shows that in the poorest quintile, 98% of the children for whom preschool information were recorded have never been to preschool (n=268). In the top quintile, only 14% have never attended. Analysing the effects of preschool attendance on the cognitive outcomes of the poorest children in the sample will not be possible as no children in the bottom quintile attended preschool. However, it will be important to examine the preschool effect on the wealth-growth relationship overall.

6.3 Mediated effects of wealth on children's vocabulary development

Drawing on the findings of the previous analysis, the parameter estimates were obtained by estimating the linear launch model with both intercept and slope random effects. Given the large and significant intercept variance, we are particularly interested in explaining individual differences in the children's initial starting points at five years of age. There is little slope variance to explain, as evidenced in the results of the base model. Details about how the models are specified are discussed in the methodology chapter (section 3.5). For ease of reference, the general mediation model discussed in Chapter 3 is specified again as follows:

$$PPVT_{ti} = \gamma_{00} + \gamma_{10}AGE_{ti} + \gamma_{01}W_i + \gamma_{11}W_iAGE_{ti} + \gamma_{02}M_i + \gamma_{12}M_iAGE_{ti} + u_{0i} + u_{1i} + \varepsilon_{ti}$$

$$M_i = \gamma_{20} + \gamma_{21} W_i + u_{2i}$$

To reiterate, the parameters γ_{00} and γ_{10} are the expected level of initial ability and the rate of change per month. The parameters γ_{01} and γ_{11} represents the mean difference in initial ability, controlling for W_i. Similarly, γ_{02} and γ_{12} represent the mean effect on the intercept, controlling for the mediator, M_i. The parameter γ_{20} is the intercept of the mediator variable and γ_{21} is the mean effect of initial wealth on the mediator variable. The residuals are u_{0i}, u_{1i}, and u_{2i}. Including age in months as a time-varying covariate results in the modification shown below. The slopes for "age in months" vary over time points rather than individuals. This fixed effect parameter is specified as β_t.

$$PPVT_{ti} = \gamma_{00} + \gamma_{10} AGE_{ti} + \beta_t AGE_{ti} + \gamma_{01} W_i + \gamma_{11} W_i AGE_{ti} + \gamma_{02} M_i + \gamma_{12} M_i AGE_{ti} + u_{0i} + u_{1i} + \varepsilon_{ti}$$

$$M_i = \gamma_{20} + \gamma_{21} W_i + u_{2i}$$

6.3.1 Nutrition as a mediating factor

Height-for-age (HAZ) measured at age five, is selected as the mediating nutrition variable. HAZ reflects a child's overall nutritional condition and is widely used as an indicator of social deprivation (WHO, 1986). Low height-for age, or stunted growth, "reflects a process of failure to reach linear growth potential as a result of suboptimal health and/or nutritional conditions" (WHO, 1997: 46). The process of stunting slows down around three years of age. Thus, low height-for-age among older children reflects a state of having failed to grow (WHO, 1997). Low weight-for-height, or wasting, which is usually very low even in poor countries, indicates a severe process of weight loss resulting from a recent severe event (WHO, 1997). Thus, measures such as BMI are particularly useful for understanding longitudinal patterns of children's growth and development (Furlong et al., 2016). In a change-to-change model of children's development, BMI-for-age would be a suitable indicator.

The results are shown for the full sample in Table 6.8. Controlling for HAZ, the expected level of initial ability is 353 points (γ_{00}= 352.44, $p < 0.01$). The average linear growth rate is 8.33 points per month (γ_{10}= 8.33, $p < 0.05$). Overall, wealth at one year of age predicts the initial vocabulary level (γ_{01} = 194.60, $p < 0.01$), but not the growth rate, which is consistent with the launch model estimates (nonsignificant γ_{11}). Height-for-age predicts the initial level (γ_{02} = 8.24, $p < 0.01$), and the growth rate (γ_{12} = -0.43, $p < 0.05$). The effect of HAZ on wealth is positive and significant (γ_{21} = 1.43, $p < 0.01$).

The difference between the value of γ_{01} in the mediated and unmediated launch model represents the indirect effect of the main predictor on the intercept, controlling for the mediator (i.e. c – c' in Figure 3.10). The mediation effect of HAZ on the intercept is large and significant (c – c' = 11.80, $p < 0.01$). In other words, including HAZ in the model reduces the size of the wealth effect on the intercept by 11.80 points. The indirect effect of HAZ on the slope is small but significant (c – c' = -0.62, $p < 0.05$). Recalling from the launch model results, initial wealth explained 63% of the intercept variance. Height-for-age explains an additional 2% of the variance.

Table 6.8 Height-for-age mediated launch model estimates, full sample (n=1,236)

	Estimate	S.E
Intercept (γ_{00})	352.44***	4.54
Slope (γ_{10})	8.33**	3.44
Intercept on W1 (γ_{01})	194.60***	8.21
Slope on W1 (γ_{11})	0.43	1.01
Intercept on HAZ (γ_{02})	8.24***	1.37
Slope on HAZ (γ_{12})	-0.43**	0.17
HAZ on Wealth (γ_{21})	1.43***	0.15
Indirect effect from WI to intercept (c – c')	11.80***	2.31
Indirect effect from WI to slope (c – c')	-0.62**	0.25
Age 5 (β_1)	56.76***	5.20
Age 8 (β_2)	39.10***	2.90
Age 12 (β_3)	20.79***	3.11
Age 15 (β_4)	14.16***	3.20
Level-1 residual variances		
Age 5 (ε_{5i})	2863.03***	175.14
Age 8 (ε_{8i})	2887.46***	122.29
Age 12 (ε_{12i})	745.94***	48.73
Age 15 (ε_{15i})	414.00***	66.18
Level-2 variance components		
Intercept variance (σ_0^2)	807.96***	107.98
Intercept*slope covariance (σ_{10})	-32.35**	12.77
Slope variance (σ_1^2)	14.90***	2.05
HAZ variance (σ_2^2)	1.02***	0.05
R-square		
ppvt_5	0.488***	0.023
ppvt_8	0.467***	0.019
ppvt_12	0.781***	0.013
ppvt_15	0.885***	0.018
HAZ	0.067***	0.013
intercept	0.662***	0.036
slope	0.013	0.010
Goodness of fit		
-2LL linear	55589.33	
RMSEA†	0.064	
CFI‡	0.964	
TLI‡	0.950	

*Correlation coefficient is statistically significant *p=0.1 **p=0.05 ***p=0.01. †The RMSEA ranges from 0 to 1. A value of 0.06 or less is indicative of good model fit; up to 0.08 is acceptable. ‡The CFI and TLI range from 0 to 1. A value of 0.90 or higher indicates acceptable model fit.

Table 6.9 reports the estimates for each language group separately. In both Amharic and Oromifa, HAZ predicts the initial level of vocabulary (Amharic: γ_{02} = 9.84, p < 0.01; Oromifa: γ_{02} = 9.49, p < 0.01), and the growth rate (Amharic: γ_{12} -0.73, p < 0.01; Oromifa: γ_{12} = -0.65, p < 0.05).

HAZ mediates the effects of wealth on the children's initial status, most notably in Oromifa (Amharic: c – c' = 12.82, p < 0.01; Oromifa: c – c' = 21.11, p < 0.01), and the growth rate (Amharic: c – c' = -0.95, p < 0.01; Oromifa: c – c' = -1.44, p < 0.10). However, in Oromifa the indirect effect on the slope is only weakly significant. The result indicates that including HAZ in the model reduces the steepness of the slope in Amharic. In Tigrinya, height-for-age does not mediate the wealth effect, model fit is poor, and the coefficients are not significant.

Height-for-age is particularly effective in explaining wealth effects on individual differences in Oromifa language. Recalling from chapter 5 that the launch model explained 74% of the variance in the intercept in Amharic. It explained 11% of the variance in onset ability in Oromifa and 15% in Tigrinya. Including HAZ as a mediator variable helps to explain an additional 2% of the intercept variance in Amharic. In Oromifa, the mediated launch model explains 21% of the variance in onset ability (i.e. 10% more than the unmediated launch model). In Tigrinya, HAZ does not contribute to explaining the intercept variance.

Table 6.9 Height-for-age mediated launch model estimates, by language group

	Amharic (n=664)		Oromifa (n=228)		Tigrinya (n=344)	
	Estimate	S.E	Estimate	S.E	Estimate	S.E
Intercept (γ_{00})	348.10***	6.22	360.57***	12.40	330.25***	8.68
Slope (γ_{10})	19.09***	4.34	7.76	7.25	1.37	5.90
Intercept on W1 (γ_{01})	217.02***	9.12	94.45***	31.80	116.54***	27.51
Slope on W1 (γ_{11})	-3.80***	1.11	4.98	4.03	-0.67	2.85
Intercept on HAZ (γ_{02})	9.84***	1.86	9.49***	2.42	4.43	2.92
Slope on HAZ (γ_{12})	-0.73***	0.22	-0.65**	0.31	0.06	0.31
HAZ on Wealth (γ_{21})	1.30***	0.16	2.23***	0.58	0.25	0.51
Indirect effect - WI to intercept	12.82***	3.01	21.11***	7.57	1.09	2.29
Indirect effect - WI to slope	-0.95***	0.32	-1.44*	0.79	0.02	0.08
Age 5 (β_1)	53.17***	7.24	49.66***	11.76	69.47***	8.91
Age 8 (β_2)	34.46***	3.68	24.55***	6.16	47.61***	5.15
Age 12 (β_3)	13.53***	3.89	17.81***	6.57	25.02***	5.44
Age 15 (β_4)	5.88	4.00	12.38*	6.75	19.06***	5.53
Level-1 residual variances						
Age 5 (ε_{5i})	2744.32***	225.25	2204.24***	299.04	2750.60***	348.33
Age 8 (ε_{8i})	2773.11***	170.54	1487.94***	163.80	2870.44***	212.64
Age 12 (ε_{12i})	642.14***	58.89	1042.73***	132.55	392.06***	65.20
Age 15 (ε_{15i})	505.74***	79.64	613.21***	166.42	604.70***	119.52
Level-2 variance components						
Intercept variance (σ_0^2)	739.22***	139.32	710.84***	233.47	1188.91***	214.73
Intercept*slope covariance (σ_{10})	-29.00*	16.78	-25.18	24.90	-45.60**	22.85
Slope variance (σ_1^2)	10.21***	2.55	9.14**	3.59	8.19**	3.32
HAZ variance (σ_2^2)	0.89***	0.05	1.35***	0.12	0.99***	0.10
R-square						
ppvt_5	0.561***	0.027	0.354***	0.068	0.423***	0.047
ppvt_8	0.521***	0.024	0.394***	0.051	0.355***	0.035
ppvt_12	0.804***	0.017	0.512***	0.040	0.759***	0.035
ppvt_15	0.843***	0.024	0.700***	0.077	0.687***	0.052
HAZ	0.080***	0.019	0.044*	0.023	0.001	0.003
intercept	0.773***	0.037	0.289***	0.089	0.137**	0.056
slope	0.129**	0.053	0.073	0.071	0.001	0.007
Goodness of fit						
RMSEA†	0.060		0.051		0.090	
CFI‡	0.974		0.952		0.882	
TLI‡	0.963		0.933		0.835	

*Correlation coefficient is statistically significant *p=0.1 **p=0.05 ***p=0.01. †The RMSEA ranges from 0 to 1. A value of 0.06 or less is indicative of good model fit; up to 0.08 is acceptable. ‡The CFI and TLI range from 0 to 1. A value of 0.90 or higher indicates acceptable model fit.

6.3.2 Mothers' education

The results are shown for the full sample in Table 6.10. Controlling for mothers' educational attainment, the expected initial ability is 354 points (γ_{00}= 353.53, $p < 0.01$). The average linear growth rate is 8.66 points per month (γ_{10}= 8.66, $p < 0.01$). For the full sample, wealth at one year of age predicts the initial level of vocabulary abilty at five years (γ_{01} = 163.21, $p < 0.01$), but not the growth rate (nonsignificant γ_{11}). Mothers' education predicts the initial status (γ_{02} = 3.25, $p < 0.01$), but not the growth rate (γ_{12} is nonsignificant).

Mothers' education mediates the effect of wealth on initial ability (c – c' = 40.96, $p < 0.01$). In other words, controlling for mothers' education, the size of the wealth effect on the intercept is reduced by 41 points. The results of the unmediated launch model showed that initial wealth explained 63% of the intercept variance. Including mothers' education as a mediating variable explains 66% of the variance.

Table 6.10 Mothers' education launch model estimates, full sample (n=1,292)

	Estimate	S.E
Intercept (γ_{00})	353.53***	4.42
Slope (γ_{10})	8.66***	3.37
Intercept on W1 (γ_{01})	163.21***	10.49
Slope on W1 (γ_{11})	0.14	1.30
Intercept on MOMED (γ_{02})	3.25***	0.59
Slope on MOMED (γ_{12})	0.009	0.07
MOMED on Wealth (γ_{21})	3.25***	0.59
Indirect effect from WI to intercept (c – c')	40.96***	7.49
Indirect effect from WI to slope (c – c')	0.11	0.87
Age 5 (β_1)	53.39***	5.06
Age 8 (β_2)	37.93***	2.79
Age 12 (β_3)	20.10***	3.02
Age 15 (β_4)	13.57***	3.11
Level-1 residual variances		
Age 5 (ε_{5i})	2835.97***	171.59
Age 8 (ε_{8i})	2931.84***	121.90
Age 12 (ε_{12i})	749.59***	47.27
Age 15 (ε_{15i})	419.45***	65.07
Level-2 variance components		
Intercept variance (σ_0^2)	782.57***	103.90
Intercept*slope covariance (σ_{10})	-34.96***	12.26
Slope variance (σ_1^2)	15.03***	1.99
MOMED variance (σ_2^2)	7.58***	0.35
R-square		
ppvt_5	0.482***	0.023
ppvt_8	0.459***	0.018
ppvt_12	0.780***	0.013
ppvt_15	0.884***	0.017
MOMED	0.425***	0.023
intercept	0.665***	0.035
slope	0.000	0.001
Goodness of fit		
-2LL linear	60489.22	
RMSEA†	0.063	
CFI‡	0.970	
TLI‡	0.958	

*Correlation coefficient is statistically significant *p=0.1 **p=0.05 ***p=0.01. †The RMSEA ranges from 0 to 1. A value of 0.06 or less is indicative of good model fit; up to 0.08 is acceptable. ‡The CFI and TLI range from 0 to 1. A value of 0.90 or higher indicates acceptable model fit.

Table 6.11 shows the results by language group. Mothers' education predicts the initial status in Amharic (γ_{02} = 2.95, $p < 0.01$) and Tigrinya (γ_{02} = 5.66, $p < 0.01$). In Amharic, it also predicts the growth rate (γ_{12} -0.20, $p < 0.05$).

Mothers' education mediates the effects of wealth on the children's initial status in both languages (Amharic: c – c' = 39.46, $p < 0.01$; Tigrinya: c – c' = 37.09, $p <0.01$), and the growth rate in Amharic (c – c' = -2.64). By including mothers' education in the model, the initial wealth effect on the slope becomes nonsignificant in Amharic.

In Oromifa, mothers' education does not mediate the wealth effect. The amount of explained intercept variance increases from 74% to 76% in Amharic, from 11% to 12% in Oromifa and from 15% to 23% in Tigrinya.

Table 6.11 Mothers' education mediated launch model estimates, by language group

	Amharic (n=681)		Oromifa (n=250)		Tigrinya (n=361)	
	Estimate	S.E	Estimate	S.E	Estimate	S.E
Intercept (γ_{00})	345.89***	6.02	357.29***	11.95	343.89***	9.16
Slope (γ_{10})	20.50***	4.28	8.82	6.96	0.96	5.68
Intercept on W1 (γ_{01})	188.47***	12.88	96.42***	31.50	82.19***	26.58
Slope on W1 (γ_{11})	-1.93	1.60	3.33	3.78	1.23	2.84
Intercept on MOMED (γ_{02})	2.95***	0.73	1.54	1.34	5.66***	1.82
Slope on MOMED (γ_{12})	-0.20**	0.09	0.10	0.17	-0.22	0.18
MOMED on Wealth (γ_{21})	13.36***	0.49	5.92***	2.09	6.56***	1.54
Indirect effect - WI to intercept	39.46***	9.87	9.13	8.10	37.09***	14.61
Indirect effect - WI to slope	-2.64**	1.17	0.60	1.03	-1.43	1.31
Age 5 (β_1)	52.71***	6.93	50.39***	11.53	62.64***	8.94
Age 8 (β_2)	33.43***	3.60	24.21***	5.99	45.98***	4.82
Age 12 (β_3)	12.37***	3.85	17.15***	6.35	24.47***	5.15
Age 15 (β_4)	4.73	3.96	11.62*	6.51	18.68***	5.26
Level-1 residual variances						
Age 5 (ε_{5i})	2682.93***	213.17	2222.10***	286.59	2954.83***	361.89
Age 8 (ε_{8i})	2806.40***	171.45	1395.96***	151.29	2939.55***	209.69
Age 12 (ε_{12i})	653.37***	58.06	1031.90***	127.02	398.78***	63.13
Age 15 (ε_{15i})	512.23***	79.78	652.78***	161.08	596.83***	117.73
Level-2 variance components						
Intercept variance (σ_0^2)	742.55***	135.59	796.27***	229.03	1043.62***	198.63
Intercept*slope covariance (σ_{10})	-30.73*	16.20	-29.09	23.85	-33.22	21.82
Slope variance (σ_1^2)	10.34***	2.45	8.93***	3.40	7.44**	3.27
MOMED variance (σ_2^2)	9.28***	0.53	5.35***	0.79	3.23***	0.50
R-square						
ppvt_5	0.562***	0.027	0.344***	0.065	0.383***	0.047
ppvt_8	0.513***	0.023	0.403***	0.049	0.343***	0.033
ppvt_12	0.799***	0.016	0.512***	0.038	0.760***	0.033
ppvt_15	0.840***	0.024	0.685***	0.074	0.698***	0.051
MOMED	0.468***	0.026	0.074*	0.046	0.137***	0.049
intercept	0.768***	0.036	0.158**	0.073	0.208***	0.072
slope	0.113**	0.048	0.027	0.044	0.020	0.035
Goodness of fit						
RMSEA†	0.057		0.049		0.079	
CFI‡	0.978		0.952		0.911	
TLI‡	0.970		0.933		0.876	

*Correlation coefficient is statistically significant *p=0.1 **p=0.05 ***p=0.01. †The RMSEA ranges from 0 to 1. A value of 0.06 or less is indicative of good model fit; up to 0.08 is acceptable. ‡The CFI and TLI range from 0 to 1. A value of 0.90 or higher indicates acceptable model fit.

6.3.3 Preschool

Table 6.12 presents the findings of the mediated launch model with preschool exposure included as a covariate in the model. Controlling for pre-schooling, the expected level of initial ability is 338 points (γ_{00}= 337.99, $p < 0.01$). The average linear growth rate is 10.83 points per month ($p < 0.10$). Pre-schooling predicts the initial vocabulary level at age five (γ_{02} = 41.26, $p < 0.01$) and the growth rate (γ_{12}= 1.27, $p < 0.05$).

On average, children who attended preschool score 41 points higher at onset than children who did not. Moreover, those who attended preschool grow at a faster rate over the duration of the survey, by 1.27 points per period.

The indirect effect of pre-schooling on the intercept is 72 points (i.e. c – c' = 71.59, $p < 0.01$). Thus, the wealth effect on the intercept decreases by 72 points after controlling for preschool exposure. The mediated launch model explains 70% of the intercept variance., in comparison with 63% in the unmediated model.

Table 6.12 Preschool mediated launch model estimates, full sample (n=1,292)

	Estimate	S.E
Intercept (γ_{00})	337.99***	4.79
Slope (γ_{10})	10.83***	3.29
Intercept on W1 (γ_{01})	132.79***	10.80
Slope on W1 (γ_{11})	-1.97	1.38
Intercept on PRESCH (γ_{02})	41.26***	4.67
Slope on PRESCH (γ_{12})	1.27**	0.58
PRESCH on Wealth (γ_{21})	1.74***	0.04
Indirect effect from WI to intercept (c – c')	71.59**	8.16
Indirect effect from WI to slope (c – c')	2.20**	1.02
Age 5 (β_1)	51.66***	5.05
Age 8 (β_2)	35.39***	2.70
Age 12 (β_3)	17.44***	2.93
Age 15 (β_4)	10.87***	3.02
Level-1 residual variances		
Age 5 (ε_{5i})	2858.55***	170.42
Age 8 (ε_{8i})	2932.49***	120.50
Age 12 (ε_{12i})	741.68***	46.24
Age 15 (ε_{15i})	429.65***	63.74
Level-2 variance components		
Intercept variance (σ_0^2)	692.44***	103.86
Intercept*slope covariance (σ_{10})	-38.46***	12.47
Slope variance (σ_1^2)	14.60***	1.97
PRESCH variance (σ_2^2)	0.10***	0.01
R-square		
ppvt_5	0.478***	0.023
ppvt_8	0.457***	0.018
ppvt_12	0.781***	0.013
ppvt_15	0.881***	0.017
PRESCH	0.524***	0.023
intercept	0.703***	0.036
slope	0.011	0.010
Goodness of fit		
-2LL linear	54864.38	
RMSEA†	0.065	
CFI‡	0.970	
TLI‡	0.957	

*Correlation coefficient is statistically significant *p=0.1 **p=0.05 ***p=0.01. †The RMSEA ranges from 0 to 1. A value of 0.06 or less is indicative of good model fit; up to 0.08 is acceptable. ‡The CFI and TLI range from 0 to 1. A value of 0.90 or higher indicates acceptable model fit.

Pre-schooling predicts the intercept across language groups (Table 6.13). The preschool effect on the intercept is most notable in Tigrinya (γ_{02} = 71.47, $p < 0.01$). Preschool mediates the wealth effect on the intercept across language groups to different degrees. In Amharic, preschool exposure mediates the effect of initial wealth on the intercept by 81 points. In Oromifa, it reduces the effect by 37 points. In Tigrinya, the wealth effect is reduced by 13 points.

Taken together, initial wealth and pre-schooling explain 80% of the variance in the children's initial scores in Amharic, 25% of the variance in Oromifa and 23% of the variance in Tigrinya.

Table 6.13 Preschool mediated launch model estimates, by language group

	Amharic (n=681)		Oromifa (n=250)		Tigrinya (n=361)	
	Estimate	S.E	Estimate	S.E	Estimate	S.E
Intercept (γ_{00})	324.97***	6.69	345.25***	11.30	337.03***	8.64
Slope (γ_{10})	22.76***	4.28	9.88	6.84	3.27	5.68
Intercept on W1 (γ_{01})	146.85***	13.74	68.58***	26.84	107.28***	27.67
Slope on W1 (γ_{11})	-2.06	1.71	4.70	3.57	0.50	2.84
Intercept on PRESCH (γ_{02})	44.43***	6.03	49.41***	11.71	71.47***	17.25
Slope on PRESCH (γ_{12})	-1.39**	0.73	-0.95	0.56	-4.13***	1.60
PRESCH on Wealth (γ_{21})	1.83***	0.04	0.75***	0.23	0.18**	0.09
Indirect effect - WI to intercept	81.28***	10.88	37.17***	14.41	13.03*	7.15
Indirect effect - WI to slope	-2.54*	1.33	-0.71	1.27	-0.75	0.56
Age 5 (β_1)	51.54***	6.87	49.27***	11.16	58.09***	8.99
Age 8 (β_2)	31.83***	3.55	23.30***	5.79	43.56***	4.80
Age 12 (β_3)	10.71***	3.81	16.27***	6.18	22.34***	5.13
Age 15 (β_4)	3.06	3.93	10.75*	6.35	16.64***	5.25
Level-1 residual variances						
Age 5 (ε_{5i})	2708.17***	212.48	2158.93***	276.44	2872.72***	360.97
Age 8 (ε_{8i})	2790.88***	168.77	1395.24***	149.53	2962.91***	212.90
Age 12 (ε_{12i})	660.69***	58.02	1058.44***	127.78	412.89***	65.48
Age 15 (ε_{15i})	498.27***	78.12	633.53***	165.38	582.16***	119.27
Level-2 variance components						
Intercept variance (σ_0^2)	627.70***	135.23	670.19***	204.44	1050.81***	202.11
Intercept*slope covariance (σ_{10})	-30.28*	16.58	-28.14	23.66	-34.22	22.01
Slope variance (σ_1^2)	10.65***	2.50	9.40***	3.53	7.82**	3.31
PRESCH variance (σ_2^2)	0.10***	0.01	0.07***	0.01	0.02	0.01
R-square						
ppvt_5	0.559***	0.026	0.354***	0.067	0.383***	0.048
ppvt_8	0.514***	0.023	0.405***	0.048	0.336***	0.033
ppvt_12	0.797***	0.016	0.506***	0.037	0.752***	0.034
ppvt_15	0.844***	0.024	0.693***	0.077	0.703***	0.052
PRESCH	0.613***	0.026	0.094***	0.053	0.018***	0.015
intercept	0.804***	0.037	0.309***	0.100	0.211***	0.067
slope	0.098**	0.044	0.026	0.040	0.044	0.038
Goodness of fit						
RMSEA†	0.059		0.063		0.090	
CFI‡	0.978		0.931		0.884	
TLI‡	0.969		0.903		0.837	

*Correlation coefficient is statistically significant *p=0.1 **p=0.05 ***p=0.01. †The RMSEA ranges from 0 to 1. A value of 0.06 or less is indicative of good model fit; up to 0.08 is acceptable. ‡The CFI and TLI range from 0 to 1. A value of 0.90 or higher indicates acceptable model fit.

6.4 Summary

This section examined how the vocabulary developmental effect of wealth in early childhood is influenced by three key contextual factors: children's nutritional status, the home environment and preschool exposure. Nutritional outcomes were measured by the height-for-age z-score. The quality of the home learning environment is proxied by mothers' educational level. Children's preschool experience is measured by a simple dichotomous item asking parents whether their child ever attended preschool. The overall purpose of this final stage of analysis is to identify interventions to address the wealth gap at different stages of childhood.

The results of the previous chapter pointed to a launch model of developmental change which suggests that to promote children's development and future learning in Ethiopia, early interventions that address initial gaps in vocabulary ability are needed. The launch model was confirmed by: (i) a large amount of variation in the children's initial starting points; (ii) a relatively small, but significant variance in the growth rates; and (iii) the finding that children's initial wealth status explained a large proportion of the variance in onset ability. Changes in wealth over time did not further help to explain individual differences in vocabulary growth. As the data were most efficiently described by the launch model, the focus of the mediation study was to examine the pathways through which living standards influence development in early childhood.

Height-for-age predicted the initial level of vocabulary ability and mediated the wealth effect on the intercept. The HAZ reduces the size of the wealth effect on the intercept by 11.80 points overall. The indirect effect of HAZ on the slope is small but significant. Explanation of the intercept variance is marginally improved by 2% with the addition of HAZ in the model. However, there are differences across language groups. In both Amharic and Oromifa, HAZ predicts the initial status and the growth rate. It further mediates the effects of wealth on the children's initial status as well as the growth rate. In both languages the growth rate becomes less steep once controlling for HAZ in the model. In Tigrinya, HAZ does not mediate the wealth effect, model fit is poor, and the coefficients are not significant. Including HAZ as a mediator variable helps to explain an additional 2% of the intercept variance in Amharic (i.e. increasing from 74% to 76%). In Oromifa, the mediated launch model explains 21% of the variance in onset ability (i.e. 10% more than the unmediated launch model).

Table 6.14 Summary of mediated wealth effects, controlling for HAZ

	Amharic (n=664)		Oromifa (n=228)		Tigrinya (n=344)	
	Estimate	S.E	Estimate	S.E	Estimate	S.E
Total effect of W1 on intercept	229.83***	9.06	115.57***	30.22	117.63***	27.71
Direct effect of W1 on intercept	217.02***	9.12	94.45***	31.80	116.54***	27.51
Indirect effect of W1 on intercept	12.82***	3.01	21.11***	7.57	1.09	2.29
Total effect of W1 on slope	-4.75***	1.11	3.54	3.87	-0.65	2.85
Direct effect of W1 on slope	-3.80***	1.11	4.98	4.03	-0.67	2.85
Indirect effect of W1 on slope	-0.95***	0.32	-1.44*	0.79	0.02	0.08
R-square						
HAZ	0.080***	0.019	0.044*	0.023	0.001	0.003
intercept	0.773***	0.037	0.289***	0.089	0.137**	0.056
slope	0.129**	0.053	0.073	0.071	0.001	0.007

*Correlation coefficient is statistically significant *p=0.1 **p=0.05 ***p=0.01. †The RMSEA ranges from 0 to 1.

The findings of the analysis showed that mothers' educational attainment mediated the wealth effect on the children's initial vocabulary ability. The results obtained from estimating the full sample model showed that the mediated effect wealth on the intercept was large and significant, 40.96 points overall.   Breaking the analysis down by language group revealed that while mothers' educational attainment is effective in mediating the wealth effect on the intercept in Amharic and Tigrinya, it does not help to explain the wealth gap in Oromifa. The indirect effect of initial wealth on the intercept is 39.46 points in Amharic and 37.09 points in Tigrinya. Moreover, by including mothers' education in the model, the initial wealth effect on the slope becomes nonsignificant in Amharic. The amount of explained intercept variance increases from 74% to 76% in Amharic, from 11% to 12% in Oromifa and from 15% to 23% in Tigrinya.

Table 6.15 Summary of mediated wealth effects, controlling for mothers' education

	Amharic (n=681)		Oromifa (n=250)		Tigrinya (n=361)	
	Estimate	S.E	Estimate	S.E	Estimate	S.E
Total effect of W1 on intercept	227.93***	8.85	105.56***	28.49	119.27	27.67
Direct effect of W1 on intercept	188.47***	12.88	96.42***	31.50	82.19***	26.58
Indirect effect of W1 on intercept	39.46***	9.87	9.13	8.10	37.09**	14.61
Total effect of W1 on slope	-4.57***	1.09	3.93	3.66	-0.21	2.84
Direct effect of W1 on slope	-1.93	1.60	3.33	3.78	1.23	2.84
Indirect effect - WI to slope	-2.64**	1.17	0.60	1.03	-1.43	1.31
R-square						
MOMED	0.468***	0.026	0.074*	0.046	0.137***	0.049
intercept	0.768***	0.036	0.158**	0.073	0.208***	0.072
slope	0.113**	0.048	0.027	0.044	0.020	0.035

*Correlation coefficient is statistically significant *p=0.1 **p=0.05 ***p=0.01. †The RMSEA ranges from 0 to 1.

Preschool exposure was found to be particularly important for explaining the wealth effect on vocabulary outcomes. Pre-schooling significantly predicts the initial onset ability at age five and the growth rate across language groups. On average, children who attended preschool score 41 points higher at onset than children who did not. Moreover, those who attended preschool grow at a faster rate over the duration of the survey, by 1.27 points per period. Controlling for preschool participation, the indirect effect of wealth on the intercept is 72 points overall. The mediated launch model explains 70% of the intercept variance, in comparison with 63% in the unmediated model. Preschool also substantially reduces the wealth effect on the intercept across language groups. In Amharic, preschool exposure mediates the effect of initial wealth on the intercept by 81 points. In Oromifa, the indirect effect is 37 points. In Tigrinya, the wealth effect is reduced by 13 points. Taken together, initial wealth and pre-schooling explain 80% of the variance in the children's initial scores in Amharic, 25% of the variance in Oromifa and 23% of the variance in Tigrinya.

Table 6.16 Summary of mediated wealth effects, controlling for preschooling

	Amharic (n=681)		Oromifa (n=250)		Tigrinya (n=361)	
	Estimate	S.E	Estimate	S.E	Estimate	S.E
Total effect of W1 on intercept	228.14***	8.88	105.76***	28.78	120.31***	27.77
Direct effect of W1 on intercept	146.86***	13.74	68.58***	26.84	107.28***	27.67
Indirect effect of W1 on intercept	81.28***	10.88	37.17***	14.41	13.03*	7.15
Total effect of W1 on slope	-4.60***	1.09	3.99	3.69	-0.25	2.85
Direct effect of W1 on slope	-2.06	1.71	4.70	3.57	0.50	2.84
Indirect effect - WI to slope	-2.54*	1.33	-0.71	1.27	-0.75	0.56
R-square						
PRESCH	0.613***	0.026	0.094***	0.053	0.018***	0.015
intercept	0.804***	0.037	0.309***	0.100	0.211***	0.067
slope	0.098**	0.044	0.026	0.040	0.044	0.038

*Correlation coefficient is statistically significant *p=0.1 **p=0.05 ***p=0.01. †The RMSEA ranges from 0 to 1.

Chapter 7 Discussion and conclusions

This chapter discusses the findings of the research in relation to each of the three sets of research questions. The results are interpreted with reference to the literature review. It discusses the lessons emerging from the findings of the study as well as the limitations of the research. Future research directions are also identified.

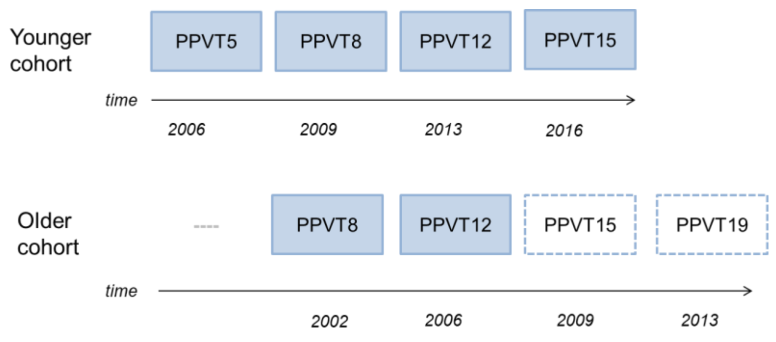
7.1 Summary of main findings

The overall aim of the D.Phil study is to better understand how poverty, experienced at different stages of childhood, shapes children's vocabulary developmental paths and the different pathways through which children's development and learning can be promoted in contexts of poverty and adversity. This first section of the discussion chapter presents the main findings of the study in the context of the broader academic literature.

7.1.1 Measuring children's vocabulary development

The first set of research questions address the effectiveness of the PPVT instrument that was used to measure vocabulary ability among the multicultural/lingual sample of children over the survey period. Initially, the psychometric analysis was meant to examine the performance of the measure when used to assess two cohorts of children, a younger cohort and an older cohort, with a view to construct a comparable metric. The original research design adopted an accelerated longitudinal approach using data from both cohorts (Figure 7.1). However, the PPVT was dropped from the older cohort survey due to the large ceiling effect that was observed in Round 3 (section 3.3.2 and Appendix 2). For the current research, the accelerated longitudinal approach was subsequently abandoned to focus on the younger cohort survey data. The decision to focus on the younger cohort of children is described in more detail in Chapter 3 (section 3.3.3).

Figure 7.1 Initial two cohort accelerated longitudinal research design



It is important to note that from the full scale of 204 items, only a subset of 55 items that showed good psychometric properties in Rounds 2 and 3 were retained for the younger cohort surveys (one-quarter of the items). Eliminated from the scale were items that did not demonstrate "cognitive equivalence (or level of difficulty)" with the original tests or between languages (Leon and Singh, 2017: 6). The PPVT-III Form III-A, developed for English speakers, was translated into the main Ethiopian languages. The rules to set the basal and ceiling items were also dropped (Leon and Singh, 2017). Normally, administration of the PPVT is such that once a set of items is successfully completed, the next set is offered until the child reaches a ceiling. The starting set for a child is found when a child makes no more than one error in a set of 12 items (Dunn et al., 1997: 11). If the child receives more than one error in the starting set then the examiner is required to drop back to the previous set beginning with the first item in that set (Young Lives, 2009: 40). This procedure is continued until the start set is established. The ceiling set is found when eight or more errors are made in a set of 12 items (Dunn et al., 1997: 11).

The significant changes to PPVT instrument and the criteria for marking children's performance meant that understanding the validity and reliability of the scale became a much more prominent research priority than what was originally envisaged. As pointed out by Borsboom (2005), "… the problem of validity cannot be solved by psychometric techniques" (Borsboom, 2005: 151). Measurement and modelling decisions should be based on substantive theories about assessment, development and/or learning

(Hopfenbeck, et al., 2017; Baird et al., 2017; Rutowski and Rutowski, 2016; McGrane, 2015a; McGrane 2015b). Thus, preparing the groundwork for this component of the analysis involved engaging with measurement theory and identifying the issues debated in the literature that pertain to cross-cultural assessment, particularly in low-income countries. Due to the absence of qualitative information to support the construct validity of the Young Lives PPVT scale, i.e. the picture stimuli used during testing or the translated words, the reliability and validity analysis relied mainly on statistical evidence, drawing on insights from the educational and psychological measurement literature.

The first research question (RQ1.1) examines the extent to which the PPVT-III a valid and reliable measure of vocabulary ability among the sample of children. The psychometric properties of the PPVT data were assessed according to the Rasch measurement model. Due to its emphasis on item functioning and the fundamental properties of good measures, the more restrictive Rasch framework was considered better suited for identifying anomalies in the data and establishing the validity of the scale compared with alternative paradigms, such as 2- or 3-parameter item response models. However, it is recognised that different scaling procedures, i.e. Rasch, 1PL, 2PL, 3PL, will produce different person estimates and will affect the measure of change over time (Goldstein, 2004; Brown et al., 2007; Bond and Lang, 2013; Kreiner and Christensen, 2014; Gebhardt and Adams, 2007).

Table 7.1 provides a summary of how the findings of the Rasch analysis by language group. In terms of overall model fit, the Rasch analysis of the test score data showed good psychometric properties and evidence of progress over the period. The measure conforms with theoretical expectations and correlates well with other measures of achievement available in the data structure. However, anomalies in the data were identified, with implications for growth modelling. For all language groups, person separation is highest at age eight and then declines suggesting that the spread of persons relative to items is most consistent with the intention of the instrument at age eight.

Person ability estimates are underestimated by the range of items available in the scale in later rounds of the survey. In Oromifa and Tigrinya, the person residual mean and standard deviation are far from standard normal in the first round of testing, when the children are five years old suggesting that the test worked less well for this age group in the two languages. The findings also showed evidence of a ceiling effect at age 12, which is worsened by age 15 for all subgroups, resulting in an underestimation of individual differences in growth/endpoints. The removal of three-quarters of the items from the original test has likely exacerbated the ceiling effect, in turn affecting the validity of the construct. Scholars have debated whether statistical misfit should be accounted for through statistical modelling or understood as substantive anomalies (Andrich, 2004; Traub, 1983; Goldstein, 1979).   An alternative approach would have been to try to understand the statistical misfit in the earliest rounds of the survey and to “generate new data that better conform to the model while also enhancing the substantive validity of the variable” (Andrich, 2004: 15).

Table 7.1 Summary of Rasch analysis findings, by language group

	Round	Amharic (n=692)	Oromifa (n=260)	Tigrinya (n=369)
Overall fit and targeting				
person separation	2	0.94	0.89	0.89
	3	0.95	0.90	0.94
	4	0.89	0.85	0.85
	5	0.83	0.85	0.85
item residual mean (SD)	2	-1.09 (2.91)	-0.23 (0.41)	-0.23 (0.42)
	3	-0.69 (1.99)	-0.32 (1.56)	-0.57 (1.41)
	4	-0.63 (2.91)	-0.14 (1.10)	-0.16 (1.60)
	5	-0.67 (2.63)	-0.16 (1.22)	-0.29 (1.52)
person residual mean (SD)	2	-0.25 (0.52)	-1.27 (4.94)	-1.65 (5.59)
	3	-0.51 (0.92)	-0.42 (0.87)	-0.51 (0.86)
	4	-0.25 (0.84)	-0.21 (0.84)	-0.21 (0.93)
	5	-0.26 (0.71)	-0.22 (0.76)	-0.20 (0.66)
person location mean (SD)	2	-2.32 (1.79)	-2.54 (1.40)	-2.65 (1.57)
	3	1.06 (1.93)	-0.78 (1.15)	0.36 (1.70)
	4	1.86 (1.31)	1.52 (1.04)	1.26 (0.99)
	5	2.38 (1.35)	1.76 (1.12)	1.69 (1.05)
Item and person fit				
number of extreme persons	2	0	0	1
	3	3	0	0
	4	2	2	1
	5	0	0	0
number of misfitting items	2	10	0	0
	3	3	0	0
	4	9	0	0
	5	9	0	0
Item stability				
items showing DIF for time	N/A	80%	71%	86%

The results also suggest the possibility of a multidimensional scale. Some of the harder items on the test may be measuring some other factor that is correlated with the PPVT, in turn giving higher ability children some advantage. Vocabulary includes several different dimensions of knowledge (Kieffer and Stahl, 2015), which are difficult to assess with picture-based measures (Kieffer and Lesaux, 2010). One possible factor may be that children living in urban areas benefit from some experience or exposure that gives them an advantage over rural children. Subsequent analyses of urban-rural differences in vocabulary outcomes support this hypothesis, namely the cumulative disadvantage faced by children living in rural areas (Figure 5.3). The difference is most pronounced among the Amharic subsample, which includes children living in the capital city, Addis Ababa. However, this would need to be examined in future research. It has been pointed out that deciding how to handle multidimensionality is “a relative issue and depends on the purpose of the scale or assessment” (Marais and Andrich, 2008a: 2012). Unidimensionality is regarded as a continuum, and a researcher must determine at which point multidimensionality threatens the interpretation of results given the aims of the research (Smith, 2002). For the present study, multidimensionality is acknowledged as a potential measurement disturbance that requires further consideration in future studies.

The second and third research questions (RQ1.2 and RQ1.3) investigate how consistently the items on the PPVT functioned across the multilingual sample of children and across time points, and to what extent growth trajectories can be reliably estimated, considering the linguistic differences. Thus, this aspect of the investigation questions the assumptions underlying subgroup comparisons. The analysis of Differential Item Functioning (DIF) showed a large and significant DIF effect for language. Therefore, the scale should not be considered unbiased across groups. Given the extent of the DIF effect, separate language group analyses were deemed the most appropriate way of interpreting the data.

The influence of language and other cultural factors on assessment outcomes is widely acknowledged in the literature (Berry, 1969; Poortinga, 1989; Hambleton and Patsula, 1998; Van de Vijver and Poortinga, 1998). However, in low-income countries, test localisation processes, which are essential for ensuring that assessment instruments “suit the context in which they are administered” are often sidestepped or under-resourced, and rarely documented (Nag et al., 2017: v). The Young Lives study has been cited as an excellent example of how fairness in testing can be ensured at each stage of the assessment cycle (Nag et al., 2017). Efforts to adapt the original instruments are evident in the establishment of a national panel of experts to guide the selection of items; the use of translators who were fluent in both English and the target language; field-testing and item revisions; training of assessors and equipping assessors with manuals and ongoing supervisory support (Cueto and Leon, 2012; Cueto et al., 2009). However, score interpretation is challenged at a theoretical level by uncertainties about the underlying concept under study (i.e. vocabulary ability vs. cognitive ability), and how the construct is aligned to both national and international frameworks of children's development. At technical level, results interpretation is further challenged by issues relating to how the construct has been measured, i.e. the removal of a large proportion of test items in later rounds of the survey, changes to translation and scoring procedures, and how the instrument has been adapted, i.e. the absence of the PPVT picture stimuli, the translated words or documentation of the meaning underlying the items in each language (as done in Haitana et al., 2010);

Apart from cultural and linguistic sources of error, item functioning may not remain constant over time even when identical data collection protocols are used on the same sample of respondents (Wright, 1996). Thus, the analysis was extended to investigate item stability over time to determine whether the scores from the repeated test can be reliably calibrated onto a single metric. A DIF analysis for time showed evidence of substantial person-time dependencies, suggesting that pooling the data would not be the most appropriate equating procedure. Moreover, cross-plotting item difficulty estimates at each time point showed that the test performed most consistently over time in Amharic, and least consistently in Tigrinya. In Tigrinya, the extent of item instability may suggest changes in the meaning underlying the items over time or possible errors in the testing process, such as changes to administration procedures. It has been argued that the measurement of change based on repeated measures data is sensitive to the choice of linking items and test equating procedures (Gebhardt and Adams, 2007). Consequently, in each of the three languages, item difficulty estimates were anchored on the most stable item estimates to provide the most accurate description of person change (Wright, 1996).

7.1.2 Wealth effects on children's vocabulary development

The second set of research questions are about understanding the extent to which household living standards, as measured by household wealth, influence developmental trajectories. Despite the abundance of research about the poverty effects on cognitive outcomes in early childhood, little is known about how disparities in development evolve over the course of childhood or how the relationship between poverty, children's development and learning changes over time. Moreover, empirical research in low-income countries does not explicitly draw on theories of child development and related conceptual models for testing hypothesis.

The first research question (RQ2.1) examines whether individual developmental trajectories differ in relation to the overall mean trend, in terms of initial status and growth. Table 7.2 provides a summary of how the results of the growth trajectory analysis differed between language groups. The results of the base model, in which only time is included as a predictor variable, shows good data-model fit in Amharic and Oromifa. The results confirmed large and significant between-child variability in onset ability at age five. The between-child variance in the growth rate was small but significant. The small amount of slope variance reinforces the premise that environmental and genetic factors in early life are most predictive of individual differences in developmental trajectories (Nelson, 2000; Guo, 1998). However, it has been argued that vocabulary ability (i.e. the PPVT test scores) is “every bit as acquired an ability as reading and mathematics” (Farkas and Beron, 2004: 467), and that the process of vocabulary development continues throughout childhood (Nagy and Herman, 2005; Stahl and Fairbanks, 1986; Beck et al., 1982; Stahl and Nagy, 2007; Beck and McKeown, 2007; Kieffer and Stahl, 2015). Thus, it is important to recognise that the results may also reflect the observed ceiling effect in the test score data which implies that individual growth trajectories are underestimated in the model. The results of the model showed that the growth rate was only significant in

Amharic. A nonsignificant growth rate is an important finding as it illustrates the differences in the way the test performed across linguistic groups.

Table 7.2 Summary of growth modelling results, by language group

	Amharic		Oromifa		Tigrinya	
	Estimate	S.E	Estimate	S.E	Estimate	S.E
Base model						
Intercept	336.47***	7.01	333.03***	9.76	315.20***	7.26
Slope	19.71***	5.65	10.02*	7.02	-1.30	6.14
Intercept variance	3126.90***	233.22	901.46***	248.11	1354.78***	243.27
Slope variance	9.42***	2.77	7.30**	3.39	7.17**	3.38
RMSEA†	0.057		0.054		0.104	
CFI‡	0.975		0.952		0.879	
TLI‡	0.967		0.938		0.843	
Launch model						
Intercept	349.24***	6.30	356.54***	9.73	336.43***	7.38
Slope	20.16***	4.28	8.56	6.89	1.12	5.86
Intercept on W1	228.29***	8.87	105.65***	28.51	119.17***	27.64
Slope on W1	-4.62***	1.09	3.83	3.65	-0.21	2.84
Intercept variance	801.27***	142.59	802.03***	226.96	1150.43***	211.04
Slope variance	10.22***	2.52	8.58***	3.44	7.77**	3.34
RMSEA†	0.061		0.055		0.091	
CFI‡	0.977		0.951		0.902	
TLI‡	0.969		0.933		0.866	
Change-to-change model						
Intercept	338.36***	6.23	335.23***	9.92	331.60***	8.29
Slope	19.86***	4.11	9.55	6.55	0.67	5.67
Intercept on W1	163.45***	16.53	96.82***	33.06	53.06*	27.64
Slope on W1	-1.48	1.95	0.863	4.18	-2.07	3.49
Wealth 5 fixed effect	55.15***	18.15	22.02	26.06	104.87***	27.79
Wealth 8 fixed effect	107.48***	14.22	21.22	22.88	27.08	21.07
Wealth 12 fixed effect	58.14***	9.64	46.43**	21.55	57.32***	11.25
Wealth 15 fixed effect	57.90***	10.79	94.53***	22.77	62.14***	13.52
Intercept variance	717.47***	141.98	852.08***	226.96	1132.23***	197.88
Slope variance	10.34***	2.49	8.93***	3.44	9.72**	3.28
RMSEA†	0.065		0.080		0.091	
CFI‡	0.960		0.854		0.857	
TLI‡	0.946		0.802		0.807	

*Correlation coefficient is statistically significant *p=0.1 **p=0.05 ***p=0.01. †The RMSEA ranges from 0 to 1. A value of 0.06 or less is indicative of good model fit; up to 0.08 is acceptable. ‡The CFI and TLI range from 0 to 1. A value of 0.90 or higher indicates acceptable model fit.

The mean growth trajectories were plotted by wealth quintile to examine the magnitude and significance of the gap between the two quintiles to observe whether the developmental disadvantage due to poverty in early childhood is cumulative, as learning is acquired and deficits tend to widen with age (Jensen, 1966 and 1974a; Merton, 1968; DiPrete and Eirich, 2006). It may be the case that the gap narrows with time (Hall et al., 2009; Barnett and Hustedt, 2005; Puma et al., 2010). Alternatively, the gap may remain constant throughout childhood as expected for more stable cognitive outcomes (Guo, 1998; Farkas and Beron, 2004). The results in the literature from low-income countries have been mixed. A recent multi-country study using the Young Lives data showed that wealth disparities in cognitive outcomes remained constant in India, widened in Ethiopia, and narrowed in Vietnam and Peru (Reynolds et al., 2017). A widening developmental gap was also found in Bangladesh (Hamadani et al., 2014), Madagascar (Galasso et al., 2017), and Colombia (Rubio-Codina and Grantham-McGregor, 2019). A constant gap up to the age of 12-13 was found in Ecuador (Schady et al., 2015). The results of the current study showed that children in the top quintile were assessed higher on average than children in the bottom. In Amharic, the average gap between top and bottom quintiles narrowed over the course of childhood, by 0.47 SD, and remained constant in Tigrinya and Oromifa. Discrepancies across and within studies may be due real differences in children's outcomes or limitations regarding the measurements of child development (Rubio-Codina and Grantham-McGregor, 2017: 2).

The second research question (RQ2.2) explores wealth status in early childhood as a predictor of initial ability and the growth rate and examines whether patterns of development differ between the top and bottom wealth quintiles. This "launch" model of developmental change is based on assumptions about the importance of early experiences for later development and the stability of contexts over time (Skinner et al., 1998). In other words, the model assumes that the initial level of an antecedent variable (i.e. wealth) launches the developmental trajectory (i.e. PPVT scores). Consistent with the findings of the base model, the results of the launch model showed that the Tigrinya scores did not fit the model well, based on the goodness of fit indices. The findings showed that wealth, measured when the children were one year old, had a large and significant effect on onset vocabulary scores. The results are consistent with existing evidence from low-income countries (section 2.3.1, table 2.1). Wealth predicted the growth rate in Amharic only. In Amharic, for every unit increase in the wealth index score, the growth rate decreased by 4.62 points ($p < 0.01$). Initial wealth also explained 54% of the covariance between initial status and growth in Amharic. Therefore, wealth status in early childhood helps to explain the convergence in the growth trajectories between top and bottom wealth quintiles. The launch model explains 74% of the variance in onset ability in Amharic, 11% in Oromifa and 15% in Tigrinya. Wealth as a “launcher” variable does not explain individual differences in the growth rate.

The third research question (RQ2.3) examines the extent to which changes in wealth over time are associated with patterns of vocabulary development. Individual differences in progress are modelled as a function of both initial wealth and wealth at each subsequent time point to examine the effects on the growth factors. The ‘Change-to-Change’ model of development assumes that changes in a target trajectory are influenced, not only by the initial level of a predictor variable but by the pattern of change in that variable, based on the reasoning that a nurturing early environment is vital but it “may not be sufficient to support a positive trajectory” (Skinner et al., 1998: 31). Moreover, persistent or long duration poverty has been shown to have a more detrimental effect than short term poverty (Dickerson and Popli, 2012; Brooks-Gunn and Duncan, 1997; Duncan et al., 1994; Korenman et al., 1995; Smith et al., 1997). This model was applied to attempt to explain the small but significant variance in the children’s growth rates. However, including wealth as a time-varying covariate reduces data-model fit across language groups. While the fixed effect of wealth is significant, changes in wealth over time do not explain individual differences in the growth factors.

Taken together, the findings suggest that the children's vocabulary trajectories are better described by the launch model of developmental change, suggesting that the earliest years of life are most important, and a child's standard of living in the earliest years of life is an important predictor of later cognitive outcomes. However, the small variation in the growth rate may also be due, in part, to statistical measurement error resulting in part from the large ceiling effect observed at age 12 and 15, low variance in the slope growth factor, and an underestimation of the growth trajectories, limiting the possibility of examining vocabulary developmental patterns over time.

7.1.3 Mediated wealth effects on children's vocabulary development

The final sub-study examined the mediated effects of wealth on the children's vocabulary developmental trajectories to identify the mechanisms through which wealth influences development. Specifically, the research question (RQ3.1) investigates how the relationship between wealth and vocabulary development is explained by the children's nutritional status, their home environment and preschool exposure? The findings of the growth models were consistent with a launch model of developmental change. Therefore, the mediation study did not seek to examine and identify pathways at later stages of childhood, such as self-concept, motivation and opportunity, which "tend to be affected by an impoverished environment from early adolescence on" (Guo, 1998: 265). It is argued that in earlier childhood, children are generally eager to learn and motivation is not yet a limiting factor (Entwisle and Alexander, 1989). Such factors become increasingly important predictors of learning outcomes over time as children grow more aware of their socioeconomic contexts (Henderson and Dweck, 1990; Zimmerman, 1990; Schunk et al., 2014; Breen and Goldthorpe, 1997; Guo, 1998; Ogbu, 1986; Mickelson, 1989, 1990). In early childhood, interactions in the child's immediate environment, such as biological characteristics and family experiences are prioritised (Mayes and Lewis, 2013: 19). For vocabulary development, proximal influences include targeted vocabulary instruction, opportunities for independent reading, communication with parents and peers, classroom lectures and discussion, reading, and television (Nagy and Hernan, 1987; Stahl and Fairbanks, 1986; Beck et al., 1982; Stahl and Nagy, 2007; Beck and McKeown, 2007; Kieffer and Stahl, 2015). Learning words from oral contexts is particularly important in early childhood, i.e. through conversation or from listening to books read aloud (Leung, 1992; McKeown, 1985; Werner and Kaplan, 1950). Thus, the factors known to influence cognitive and vocabulary processes in the early years and vary with wealth were modelled as mediator variables. Nutritional outcomes were measured by the height-for-age z-score (HAZ). The quality of the home learning environment was proxied by a measure of mothers' educational level. Children's preschool experience was measured by a simple dichotomous item asking parents whether their child ever attended preschool.

Table 7.3 provides a summary of how the mediated growth trajectory analysis differed between language groups. Nutrition has been cited as one of the most proximal risk factors encountered in the first five years of life, specifically: stunting, inadequate cognitive stimulation, iodine deficiency and iron deficiency anaemia (Walker et al., 2007), with long term implications for children's development and learning (Grantham-McGregor et al., 2007). The results of the current mediation analysis showed that height-for-age predicted the initial level of vocabulary ability and partially mediated the wealth effect on onset ability scores and the growth rate in both Amharic and Oromifa, consistent with findings from other low-income countries (Fernald et al., 2012; Reynolds, 2017). In Tigrinya, the indirect effect of HAZ was nonsignificant. Overall, the results showed that HAZ explained an additional 0-10% of the variance in onset ability scores.

Table 7.3 Summary of mediated growth modelling results, by language group

	Amharic		Oromifa		Tigrinya	
	Estimate	S.E	Estimate	S.E	Estimate	S.E
Nutrition/HAZ						
Total effect of W1 on intercept	229.83***	9.06	115.57***	30.22	117.63***	27.71
Direct effect of W1 on intercept	217.02***	9.12	94.45***	31.80	116.54***	27.51
Indirect effect of W1 on intercept	12.82***	3.01	21.11***	7.57	1.09	2.29
Total effect of W1 on slope	-4.75***	1.11	3.54	3.87	-0.65	2.85
Direct effect of W1 on slope	-3.80***	1.11	4.98	4.03	-0.67	2.85
Indirect effect of W1 on slope	-0.95***	0.32	-1.44*	0.79	0.02	0.08
R-square						
HAZ	0.080***	0.019	0.044*	0.023	0.001	0.003
intercept	0.773***	0.037	0.289***	0.089	0.137**	0.056
slope	0.129**	0.053	0.073	0.071	0.001	0.007
Mother's education						
Total effect of W1 on intercept	227.93***	8.85	105.56***	28.49	119.27	27.67
Direct effect of W1 on intercept	188.47***	12.88	96.42***	31.50	82.19***	26.58
Indirect effect of W1 on intercept	39.46***	9.87	9.13	8.10	37.09**	14.61
Total effect of W1 on slope	-4.57***	1.09	3.93	3.66	-0.21	2.84
Direct effect of W1 on slope	-1.93	1.60	3.33	3.78	1.23	2.84
Indirect effect - WI to slope	-2.64**	1.17	0.60	1.03	-1.43	1.31
R-square						
MOMED	0.468***	0.026	0.074*	0.046	0.137***	0.049
intercept	0.768***	0.036	0.158**	0.073	0.208***	0.072
slope	0.113**	0.048	0.027	0.044	0.020	0.035
Preschool						
Total effect of W1 on intercept	228.14***	8.88	105.76***	28.78	120.31***	27.77
Direct effect of W1 on intercept	146.86***	13.74	68.58***	26.84	107.28***	27.67
Indirect effect of W1 on intercept	81.28***	10.88	37.17***	14.41	13.03*	7.15
Total effect of W1 on slope	-4.60***	1.09	3.99	3.69	-0.25	2.85
Direct effect of W1 on slope	-2.06	1.71	4.70	3.57	0.50	2.84
Indirect effect - WI to slope	-2.54*	1.33	-0.71	1.27	-0.75	0.56
R-square						
PRESCH	0.613***	0.026	0.094***	0.053	0.018***	0.015
intercept	0.804***	0.037	0.309***	0.100	0.211***	0.067
slope	0.098**	0.044	0.026	0.040	0.044	0.038

*Correlation coefficient is statistically significant *p=0.1 **p=0.05 ***p=0.01.

Home environmental factors that have been shown to mediate the effect of socioeconomic factors on cognitive development include the availability of stimulating materials, conversations and activities with adults, and parental warmth and affection (Bradley and Corwyn, 1984). In the current study, mothers' education is used to proxy the quality of the home learning environment, due to the absence of more direct indicators. The findings showed that mothers' educational attainment mediated the wealth effect on the children's initial ability in Amharic and Tigrinya but did not help to explain the wealth gap in Oromifa. The mediated effect of wealth on the intercept was large and significant overall, consistent with the results from other mediation studies in low-income countries (Hamadani et al., 2014; Galasso et al., 2017; Rubio-Codina and Grantham-McGregor, 2019). Overall, the results showed that 1-8% the of the variance in onset ability was explained by mothers' education.

High-quality pre-schooling has been shown to promote improved cognitive development among children living in families with low SES (Sammons et al., 2014; Sylva et al., 2014; Hall et al., 2009). In the current study, preschool exposure was found to be a particularly important mediator variable. In all languages, pre-schooling significantly predicts the initial level of vocabulary ability and the growth rate, reinforcing findings from other relevant studies in low-income countries (Lopez-Boo, 2016). On average, children who attended preschool score 41 points higher at onset than children who did not. Moreover, those who attended preschool grow at a faster rate over the duration of the survey, by 1.27 points per period. The results showed that pre-schooling explained 6-14% of the wealth effect.

7.2 Emerging insights and lessons learned

This section discusses the issues emerging from the findings of the D.Phil study, reinforcing debates within the academic literature. The recommendations put forward in this section serve as a starting point for what could become a set of tools and resources for institutions researching children's development and learning in low-income developing countries, particularly in multicultural contexts. The list is not exhaustive. It highlights the prominent measurement issues discussed in the literature that are corroborated by the results of the present study.

1. Deciding why, what and how to measure children's cognitive outcomes based on substantive educational and psychological theory. The first step of an assessment cycle should focus on clearly defining the purpose of measurement and the underlying concepts that will be measured. In recent decades there has been an expansion of large-scale measurement of children's development and learning in low-income countries. Scholars argue that large-scale assessments add value when they can inform improvements to the quality of educational systems (Tesfay, 2017, based on Nag et al., 2017). However, the aims of assessment, the broad concepts of inquiry and the specific constructs being measured are not always explicitly stated. Often, the starting point is deciding which instrument to use. Viswanathan (2010) refers to construct definition and domain delineation as the first intangible of measurement and the starting point for the measure development process (Viswanathan, 2010). Construct definition and domain delineation involves understanding what is being measured by clearly defining the concepts of inquiry before designing the measure and generating the items. This preliminary step is crucial as it "enables understanding the conceptual as it relates to the operational" (Viswanathan, 2010).

2. Test localisation as a process of understanding the influence of cultural factors, such as language, on testing outcomes. Inadequate attention is given to cultural sources of influence on instrument adaptation and score interpretation. In many low-income countries, national frameworks for children's development and learning do not exist or are in early phases of development. Nationally owned instruments and technical expertise to administer these tools may also be lacking. Transferring instruments from external contexts may be necessary to take stock of the situation and strengthen policies for children. Developing meaningful measures to inform educational policy decisions remains a challenge in low-income countries (Nag et al., 2017). Tests that have demonstrated good psychometric properties in Western contexts are increasingly being adapted for use in low-income countries. However, transferring instruments from one context to another requires rigorous piloting, adaptation, and reliability and validity testing to ensure the instruments are culturally and contextually meaningful, adhere to local testing conventions, well-targeted, and feasible for testing given local technical and human resource capacities. Van de Vijver and Leung (1997) distinguished between three types of bias in cross-cultural social research: construct, method, and item bias, which were discussed in detail in the literature review. Construct bias occurs when the definition of the construct is not the same across cultures. Method bias can take three forms: (i) sample bias, i.e. incomparability of samples; (ii) instrument bias, i.e. lack of familiarity with the item format or testing process; (iii) administration bias, i.e. communication problems between assessors and respondents. Item bias occurs when individuals with the same level of ability perform differently on an item. Item bias may be due to differences in the way items are translated across cultures or the inclusion of irrelevant items on a test. The process of developing meaningful measures that minimise the effect of bias is iterative, and involves consultations with a range of experts who have an understanding of what children's development and learning means locally, in terms of what children are expected to know and do, as well as an understanding of the practical constraints within the context.

3. Transparency and reporting of test validation processes. In low-income developing countries, commercial tests adaptation processes are less well documented than small scale researcher developed tools (Nag et al., 2017). Most large-scale assessments emphasise translation efforts with little description of test design, adaptation, and piloting processes, or the "intangibles of measurement" which refers to the processes by which the data are obtained (Viswanathan, 2005). Even the more tangible aspects of measurement, the psychometric properties of instruments, are rarely reported as a way of validating them: i.e. how well the test targets children's ability levels, any floor or ceiling effects that might affect conclusions drawn from data. Such findings would normally be addressed at the piloting stage, where decisions are made regarding whether to modify the test items or the assumptions underlying the test. Psychometric analysis further allows researchers to observe differences in the way subgroups respond to the items on the test and to then make modifications to reduce biases. In developing world educational research, the reader is often left to assume that the measure is working as intended without any statistical or qualitative information to support the judgement.

4. *Identifying mitigating pathways requires information about the relationships and experiences through which cognitive outcomes are shaped*. Developmental theory emphasises multiple layers of context that are important for children's development and learning: biological factors, the family, peers, the school, and the community. Understanding what shapes a specific cognitive outcome requires knowledge of the developmental process and the contexts in which development takes place. For vocabulary development, important influences relate to oral communication with parents and caregivers, particularly in the younger years. For older children, experiences in the classroom, reading and targeted instruction become increasingly important factors. The findings of the current study reinforced existing research about the importance of children's nutrition, mothers' education, and preschool participation in mitigating the effects of poverty on children's cognitive development. However, inorder to expand this knowledge base beyond, research on children's development in low-income countries will need to increasingly draw on substantive theories and local understandings of child development.

7.3 Limitations of the research

During the research, three key limitations were identified which had implications for the analysis. First, the change-to-change model of development could not be explored further due to the small slope variance. Thus, it was not possible to examine how poverty launches children's developmental trajectories and further exerts its impact in later childhood. The implication is that the research was not able to contribute to identifying mediating mechanisms in middle childhood and adolescence. Rather, the magnitude and significance of the variance in onset ability implied that the focus of the study should be on identifying early interventions that address initial gaps in vocabulary ability in Ethiopia.

The absence of slope variance may reflect developmental theory, that ability develops most rapidly in the early years, or it may be an artefact of anomalies in the data, i.e. the ceiling effect. It has been argued that vocabulary ability is "every bit as acquired an ability as reading and mathematics" (Farkas and Beron, 2004: 467). The observed ceiling effect limits the possibility to challenge/test these assumptions.

Related to this issue, a second limitation is that persons with extreme fit residuals outside of the acceptable range of ±2.5 were not removed from the data set prior to the Rasch analysis. Extreme persons are not directly estimable by the Rasch model and considered as outliers. Addressing person misfit improves data-model fit and the reliability of the scale. The findings showed no major person misfit overall, as evidenced by a pooled person fit residual that is close to standard normal (Amharic: $\bar{z}_i$ = -0.32, SD = 0.88; Oromifa: $\bar{z}_i$ = -0.31, SD = 0.94; Tigrinya: $\bar{z}_i$ = -0.31, SD = 0.90). However, in Oromifa and Tigrinya, the test did not perform as well in the first round of administration (Table 7.1). Moreover, a closer look at the individual fit statistics showed five extreme cases in Amharic, two in Oromifa and two in Tigrinya. In each of the language groups, person misfit is characterised by respondents having a perfect total raw score on the PPVT (i.e. all items answered correctly), with the exception of one respondent in Tigrinya who received a raw score of zero in the first administration of the test.

One of the main advantages of the Young Lives survey is the breadth of the data set. The survey provides information on many themes that are important for child development, such as home socioeconomic contexts, child health, nutrition and education, pregnancy, breastfeeding, and infant health. However, as a third limitation, the survey does not measure each construct comprehensively, but rather screens for different aspects of life. Moreover, the social and cultural processes that are known to shape vocabulary development and learning are not prioritised in the survey, i.e. conversation with parents and peers, listening to stories or books read aloud, classroom lectures and discussion, reading, and television (Nagy and Herman, 1987; Leung, 1992; McKeown, 1985; Werner and Kaplan, 1950). Thus, the present study was not able to explore different aspects of infant health and nutrition, family and community interactions, the home learning environment, language instruction and children's preschool experiences. The data structure does include some information on pregnancy, delivery, breastfeeding practices, and infant health. However, this information is narrow, based on caregiver reports and retrospectively collected, i.e. recorded when the children were one year old. Moreover, the variables are also significantly skewed and there is a large amount of missing data. Alternatively, the current study focused on three variables with good measurement properties: height-for-age as a measure of nutrition/stunting, mothers' education as a proxy measure of home factors, and preschool engagement.

7.4 Future research directions

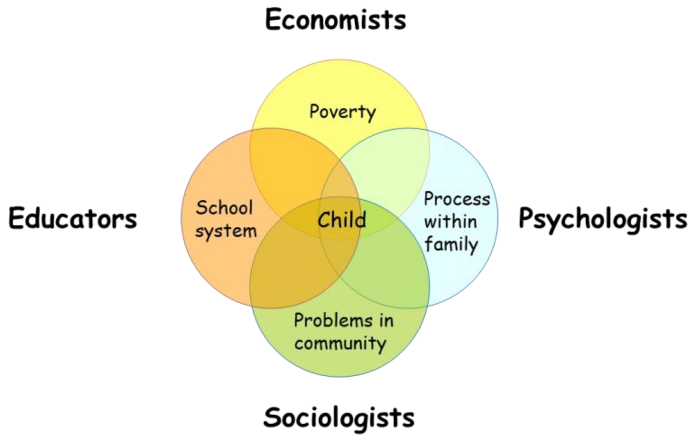
For several years I worked as a social policy officer with UNICEF in sub-Saharan Africa, managing different stages of the policy cycle. Designing research and coordinating advocacy initiatives provided exposure to real world challenges and prepared the ground for me to engage more substantively with the issues that I encountered in the field. Prior to working with UNICEF, I trained in Economics and International Development Studies. When launching into the D.Phil, my perspective as a researcher was significantly guided by these experiences. Throughout the research process my understanding of children's development and learning has become less utilitarian and much more holistic/multidisciplinary. In the next phase of my work, I hope to transform the lessons learned from the D.Phil study into practical resources for researchers working on the measurement of children's development and learning in Africa.

In terms of applied research, the findings suggest different research directions. Building on the issues raised in the previous section, an important future research priority would be to examine the longitudinal relationship between the scores obtained from different tests, for example, the PPVT scores and the results of the curriculum-based mathematics and literacy assessments that were also administered. The survey data includes an Early Grade Reading Assessment (EGRA), which was administered when the children were eight years old. The EGRA was followed by more substantive literacy tests at 12 and 15 years. To assess mathematics, the children were given a basic maths test at eight years and more comprehensive tests at 12 and 15 years. The primary aim of such a study would be to explore the relationship between cognitive development and academic achievement, and the extent to which academic achievement serves as a mitigating mechanism of socioeconomic risk (Asbury et al., 2005: 658; Noble et al., 2006). Such a study would also contribute to validating the assessments, which, as the current study has shown, is an under researched area in low-income countries.

7.5 Final thoughts

The findings highlight the need for meaningful outcome measures and in-depth data on the environments that shape development and learning for the design and evaluation of effective interventions, programmes, and policies. Most importantly, the study's findings highlight the importance of multidisciplinary research on children's development, which brings together experts from different fields, each drawing on the scientific knowledge within their own discipline, to develop robust research approaches that are centred on achieving better outcomes for children (Figure 7.2).

Figure 7.2 Perspectives of how the environment shapes individual differences



Source: Sameroff and Fiese, cited by Shonkoff & Meisels, 2000

List of references

Aber, J. L., Bennett, N. G., Conley, D. C., and Li, J. (1997) The effects of poverty on child health and development. *Annual review of public health*, *18* (1), pp. 463-483.

Adams, R. and Wu, M. (2002) PISA 2000 technical report. Paris, France: Organisation for Economic Co-operation and Development.

Adams, R. J., Wu, M. L. and Carstensen, C. H. (2007) Application of multivariate Rasch models in international large scales educational assessment. In M. v. Davier and C. H. Carstensen (Eds.), Multivariate and mixture distribution Rasch models: Extensions and applications (pp. 271-280): Kiel, Germany: Springer.

Alexander, K. L., Entwisle, D. R. and Olson, L. S. (2007) Lasting consequences of the summer learning gap. *American Sociological Review*, *72* (2), pp. 167-180.

Akire, S., and Santos, M. (2010) Acute Multidimensional Poverty: A New Index for Developing Countries. Human Development Research Paper 2010/11. UNDP-HDRO. New York. http:// hdr.undp.org/en/content/acute-multidimensional-poverty. [Accessed 18 March 2016]

Alkire, S., Roche, J. M., Santos, M. E., and Seth, S. (2011) Multidimensional Poverty Index 2011: Brief Methodological Note. OPHI Briefing 07. Oxford, UK: University of Oxford.

Alkire, S., Jindra, C., Robles, G., and Vaz, A. (2016). Multidimensional Poverty Index 2016: Brief methodological note and results. OPHI Briefing 47. Oxford, UK: University of Oxford.

Alwin, D. F., and Thornton, A. (1984) Family Origins and the Schooling Process: Early versus Late Influence of Parental Characteristics. *American Sociological Review*, 49, pp.784-802.

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

Anderson, R. C. and Freebody, P. (1981) Vocabulary knowledge. In T. Guthrie (Ed.), Comprehension and teaching: Research reviews (pp. 77-117). Newark: International Reading Association.

Andrich, D. (1982) An index of person separation in latent trait theory, the traditional KR.20 index, and the Guttman scale response pattern. *Education Research and Perspectives*, 9, 95-104.

Andrich, D. (1985). A latent trait model for items with response dependencies: Implications for test construction and analysis. In S. E. Embretson (Ed.), Test design (pp. 245-275). New York: Academic Press.

Andrich, D. (2004). Controversy and the Rasch model: a paradigm of incompatible paradigms. *Medical Care*, 42(1), pp.1-7.

Andrich, D. and Kreiner S. (2010). Quantifying response dependence between two dichotomous items using the Rasch model. *Applied Psychological Measurement*, 34(3), pp.181-192.

Andrich, D. and Hagquist, C. (2012) Real and artificial differential item functioning. *Journal of Educational and Behavioural Statistics*, 37(3), pp.387-416.

Andrich, D., Marais, I. and Humphry, S. (2012). Using a theorem by Andersen and the dichotomous Rasch model to assess the presence of random guessing in multiple choice items. Journal of Educational and Behavioural Statistics, 37(3), 417-442.

Andrich, D. Sheridan, B. and Luo, G. (2014) RUMM2030: Rasch unidimensional models for measurement. RUMM Laboratory, Perth.

Andrich, D., Sheridan, B. and Luo, G. (2009) Displaying RUMM2030. RUMM Laboratory Pty Ltd., Perth.

Andrich, D., Sheridan, B. and Luo, G. (2009) Interpreting RUMM2030. RUMM Laboratory Pty Ltd., Perth.

Araya, M. and Woldehanna, T. (2019) Poverty and Inequality in Ethiopia, 1995/6–2015/16. In F. Cheru, C. Cramer and A. Oqubay (Eds.). The Oxford Handbook of the Ethiopian Economy (pp. 1-48). Oxford, UK: Oxford University Press.

Arnett, J. (1989) Caregivers in Day-Care Centres: Does training matter? *Journal of Applied Developmental Psychology*, 10, pp. 541-552.

Asbury, K., and Plomin, R. (2013) *G is for genes: The impact of genetics on education and achievement* Vol. 24. John Wiley & Sons.

Asbury, K., Wachs, T. D., and Plomin, R. (2005) Environmental moderators of genetic influence on verbal and nonverbal abilities in early childhood. *Intelligence*, *33* (6), pp. 643-661.

Azubuike, O. B. and Briones, K. (2016) Young Lives Rounds 1 to 4 Constructed files. *Young Lives Technical Note*, *35*.

Baird, J., Andrich, D., Hopfenbeck, T.N., and Stobart, G. (2017) Assessment and learning: fields apart? *Assessment in Education: Principles, Policy & Practice*, 24(3), pp. 317-350.

Bane, M.J. and Ellwood, D. (1996) Slipping Into and Out of Poverty: The Dynamics of Spells. *Journal of Human Resources*, 21(1), pp. 1-23.

Barley, E. A. and Jones, P. W. (2006) Repeatability of the Rasch model of the AQ20 over five assessments. *Quality of Life Research*, 15, pp. 801-809.

Barnett, W. S. (1995) Long-term effects of early childhood programs on cognitive and school outcomes. *The future of children*, 5 (3), pp. 25-50.

Barnett, W. S., and Hustedt, J. T. (2005) Head Start's lasting benefits. *Infants & Young Children*, *18* (1), pp. 16-24.

Baron R.M., Kenny D.A. (1986) The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, pp. 1173–1182.

Beck, I.L., McKeown, M.G. (2007) Different ways for different goals, but keep your eye on the higher verbal goals. In R. Wagner, A. Muse, K. Tannenbaum (Eds.) Vocabulary acquisition: Implications for reading comprehension (pp.182-204). New York: Guilford Press.

Beck, I. L., Perfetti, C. A., and McKeown, M. G. (1982) Effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology*, 74(4), pp. 506–521.

Becker, G.S. (1964) *Human capital: a theoretical and empirical analysis, with special reference to education.* New York: National Bureau of Economic Research.

Berry, J.W. (1969) On cross-cultural comparability. *International Journal of Psychology*, 4, pp.119-128.

Biesanz, J. C., Deeb-Sossa, N., Papadakis, A. A., Bollen, K. A. and Curran, P. J. (2004) The role of coding time in estimating and interpreting growth curve models. *Psychological methods*, *9*(1), p. 30.

Birnbaum, A. (1968) Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M. R. Novick (Eds.) Statistical Theories of Mental Test Scores (pp. 397-545). Reading, Mass: Addison-Wesley.

Bland, J.M. and Altman, D.G. (1995). Multiple significance tests: the Bonferroni method. *British Medical Journal*, 310, p. 170.

Blum, A., Goldstein, H. and Guérin-Pace, F. (2001) International Adult Literacy Survey (IALS): An analysis of international comparisons of adult literacy, Assessment. *Education: Principles, Policy & Practice*, 8(2), pp. 225-246.

Bond, T.N. and Lang, K. (2013) The evolution of the black-white test score gap in grades K-3: the fragility of results. *The Review of Economics and Statistics*, 95(5), pp. 1468–1479.

Bonnie, R. J., and Backes, E.P. (2019) The promise of adolescence: Realizing opportunity for all youth. Washington, D.C.: National Academy Press.

Borsboom, D. (2005) Measuring the mind: Conceptual issues in contemporary psychometrics. Cambridge: Cambridge University Press.

Bornstein, M. H. and Putnick, D. L. (2012) Cognitive and socioemotional caregiving in developing countries. *Child development*, *83* (1), pp. 46-61.

Bradbury, B. and Jäntti, M. (1999) Child poverty across industrialized nations. Innocenti Occasional Papers, Economic and Social Policy Series 71, Florence: UNICEF International Child Development Centre.

Bradley, R. H. and Caldwell, B. M. (1984) The relation of infants' home environments to achievement test performance in first grade: A follow-up study. *Child development*, pp. 803-809.

Bradley, R.H. and Corwyn, R.F. (2002) Socioeconomic status and child development. *Annual review of psychology*, *53*(1), pp.371-399.

Bradley, R. and Corwyn, R. (2005) Caring for children around the world: A view from HOME. *International Journal of Behavioral Development*, *29* (6), pp. 468-478.

Bradshaw, J., Middleton, S., Williams, J. and Ashworth, K. (2000) The Necessities of Life for Children, Working Paper No. 2. Townsend Centre for International Poverty Research: University of Bristol.

Breen, R. and Goldthorpe, J.H. (1997) Explaining educational differentials towards a formal rational action theory. *Rationality and society*, *9*(3), pp.275-305.

Brooks-Gunn J. and Duncan G.J. (1997) The effects of poverty on children. *Future Child*. 7(2), pp. 55–71.

Brooks-Gunn, J., G. J. Duncan, and Klebanov, P. (1996) Ethnic Differences in Children's Intelligence Test Scores: Role of Economic Deprivation, Home Environment and Maternal Characteristics. *Child Development,* 67, pp. 396–408.

Brown G., Micklewright J., Schnepf S.V., and Waldmann R. (2007) International surveys of educational achievement: how robust are the findings? *Journal of the Royal Statistical Society,* 170(3), pp. 623–46.

Bruns, B., Martinez, S. Naudeau, S. and Pereira, V. (2010) Impact evaluation of Save the Children early childhood development program in Mozambique. baseline results. World Bank: Washington, DC.

Bryman, A. (2012) *Social Research Methods*. Oxford, UK: Oxford University Press.

Byrne, B.M. and Watkins, D. (2003) The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology* 34, pp. 155-175.

Byrne B. M. and van de Vijver. F. (2010) Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of non-equivalence. *International Journal of Testing* 10, pp. 107-132.

Calderon, M.C., and Kovacevic, M. (2014) The 2014 Multidimensional Poverty Index: New Specification. Human Development Research Paper. UNDP-HDRO, New York.

Campbell, A. and Rudan, I. (2011) Systematic review of birth cohort studies in Africa. *Journal of Global Health*, 1(1), pp. 46–58.

Campbell, J.M. (1998) Review of the Peabody Picture Vocabulary Test, Third Edition, *Journal of Psychoeducational Assessment* 16(4), pp. 334-8.

Campbell, J.M., Bell, S.K. and Keith, L.K. (2001) Concurrent Validity of the Peabody Picture Vocabulary Test Third Edition as an Intelligence and Achievement Screener for Low SES African American Children, *Assessment* 8(1), pp. 85-94.

Carranza, E. and Gallegos, J. (2013) Ethiopia Trends of Wellbeing 2000-2011. *Poverty Reduction and Economic Management, Africa Region*. Washington, D.C: World Bank.

Carroll, J.B. (1974) Fitting a Model of School Learning to Aptitude and Achievement Data over Grade levels. In D.R. Green (Ed.) The Aptitude-Achievement Distinction, Proceedings (pp. 53-78). CTB/McGraw-Hill.

Centers for Disease Control and Prevention (2013) Cut-offs to define outliers in the 2000 CDC growth charts. https://www.cdc.gov/nccdphp/dnpa/growthcharts/00binaries/BIV-cutoffs.pdf [Accessed 10 March 2020]

Champion, T.B., Hyter, Y.D., McCabe, A. and Bland-Stewart, L.M. (2003) A Matter of Vocabulary: Performances of Low-Income African American Head Start Children on the Peabody Picture Vocabulary Test–III, *Communication Disorders Quarterly* 24, pp.121-7.

Chase-Lansdale, P. L., Gordon, R. A., Brooks-Gunn, J. and Klebanov, P. K. (1997) Neighborhood and family influences on the intellectual and behavioral competence of preschool and early school-age children. *Neighborhood poverty*,*1*, pp. 79-118.

Chen, G., Kirkman, B.L., Kanfer, R., Allen, D. and Rosen, B. (2007) A multilevel study of leadership, empowerment, and performance in teams. *The Journal of Applied Psychology*, 92, pp. 331-346.

Cheung, Y.B. (2006) Growth and cognitive function of Indonesian children: zero-inflated proportion models. *Stat Med* 25, pp. 3011–22.

Chien, N. C., Howes, C., Burchinal, M., Pianta, R. C., Ritchie, S., Bryant, D. M., and Barbarin, O. A. (2010) Children’s classroom engagement and school readiness gains in prekindergarten. *Child development*, *81*(5), pp. 1534-1549.

Collins, W. A. (1984) Development during middle childhood: The years from six to twelve. Washington, D.C.: National Academy Press.

Corcoran, M., Gordon, R., Laren, D., and Solon, G. (1992) The association between men's economic status and their family and community origins. *Journal of Human Resources*, 27, pp. 575-601.

Cronbach, L.J. (1971) Test Validation. In R.L. Thorndike (Ed.) Educational Measurement, 2nd ed. (pp. 443-507). Washington, D.C.: American Council on Education.

Cronbach, L.J. and Shavelson, R.J. (2004) My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement* 64(3), pp.391-418.

Crookston, B. T., Penny, M. E., Alder, S. C., Dickerson, T. T., Merrill, R. M., Stanford, J. B. and Dearden, K. A. (2010) Children who recover from early stunting and children who are not stunted demonstrate similar levels of cognition. *The Journal of Nutrition*, 140(11), pp. 1996–2001.

Crookston, B. T., Forste, R., McClellan, C., Georgiadis, A., and Heaton, T. B. (2014) Factors associated with cognitive achievement in late childhood and adolescence:

the Young Lives cohort study of children in Ethiopia, India, Peru, and Vietnam. *BMC Pediatrics*, 14 (253).

CSA and UNICEF Ethiopia (2018) Multidimensional Child Deprivation in Ethiopia. Addis Ababa, Ethiopia: UNICEF and Central Statistical Agency of Ethiopia.

Cueto, S., and Leon, J. (2012) Psychometric characteristics of cognitive development and achievement instruments in Round 3 of Young Lives. Oxford, UK: Young Lives.

Cueto, S., Leon, J., Guerrero, G. and Munoz, I. (2009) Psychometric characteristics of cognitive development and achievement instruments in Round 2 of Young Lives. Oxford, UK: Young Lives.

Cueto, S., Singh, R. And Woldehanna, T. and Duc, L.T. (2016) Education trajectories in Ethiopia, India, Peru and Vietnam: from early childhood to early adulthood. Oxford, UK: Young Lives.

Currie J., and Thomas, D. (1999) Early test scores, socioeconomic status and future outcomes, Working Paper 6943. Cambridge: National Bureau of Economic Research.

Das, J. and Zajonc, T. (2010) India shining and Bharat drowning: comparing two Indian states to the worldwide distribution in mathematics achievement. *Journal of Development Economics* 92(2), pp.175–187.

Davis, F.B. (1968) Research in comprehension in reading. *Reading Research Quarterly*, pp. 499-545.

De Neubourg, C., Chai, J., de Milliano, M., Plavgo, I., and Wei, Z. (2013) Cross-country MODA study: multiple overlapping deprivations analysis technical note. New York: UNICEF.

De Moura, D. R., Costa, J. C., Santos, I. S., Barros, A. J., Matijasevich, A., Halpern, R., Dumith, S., Karam, S. and Barros, F. C. (2010a), Natural history of suspected developmental delay between 12 and 24 months of age in the 2004 Pelotas birth cohort. *Journal of Paediatrics and Child Health*, 46, pp. 329–336.

De Moura, D. R., Costa, J. C., Santos, I. S., Barros, A. J., Matijasevich, A., Halpern, R., Dumith, S., Karam, S. and Barros, F. C. (2010b), Risk factors for suspected developmental delay at age 2 years in a Brazilian birth cohort. *Paediatric and Perinatal Epidemiology*, 24, pp. 211–221.

Deaton, A. and Grosh, M. (2000) Consumption. In M. Grosh, and P. Glewwe (Eds.) Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study. The World Bank: Washington.

Deaton, A., and Zaidi, S. (2002) Guidelines for constructing consumption aggregates for welfare analysis. Living Standards Measurement Study Working Paper 135, World Bank, Washington, DC.

Denman D., Speyer, R., Munro, N., Pearce, W.M., Chen, Y-W and Cordier, R. (2017) Psychometric Properties of Language Assessments for Children Aged 4–12 Years: A Systematic Review. *Frontiers in Psychology* 8, pp. 1515.

Deregowski, J. B., and Serpell, R. (1971) Performance on a sorting task: A cross-cultural experiment. *International Journal of Psychology*, 6, pp. 273-281.

Desforges, C., with Abouchaar, A. (2003). *The Impact of Parental Involvement, Parental support and Family Education on Pupil Achievements and Adjustment*: *A Literature Review.* (Report No. 433). Department for Education and Skills: Queen's Printer.

Dickerson, A. & Popli, G.K. (2012). Persistent poverty and children's cognitive development: evidence from the UK Millennium Cohort Study. Centre for Longitudinal Studies Working Paper No. 2012/2, London: Institute of Education.

Dickson, H.G., Koehler, F. (1996) The multi-dimensionality of the FIM motor items precludes an interval scaling using Rasch analysis. Scandinavian Journal of Rehabilitation Medicine, 26, pp. 159-62.

DiPrete, T.A. and Eirich, G.M. (2006) Cumulative advantage as a mechanism for inequality: A review of theoretical and empirical developments. *Annual review of sociology*, pp.271-297.

Dockrell, J. E., and Marshall, C. R. (2015) Measurement issues: assessing language skills in young children. *Child and Adolescent Mental Health*, *20*(2), pp. 116-125.

Draper, J.F. (1974) Open Discussion on the Logic of the Aptitude — Achievement Distinction. In D.R. Green (Ed.) The Aptitude-Achievement Distinction, Proceedings (pp. 335-44). CTB/McGraw-Hill.

Drèze, J. and Sen, A. (1995) India: economic development and social opportunity. Delhi: Oxford University Press.

Duncan, G. J. and Brooks-Gunn, J. (1997) Income Effects across the Life Span: integration and interpretation. In G.J. Duncan and Brooks-Gunn, J. (Eds.) Consequences of Growing Up Poor (pp. 596-610). Russell Sage Foundation.

Duncan, G., Brooks-Gunn, J., and Klebanov, E. (1994) Economic deprivation and early childhood development. *Child Development*, 65, pp. 296-318.

Duncan, G., and Rodgers, W. (1988) Longitudinal aspects of childhood poverty. *Journal of Marriage and the Family*, 50, pp. 1007-1021.

Duncan, G. J., Yeung, W. J., Brooks-Gunn, J., and Smith, J. R. (1998) How much does childhood poverty affect the life chances of children? *American Sociological Review*, 63, pp. 406-423.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C. and Klebanov, P. (2007) School readiness and later achievement. *Developmental Psychology,* 43, pp. 1428–1446.

Dunn, L. M. and Dunn, L. M. (1997) Examiner's Manual for the PPVT-III. *Circle Pines, MN: American Guidance Service*.

Dunn, L.M., Dunn. L.M., Whetton, C. and Burley, J. (1997) British Picture Vocabulary Scale. 2nd edition. Windsor, Berks: NFER-Nelson.

Dutton, D. B., and Levine, S. (1989) Overview, methodological critique, and reformulation. In J. P. Bunker, D. S. Gomby, and B. H. Kehrer (Eds.), Pathways to health (pp. 29–69). Menlo Park, CA: The Henry J. Kaiser Family Foundation.

Embretson, S.E. (1996) The new rules of measurement. *Psychological Assessment*, 8, pp. 341–349.

Embretson, S. E. and Hershberger, S. L. (1999) *The new rules of measurement: What every psychologist and educator should know*. Psychology Press.

Enders, C.K. and Tofighi, D. (2007) Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological methods*, *12*(2), pp.121-138.

Engle, P.L., Black, M.M., Behrman, J.R., Cabral de Mello, M., Gertler, P.J., Kapiriri, L., Martorell, R., Young, M.E. (2007) Child development in developing countries 3: strategies to avoid the loss of developmental potential in more than 200 million children in the developing world. *The Lancet,* 369, 229-242.

Entwisle, D. R., and Alexander, K. L. (1989) Early schooling as a “critical period” phenomenon. *Sociology of education and socialization*, *8*, pp. 27-55.

Ercikan, K. and Lyons-Thomas, J. (2013) Adapting tests for use in other languages and cultures. In K.F. Geisinger, B.A. Bracken, J.F. Carlson, J.C. Hansen, N.R. Kuncel, S.P. Reise, M.C. Rodriquez (Eds.), APA handbook of testing and assessment in psychology, vol.3: testing and assessment in school psychology and education (pp. 545-569). Washington, D.C: American Psychological Association.

Ethiopian Central Statistical Agency and ICF International (2011) *Ethiopia Demographic and Health Survey 2011*. Addis Ababa, Ethiopia: Central Statistical Agency.

Eurostat (2014) Living conditions in Europe. Luxembourg: European Commission.

Evangelou, M., Sylva, K., Edwards, A. & Smith, T. (2008). *Supporting Parents in Promoting Early Learning: The Evaluation of the Early Learning Partnership Project.* London: DCSF.

Farah, M.J., Shera, D.M., Savage, J.H., Betancourt, L., Giannetta, J.M., Brodsky, M.L., Malmud, E.K., and Hurt, H. (2006) Childhood poverty: specific associations with neurocognitive development. *Brain Research*, 1110, pp. 166–174.

Farkas, G. and Beron, K. (2004) The detailed age trajectory of oral vocabulary knowledge: Differences by class and race. *Social Science Research*, *33*(3), pp. 464-497.

Feinstein L (2003) Inequality in the early cognitive development of children in the 1970 cohort. *Economica*, 70, pp. 73–97.

Feinstein, L. (2004). Mobility in pupils' cognitive attainment during school life.*Oxford Review of Economic Policy*, *20* (2), pp. 213-229.

Fernald, L. C., Kariger, P., Hidrobo, M., and Gertler, P. J. (2012) Socioeconomic gradients in child development in very young children: Evidence from India, Indonesia, Peru, and Senegal. *Proceedings of the* National Academy of Sciences, 109 (Suppl. 2), 17273.

Fernald, L. C., Kariger, P., Engle, P., and Raikes, A. (2009) Examining early child development in low-income countries: a toolkit for the assessment of children in the first five years of life.

Fernald L.C.H., Weber, A., Galasso, E. and Ratsifandrihamanana, L. (2011) Socioeconomic gradients and child development in a very low-income population: evidence from Madagascar. *Developmental Science*, 14(4), pp. 832–847.

Fiegehen, G. C, Lansley, P. S. and Smith, A. D. (1977) *Poverty and Progress in Britain 1953-73*, Cambridge: Cambridge University Press.

Filmer, D. and Pritchett, L.H. (2001) Estimating wealth effects without expenditure data— or tears: an application to educational enrollments in states of India. Demography 2001; 38, pp.115–32.

Filmer, D. and Naudeau, S. (2010) Impact evaluation on early childhood care and development (ECCD programs in Cambodia, baseline Report). World Bank, Washington, DC.

Fischer, G. H. and Molenaar, I. W. (1995) *Rasch Models: Foundations, Recent Developments, and Applications*. New York: Springer-Verlag.

Fraley, R. C., Waller, N. G. and Brennan, K. A. (2000) An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology*, 78, pp. 350-365.

Frisbie, D.A. (1988) NCME instructional module on reliability of scores from teacher-made tests. Nebraska-Lincoln: ITEMS.

Frongillo, E. A., Tofail, F., Hamadani, J. D., Warren, A. M. and Mehrin, S. F. (2014) Measures and indicators for assessing impact of interventions integrating nutrition, health, and early childhood development. *Annals of the New York academy of sciences*, *1308*(1), pp. 68-88.

Furlong, K.R., Anderson, L.N., Kang H., Lebovic, G., Parkin, P.C., Maguire, J.L. et al. (2016) BMI-for-age and weight-for-length in children 0 to 2 years. *Pediatrics*,138 (1), pp.1-9.

Galasso, E., Weber, A., and Fernald, L. C. (2017) Dynamics of child development: Analysis of a longitudinal cohort in a very low-income country. *The World Bank Economic Review*, 33(1), pp. 140-159.

Galindo, C. and Sheldon, S.B. (2011) School and home connections and children's kindergarten achievement gains: The mediating role of family involvement. *Early Childhood Research Quarterly*.

Gebhardt, E. and Adams, R. J. (2007) The influence of equating methodology on reported trends in PISA. *Journal of Applied Measurement, 8*(3), pp. 305–322.

Gewirtz, J.L. (1969) Mechanisms of social learning: some roles of stimulation and behavior in early human development. In D.A. Goslin (Ed.) Handbook of Socialization Theory and Research (pp. 57-212). Rand McNally.

Gierl, M. J. (2000) Construct equivalence on translated achievement tests. *Canadian Journal of Education* 25(4), pp. 280–296.

Glewwe, P., Jacoby, H. G. and King, E. M. (2001) Early childhood nutrition and academic achievement: a longitudinal analysis. *Journal of public economics*, *81*(3), pp. 345-368.

Goldstein, H. (1979) Consequences of Using the Rasch Model for Educational Measurement. *British Educational Research Journal*, 5(2), pp. 211-220.

Goldstein, H. (1980) Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology,* 33, pp. 234–246.

Goldstein, H. (2004) International comparisons of student attainment: some issues arising from the PISA debate. *Assessment in Education*, 11, pp. 319–330.

Gomes-Neto, J. B., Hanushek, E. A., Leite, R. H., and Frota-Bezzera, R. C. (1997) Health and schooling: Evidence and policy implications for developing countries. *Economics of Education Review*, *16*(3), pp. 271-282.

Goodman, R. (1997) The Strengths and Difficulties Questionnaire: a research note. *Journal of Child Psychology and Psychiatry*, 38(5), pp. 581-586.

Gordon, D., Nandy, S., Pantazis, C., Pemberton, S. and Townsend, P. (2003) The distribution of child poverty in the developing world, Bristol: University of Bristol.

Gordon, D., Pantazis, C. and Townsend, P. (2000) Changing Necessities of Life 1983-1999. Poverty and Social Exclusion Survey of Britain, Working Paper No. 3, Townsend Centre for International Poverty Research: University of Bristol.

Graber, J. A., and Brooks-Gunn, J. (1996). Transitions and turning points: Navigating the passage from childhood through adolescence. *Developmental psychology*, *32*(4), 768.

Grantham-McGregor, S. (1995) A review of studies of the effect of severe malnutrition on mental development. *The Journal of nutrition*, *125* (suppl_8), pp. 2233S-2238S.

Grantham-McGregor, S., Cheung, Y.B., Cueto, S., Glewwe, P., Richter, L. and Strupp, B. (2007) Child development in developing countries 1: developmental potential in the first 5 years for children in developing countries. *The Lancet*, 369, pp. 60-70.

Grantham-McGregor, S. M., Fernald, L. C., and Sethuraman, K. (1999a) Effects of health and nutrition on cognitive and behavioural development in children in the first three years of life. Part 1: Low birthweight, breastfeeding, and protein-energy malnutrition. *Food and nutrition Bulletin*, *20*(1), pp. 53-75.

Grantham-McGregor, S. M., Fernald, L. C., and Sethuraman, K. (1999b) Effects of health and nutrition on cognitive and behavioural development in children in the first three years of life. Part 2: infections and micronutrient deficiencies: iodine, iron, and zinc. *Food and nutrition Bulletin*, *20*(1), pp. 76-99.

Gray, J. R. and Thompson, P. M. (2004) Neurobiology of intelligence: science and ethics. *Nature Reviews Neuroscience*, *5*(6), pp.471-482.

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

Grosh, M. and Glewwe, P. (2000) Designing household survey questionnaires for developing countries: lessons from 15 years of the living standards measurement study. Washington: World Bank.

Guo G. (1998) The timing of the influences of cumulative poverty on children's cognitive ability and achievement. *Social Forces* 77, pp. 257- 88.

Hackman, D.A., and Farah, M.J. (2009) Socioeconomic status and the developing brain. *Trends in Cognitive Sciences*, 13 (2), pp. 65–73.

Hagquist, C. and Andrich, D. (2004) Is the sense of coherence-instrument applicable on adolescents? A latent trait analysis using Rasch-modelling. *Personality and Individual Differences,* 36, pp. 955–68.

Hagquist, C., Bruce, M. and Gustavsson, J. P. (2009) Using the Rasch model in nursing research: An introduction and illustrative example. *International Journal of Nursing Studies*, 46(3), pp. 380-393.

Haitana, T., Pitama, S. and Rucklidge, J. J. (2010) Cultural biases in the peabody picture vocabulary test-III: testing tamariki in a New Zealand sample. *New Zealand Journal of Psychology*, *39*(3), pp. 24-34.

Hall, J., and Sammons, P. (2013) Mediation, Moderation and Interaction. In *Handbook of Quantitative Methods for Educational Research* (pp. 267-286). Springer, Rotterdam.

Hall, J., Sylva, K., Melhuish, E., Sammons, P., Siraj-Blatchford, I., & Taggart, B. (2009). The role of pre-school quality in promoting resilience in the cognitive development of young children. *Oxford Review of Education*, *35*, 331-352.

Hamadani, J. D., Tofail, F., Huda, S. N., Alam, D. S., Ridout, D. A., Attanasio, O., and Grantham-McGregor, S. M. (2014) Cognitive deficit and poverty in the first 5 years of childhood in Bangladesh. *Pediatrics, 134(4)*, e1001–e1008.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. Merenda, and C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K. and Lee, M.K. (2013) Methods for translating and adapting tests to increase cross-language validity. In D.H. Saklofske, C.R. Reynolds and V. Schwean (Eds.), The Oxford handbook of child psychological assessment (pp. 172-181). Oxford, UK: Oxford University Press.

Hambleton, R. K. and Patsula, L. (1998) Adapting tests for use in multiple languages and cultures. *Social Indicators Research* 45, pp.153-171.

Hambleton, R. K., and Zenisky, A. L. (2011). Translating and adapting tests for cross-cultural assessments. In D. Matsumoto and F. J. R. van de Vijver. (Eds.), Cross-cultural research methods in psychology (pp. 46–70). New York: Cambridge University Press.

Harkness, J.A. (1998) *Cross-cultural survey equivalence*. Mannheim, Germany: ZUMA.

Harkness, S., and Super, C. M. (2002) Culture and parenting. In M. H. Bornstein (Ed.), *Handbook of parenting* (2nd ed.). Hillsdale, NJ: Erlbaum.

Harkness, J. A., Van de Vijver, F. and Mohler, P. (2002) *Cross-cultural Survey Methods*. Chichester: Wiley.

Harms, T., Clifford, R. and Cryer, D. (1998) Early Childhood Environment Rating Scale, Revised Edition (ECERS-R). New York: Teachers College Press.

Harris, J. R. (1995). Where is the child's environment? A group socialization theory of development. Psychological review, 102(3), 458.

Hart, B. and Risley, T. (1995) Meaningful Differences in Everyday Experience of Young American Children. Baltimore, Maryland: Paul H. Brookes Publishing Co.

Hart, B. and Risley, T. (1999). The Social World of Children Learning to Talk. Paul Brookes, Baltimore, Maryland: Paul H. Brookes Publishing Co.

Hart, B. and Risley, T. R. (2003) The early catastrophe: The 30-million-word gap by age 3. *American educator*, *27*(1), pp. 4-9.

Haughton, J. and Khandker, S. R. (2009) *Handbook on poverty and inequality*. Washington, D.C: World Bank.

Haveman, R., Wolfe, B., and Spaulding, J. (1991) Childhood events and circumstances influencing high school completion. *Demography*, 28, pp. 133-157.

Haveman, R. and Wolfe, B. (1995) The determinants of children's attainments: a review of methods and findings. *Journal of Economic Literature*, 33, pp. 1829-1878.

Heck, R. H., Thomas, S. (2015) An Introduction to Multilevel Modeling Techniques. New York: Routledge.

Heck, R. H., Tabata, L. and Thomas, S. L. (2013) Multilevel and longitudinal modeling with IBM SPSS. New York: Routledge.

Helms, J. E. (1992) Why is there no study of cultural equivalence in standardized cognitive ability testing? *American Psychologist* 47(9), pp. 1083–1101.

Henderson, V. and Dweck, C.S. (1990) Motivation and achievement. In S. Feldman and G.R. Elliot (Eds.) At the threshold: the developing adolescent (pp. 308-29). Harvard University Press.

Hill, M. S., and Duncan, G. (1987) Parental family income and the socioeconomic attainment of children. *Social Science Research*, 16, pp. 39-73.

Hill, R. H. and Porter, C. (2017) Vulnerability to drought and food price shock: evidence from Ethiopia. *World Development*, 96, pp. 65-77.

Hoekstra, R. A., Bartels, M., and Boomsma, D. I. (2007) Longitudinal genetic study of verbal and nonverbal IQ from early childhood to young adulthood. *Learning and Individual Differences*, *17* (2), pp. 97-114.

Holland, P. W., and Wainer, H. (1993) *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., and Baird, J. A. (2018) Lessons learned from PISA: A systematic review of peer-reviewed articles on the programme for international student assessment. *Scandinavian Journal of Educational Research*, *62* (3), pp. 333-353.

Hopmann, S.T., Brinek, G. and Retzl, M. (2007) PISA zufolge PISA: Hält PISA, was es verspricht? / PISA according to PISA: Does PISA keep what it promises? Vienna: LIT Verlag.

Hui, C. H. and H. C. Triandis (1985) Measurement in cross cultural psychology, *Journal of Cross-Cultural Psychology* 16, pp. 131–152.

IBM Corp. Released 2012. IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.

Jencks, C., Smith, M., Asland, H., Bane, M.J., Cohen, D., Gintis, H., Heyns, B., and S. Michelson (1972) Inequality A Reassessment of the Effect of Family and Schooling in America. Basic Books.

Jenkins, S.P. (2011). Changing Fortunes. Oxford: Oxford University Press.

Jensen, A. R. (1966) Cumulative deficit in compensatory education. *Journal of School Psychology*, *4*(3), pp. 37-47.

Jensen, A. R. (1974a) Cumulative deficit: A testable hypothesis? *Developmental Psychology*, *10*(6), pp. 996.

Jensen, A. R. (1974b) How biased are culture-loaded tests? *Genetic Psychology Monographs*, 90, pp. 185–244.

Jordan, G. E., Snow, C. E. and Porche, M. V. (2000) Project EASE: The Effect of a Family Literacy Project on Kindergarten Students' Early Literacy Skills. *Reading Research Quarterly*, Vol. 35, No. 4 (Oct. - Nov. - Dec., 2000), pp. 524-546.

Kagitcibasi, C., Sunar, D. and Bekman, S. (2001) Long-term effects of early intervention: Turkish low-income mothers and children. *Journal of Applied Developmental Psychology*, *22*(4), 333-361.

Kang, S.M. and Waller, N.G. (2005) Moderated multiple regression, spurious interaction effects, and IRT. *Applied Psychological Measurement* 29, pp.87–105.

Kiescht, M. and Vietze, M.S. (1975) Test stimuli: Representational level with middle class and head start children. *Psychology in the Schools* 12(3), pp. 309-313.

Kieffer, M. J. and Lesaux, N. K. (2012) Knowledge of words, knowledge about words: Dimensions of vocabulary in first and second language learners in sixth grade. *Reading and Writing*, 25(2), pp. 347-373.

Kieffer, M. J., and Stahl, K. D. (2015) Complexities of individual differences in vocabulary knowledge. In P. Afflerbach (Ed.) Handbook of individual differences in reading: reader, text and context (pp.120-137). London, UK: Routledge.

Kindermann, T., and Skinner, E. A. (1992) Modeling environmental development: Individual and contextual trajectories. In J. B. Asendorpf and J. Valsiner (Eds.), Stability and change in development (pp. 155-190). London: Sage.

Kindermann, T. A., and Valsiner, J. (1995) Introduction: Individual development, changing contexts, and the co-construction of person-context relations in human development. In T. A. Kindermann and J. Valsiner (Eds.), Development of person-context relations (pp. 1-12). Hillsdale, NJ: Erlbaum.

Kirsch, I., DeJong, J., LaFontaine, D., McQueen, J., Mendelovits, J. and Monseur, C. (2002) Reading for change: Performance and engagement across countries: Results from PISA 2000 (Publication No. ED474915). *Washington, DC: US Government Printing Office*.

Klasen, S., and Dotter, C. (2013) The Multidimensional Poverty Index: Achievements, Conceptual, and Empirical Issues. Human Development Research Paper. UNDP-HDRO, New York.

Klebanov, P. K., Brooks-Gunn, J., Chase-Lansdale, P. L. and Gordon, R. A. (1997) Are neighborhood effects on young children mediated by features of the home environment. *Neighborhood poverty: Context and consequences for children*, *1*, pp. 119-145.

Konishi, C., Hymel, S., Zumbo, B. D., Li, Z., Taki, M. and Slee, P. (2009) Investigating the comparability of a self-report measure of childhood bullying across countries. *Canadian Journal of School Psychology* 24(1), pp. 82–93.

Korenman, S., Miller, J., and Sjaastad, J. (1995) Long-term poverty and child development in the United States: Results from the NLSY. *Children and Youth Services Review*, *17*, pp. 127–155.

Kraemer, H. C., Stice, E., Kazdin, A., Offord, D., and Kupfer, D. (2001). How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. *American Journal of Psychiatry*, 158, 848–856.

Kreiner, S., and Christensen, K. (2014) Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika, 79*, pp. 210–231.

Krull, J. L. and MacKinnon, D. P. (2001) Multilevel modeling of individual and group level mediated effects. *Multivariate behavioral research*, *36*(2), pp. 249-277.

Laderchi, C.R. (2000) The monetary approach to poverty: a survey of concepts and methods. OPHI Working Paper Series, 58.

Lansdown, R. G., Goldstein, H., Shah, P. M., Orley, J. H., Di, G., Kaul, K. K., … Reddy, V. (1996). Culturally appropriate measures for monitoring child development at family and community level: a WHO collaborative study. *Bulletin of the World Health Organization*, 74(3), pp. 283–290.

La Paro, K. M., and Pianta, R.C. (2000) Predicting children's competence in the early school years. A meta-analytic review. *Review of Educational Research*, 70, pp. 443-84.

Lawton, G., Lundgren-Nilsson, A., Biering-Sorensen, F., Tesio, L., Slade, A., Penta M., Grimby, G., Ring, H. and Tennant, A. (2006) Cross-cultural validity of FIM in spinal cord injury. *Spinal Cord,* 44, pp. 746–52.

Lee, V. E., and Loeb, S. (1995) Where do Head Start attendees end up? One reason why preschool effects fade out. *Educational evaluation and policy analysis*, *17* (1), pp. 62-82.

León, J. and Collahua, Y. (2015) Technical note 33: the reliability and validity of achievement tests in the second Young Lives school survey in Ethiopia. Oxford, UK: Young Lives.

León, J., Miranda, A. and Cueto, S. (2015) Equating children's PPVT scores across survey rounds and age cohorts in Peru. Young Lives Technical Note, 32. Oxford, UK: Young Lives.

Leon, J. and Singh, A. (2017) Equating test scores for receptive vocabulary across rounds and cohorts in Ethiopia, India and Vietnam. *Young Lives Technical Note, 40.* Oxford, UK: Young Lives.

Leung C.B. (1992) Effects of word-related variables on vocabulary growth through CPS repeated read-aloud events. National Reading Conference Yearbook 41, pp. 491–498.

Laing, S. and Kamhi, A. (2003) Assessment of language and literacy in culturally and linguistically diverse populations. Language, Speech, and Hearing Services in the Schools, 34(1), pp. 44-55.

Linacre, J. M. (1996) Generalizability and manyfacet Rasch measurement. In G. Engelhard, Jr., and M. Wilson (Eds.), Objective measurement: Theory into practice: Vol. 3 (pp. 85- 98). Norwood, NJ: Ablex.

Little, R. J. (1988) Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, *6*(3), pp. 287-296.

Loner, E. and Peri, P. (2009) Ethnic identification in the former Soviet Union: Hypotheses and analyses. *Europe-Asia Studies* 61(8), pp. 1341–1370.

Looveer, J. and Mulligan, J. (2009) The efficacy of link items in the construction of a numeracy achievement scale—from kindergarten to year 6. *Journal of applied measurement*, *10*(3), pp.1-19.

López Boo, F. (2016) Socioeconomic status and early childhood cognitive skills: A mediation analysis using the Young Lives panel. *International Journal of Behavioral Development*, *40*(6), pp. 500-508.

Lord, F.M. and Novick. M.R. (1968) Statistical Theories of Mental Test Scores. Reading, Mass: Addison-Wesley.

Lozoff, B. and Teal, S.J.L. (2004) Longitudinal analysis of cognitive and motor effects of iron deficiency in infancy. *Pediatric Research*, 55, p. 23A

Ludlow, L.H. and O'Leary, M. (1999) Scoring omitted and not- reached items: practical data analysis implications. *Educational and Psychological Measurement,* 59(4), pp. 615- 603.

Lundeen, E. A., Behrman, J. R., Crookston, B. T., Dearden, K. A., Engle, P., Georgiadis, A., ... Stein, A. D. (2014) Growth faltering and recovery in children aged 1–8 years in four low-and middle-income countries: Young Lives. *Public Health Nutrition*, 17(09), pp. 2131–2137.

Maaz, K., Trautwein, U., Lüdtke, O. and Baumert, J. (2008) Educational transitions and differential learning environments: How explicit between-school tracking contributes to social inequality in educational outcomes. *Child Development Perspectives*, *2* (2), pp. 99-106.

Mack, J. and Lansley, S. (1985) Poor Britain. London: George Allen & Unwin.

MacKinnon, D. P. (2008) *Multivariate applications series. Introduction to statistical mediation analysis.* New York, NY: Taylor & Francis Group/Lawrence Erlbaum Associates.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G. and Sheets, V. (2002) A comparison of methods to test mediation and other intervening variable effects. *Psychological methods*, *7*(1), pp. 83-104.

MacKinnon, D. P., Warsi, G. and Dwyer, J. H. (1995) A simulation study of mediated effect measures. *Multivariate behavioral research*, *30* (1), pp. 41-62.

Macours, K., Schady, N., and Vakis, R. (2008). Cash transfers, behavioral changes, and cognitive development in early childhood: evidence from a randomized experiment. World Bank Policy Research Working Paper No. 4759, Washington, D.C.: World Bank.

Macours, K. and Vakis, R. (2010) Seasonal migration and early childhood development. *World development*, *38*(6), pp.857-869.

Malkova, L., Bachevalier, J., Webster, M. and Mishkin, M. (2000) Effects of neonatal inferior prefrontal and medial temporal lesions on learning the rule for delayed nonmatching-to-sample. *Developmental neuropsychology*, *18* (3), pp. 399-421.

Malmberg, L.E (2011) Seminar 8: Selecting statistical tests and other parts of the research process, lecture notes, Quantitative design and analysis, University of Oxford, delivered 1 December, 2011.

Maluccio, J. and Flores, R. (2005) Impact evaluation of a conditional cash transfer program: the Nicaraguan red de protección social, Research Report No 141. Washington DC, USA: International Food Policy Research Institute (IFPRI).

Marais, I. (2009) Response dependence and the measurement of change. *Journal of Applied Measurement* 10, pp. 17–29.

Marais, I. and Andrich, D. (2008a) Formalising Dimension and Response Violations of Local Independence in the Unidimensional Rasch model. *Journal of Applied Measurement*, 9(3), pp. 1-16.

Marais, I. and Andrich, D. (2008b). Effects of Varying Magnitude and Patterns of Response Dependence in the Unidimensional Rasch Model. *Journal of Applied Measurement,* 9(2), pp. 1-20.

Masten, A. S., and Gewirtz, A. H. (2006). Vulnerability and resilience in early child development. In K. McCartney and D. Phillips (Eds.) *Blackwell handbook of early childhood development* (pp. 22-43), Oxford: Blackwell Publishing.

Masten, A. S., & Powell, J. L. (2003). A resilience framework for research, policy, and practice. *Resilience and vulnerability: Adaptation in the context of childhood adversities*, 1-25.

Masters, G.N. (1988) Item discrimination: when more is worse. *Journal of Educational Measurement* 25, pp. 15-29.

Matlock-Hetzel, S. (1997) Basic concepts in item and test analysis. Austin: paper presented at the annual meeting of the Southwest Educational Research Association.

Mayes, L., and Lewis, M. (2013) The Cambridge Handbook of Environment in Human Development. Cambridge University Press: Oxford.

McDonald, R. P. (1999) *Test theory: A unified treatment*. Psychology Press: New York.

McGrane, J. A. (2015) Educational growth, learning progressions, and metrological simulacra: Putting the house in (partial) order. *Measurement: Interdisciplinary Research and Perspectives*, *13*(3-4), pp. 200-205.

McGrane, J. A. (2015) Stevens' forgotten crossroads: the divergent measurement traditions in the physical and psychological sciences from the mid-twentieth century. *Frontiers in Psychology*, *6*, pp. 1-8.

McKeown M.G. (1985) The acquisition of word meaning from context by children of high and low ability. *Reading Research Quarterly*, 20, pp. 482–496.

McLaughlin, K. A., Sheridan, M. A., and Nelson, C. A. (2017) Neglect as a violation of species-expectant experience: neurodevelopmental consequences. *Biological Psychiatry*, 82(7), pp. 462-471.

Mehta, P. D. and Neale, M. C. (2005) People are variables too: Multilevel structural equations modeling. *Psychological methods*, 10(3), pp. 259-284.

Melhuish, E, Sylva, K., Siraj-Blatchford, I., Taggart, B. and Phan, M. (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*, 95-114.

Mendez, M.A., Adair, L.S. (1999) Severity and timing of stunting in the first two years of life affect performance on cognitive tests in late childhood. *Journal of Nutrition*,129, pp. 1555–62.

Merton, R. K. (1968) The Matthew effect in science: The reward and communication systems of science are considered. *Science*, 159 (3810), pp. 56-63.

Messick, S. (1986) The once and future issues of validity: assessing the meaning and consequences of measurement, RR-86-30. Princeton, N.J: Educational Testing Service.

Messick, S. (1989) Meaning and values in test validation: the science and ethics of assessment. Educational researcher 18(2), pp.5-11.

Mickelson, R.A. (1989) Why xoes Jane dead and write so well? The anomaly of women's achievement. *Sociology of Education*, 62, pp. 47-63.

Mickelson, R.A. (1990) The attitude-achievement paradox among black adolescents. *Sociology of Education*, 63, pp.44-61.

Michell, J. (2008) Conjoint measurement and the Rasch paradox: A response to Kyngdon. *Theory & Psychology*, *18* (1), pp. 119-124.

Monseur, C., Baye, A., Lafontaine, D. and Quittre, V. (2011) PISA test format assessment and the local independence assumption. In IERI monograph series: Issues and methodologies in large-scale assessments, 4. Hamburg, Germany: IEA/ETS Research Institute (IERI).

Morrow, V. (2009) The Ethics of Social Research with Children and Families in Young Lives: Practical Experiences. Young Lives Working Paper No. 53. Oxford: Young Lives.

Nag, S. (2017) Assessment of Literacy and Foundation Learning in Developing Countries. Oxford: DFID Health and Education Advice Resource Team (HEART).

Nagy, W. (2007) Metalinguistic awareness and the vocabulary-comprehension connection. In R.K. Wagner, A.E. Muse, and K.R. Tannenbaum (Eds.). Vocabulary acquisition: Implications for reading comprehension, (pp. 52-77). New York: Guildford Press.

Nagy, W. E., and Herman, P. A. (1987) Breadth and depth of vocabulary knowledge: Implications for acquisition and instruction. *The nature of vocabulary acquisition*, 19, pp. 35.

National Center on the Developing Child (2007) *The Science of Early Childhood Development* (InBrief). Retrieved from www.developingchild.harvard.edu.

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

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

National Organisation for Examinations (2011) *Ethiopian Second National Learning Assessment of Grade 8 students.* Addis Ababa: National Organisation for Examinations.

Naudeau, S., Martinez, S., Premand, P., and Filmer, D. (2011). Cognitive development among young children in low-income countries. *No Small Matter: The Impact of Poverty, Shocks, and Human Capital Investments in Early Childhood Development*, pp. 9-50.

Nelson, C.A. (2000) The neurobiological bases of early intervention. In J.P. Shonkoff, and S.J. Meisels, Handbook of Early Childhood Intervention (pp. 204-227). Cambridge, MA: Cambridge University Press.

Nelson, C. A. and Bloom, F. E. (1997) Child development and neuroscience. *Child development*, *68* (5), pp. 970-987.

Noble, K. G., Farah, M. J., and McCandliss, B. D. (2006) Socioeconomic background modulates cognition–achievement relationships in reading. *Cognitive Development*, *21* (3), pp. 349-368.

Noble, K. G., Norman, M. F. and Farah, M. J. (2005) Neurocognitive correlates of socioeconomic status in kindergarten children. *Developmental science*, *8* (1), pp. 74-87.

Noble, K.G., McCandliss, B.D., and Farah, M.J. (2007) Socioeconomic gradients predict individual differences in neurocognitive abilities. *Developmental Science*, 10 (4), pp. 464– 480.

Nunnally, J.C. and Bernstein, I.H. (1994) Psychometric theory (3rd ed.). New York: McGraw-Hill.

O’Donnell, O., van Doorslaer, A., Wagstaff, A., and Lindelow, M. (2008) Analyzing health equity using household survey data: a guide to techniques and their implementation. Washington, DC: The World Bank.

Ogbu, J.U. (1986) The Consequences of the American Caste System. In U. Neisser (Ed.) The School Achievement of Minority Children: New Perspectives (pp. 19-56). Lawrence Erlbaum.

Oliver, B. R., and Plomin, R. (2007) Twins' Early Development Study (TEDS): A multivariate, longitudinal genetic investigation of language, cognition and behavior problems from childhood through adolescence. *Twin Research and Human Genetics*, *10* (01), pp. 96-105.

Oliveri, M.E. and Kadriye, E. (2011) Do different approaches to examining construct comparability in Multilanguage assessment lead to similar conclusions? *Applied measurement in education*, 24, pp.349-366.

Organisation for Economic Co-operation and Development (OECD). (2005). PISA 2003 technical report. Paris, France: OECD.

Organisation for Economic Co-operation and Development (OECD). (2012). PISA 2012 technical report. Paris, France: OECD.

Outes-Leon, I. and Dercon, S. (2008) Survey attrition and attrition bias in Young Lives. Young Lives Technical Note No, 5. Oxford, UK: Young Lives.

Outes-Leon, I. and Sanchez, A. (2008) An assessment of the Young Lives sampling approach in Ethiopia: Young Lives Technical Note No. 1. Oxford, UK: Young Lives.

Outes-Leon, I., Porter, C., and Sánchez, A. (2011) *Early Nutrition and Cognition in Peru: A Within-Sibling Investigation* (No. 3430). Inter-American Development Bank.

Overseas Development Institute (2011) *Ethiopia's progress in education: A rapid and equitable expansion of access.* London, UK: ODI.

Pallant, J.F. and Tennant, A. (2006). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS), *British Journal of Clinical Psychology*, 1-18.

Pallant, J.F., Miller, R.L. and Tennant, A. (2006) Evaluation of the Edinburgh Post Natal Depression Scale using Rasch analysis. *BioMed Central Psychiatry,* pp. 6:28.

Panayides, P., Robinson, C. and Timms, P. (2010) The assessment revolution that has passed England by: Rasch measurement, *British Educational Research Journal* 36(4), pp. 1-16.

Paxson, C., & Schady, N. (2007) Cognitive Development among Young Children in Ecuador The Roles of Wealth, Health, and Parenting. *Journal of Human resources*, *42*(1), pp. 49-84.

Phillips, M., Brooks-Gunn, J., Duncan, G.J., Klebanov, P. and Crane, J. (1998) Family Background, Parenting Practices, and the Black-White Test Score Gap. In C. Jencks and M. Phillips (Eds.), The Black-White Test Score Gap (pp.103-145). Washington, DC: Brookings Institution Press.

Phillips, D. A. and Shonkoff, J. P. (Eds.). (2000) *From neurons to neighborhoods: The science of early childhood development*. National Academies Press.

Piachaud, D. (1987) Problems in the Definition and Measurement of Poverty, *Journal of Social Policy*, 16(2), pp. 147-164.

Pianta R.C. and McCoy S.J. (1997) The first day of school: the predictive validity of early school screening. Journal of Applied Developmental Psychology, 18, pp. 1–22.

Pituch, K. A., Stapleton, L. M. and Kang, J. Y. (2006) A comparison of single sample and bootstrap methods to assess mediation in cluster randomized trials. *Multivariate Behavioral Research*, *41*(3), pp. 367-400.

Plomin, Robert. 1986. Development, Genetics, and Psychology. Lawrence Erlbaum.

Plomin, R., DeFries, J. C., Knopik, V. S. and Neiderheiser, J. (2013) *Behavioral genetics*. Palgrave Macmillan.

Plomin, R. and Palfrey-Davis, O. (2006) Gene–environment interactions and correlations in the development of cognitive abilities and disabilities. In J. MacCabe, O. O'Daly, R. Murray, P. McGuffin, and P. Wright (Eds.), *Beyond nature and nurture in psychiatry: Genes, environment and their interplay* (pp. 35-45). CRC Press.

Pollitt, E., Gorman, K. S., Engle, P. L., Martorell, R., Rivera, J., Wachs, T. D., and Scrimshaw, N. S. (1993) Early supplementary feeding and cognition: effects over two decades. *Monographs of the Society for Research in Child Development*, pp. i-118.

Poortinga, Y.H. (1989) Equivalence of cross-cultural data: an overview of basic issues. *International Journal of Psychology*, 24, pp.737-756.

Pradhan, M., Suryahadi, A., Sumarto, S., and Pritchett. L. (2000) Measurements of Poverty in Indonesia, 1996, 1999, and Beyond. Policy Research Working Paper Series No. 2438, World Bank, Washington, DC.

Prais, S. J. (2003) Cautions on OECD's recent educational survey (PISA). *Oxford Review of Education,* 29(2), pp. 139-163.

Preacher, K. J., Zhang, Z. and Zyphur, M. J. (2011) Alternative methods for assessing mediation in multilevel data: the advantages of multilevel SEM. *Structural Equation Modeling*, *18*(2), pp.161-182.

Preacher, K. J., Zyphur, M. J. and Zhang, Z. (2010) A general multilevel SEM framework for assessing multilevel mediation. *Psychological methods*, *15*(3), pp. 209-233.

Puma, M., Bell, S., Cook, R., Heid, C., Shapiro, G., Broene, P., Jenkins, F., Fletcher, P., Quinn, L., Friedman, J. and Ciarico, J. (2010) Head Start Impact Study. Final Report. *Administration for Children & Families*.

Rasch, G. (1980) Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Ravallion, M. (2010) Mashup indices of development. Policy Research Working Paper, 5432. Washington: World Bank.

Ravallion, M. (2011) On multidimensional indices of poverty. *Journal of Economic Inequality*, 9, pp. 235-248.

Reynolds, S. A., Andersen, C., Behrman, J., Singh, A., Stein, A. D., Benny, L., … Fernald, L. C. H. (2017) Disparities in children's vocabulary and height in relation to household wealth and parental schooling: A longitudinal study in four low- and middle- income countries. *SSM - Population Health*, 3(February), pp. 767–786.

Richter, L., Griesel, R. and De Wet, T. (1998) Growing up and going to school: birth to ten, children's cognitive and social competence at school entry. A view of school: preparation for and adjustment to school under rapidly changing social conditions. Johannesburg: The Goethe Institute.

Rippin, N. (2011) A Response to the Weaknesses of the Multidimensional Poverty Index (MPI): The Correlation Sensitive Poverty Index (CSPI). Briefing Paper 19/11, German Development Institute. http://www.die-gdi.de/uploads/media/BP_19.2011.pdf [Accessed 20 July 2015]

Roberts, J., Jergens, J. and Burchinal, M. (2005) The role of home literacy practices in preschool children's language and emergent literacy skills. *Journal of Speech, Language, and Hearing Research*, *48* (2), pp. 345-359.

Rogoff, B. (1995). Observing sociocultural activity on three planes: participatory appropriation, guided participation, and apprenticeship. In J.V. Wertsch, P. del Rio, and A. Alvarez (Eds.), *Sociocultural studies of mind* (pp. 139-164). Cambridge, UK: Cambridge University Press. Reprinted (2008) in K. Hall and P. Murphy (Eds.), *Pedagogy and practice: Culture and identities*. London: Sage.

Rogosa, D. (1988) Myths about longitudinal research. In K. W. Schaie, R. T. Campbell, W. Meredith and S. C. Rawlings (Eds.), Methodological issues in aging research. New York: Springer.

Rowntree, B.S (1901) *Poverty: A Study of Town Life*. London: Macmillan.

Rubio-Codina, M., Araujo, M. C., Attanasio, O., Muñoz, P. and Grantham-McGregor, S. (2016) Concurrent validity and feasibility of short tests currently used to measure early childhood development in large scale studies. *PloS one*, *11*(8), e0160962.

Rubio-Codina, M., Attanasio, O., Meghir, C., Varela, N., and Grantham-McGregor, S. (2015) The socioeconomic gradient of child development: Cross-sectional evidence from children 6-42 months in Bogota. *Journal of Human Resources*, 50(2), 465–483.

Rubio-Codina, M., Attanasio, O., and Grantham-McGregor, S. (2016) Mediating pathways in the socioeconomic gradient of child development: Evidence from children 6–42 months in Bogota. *International journal of behavioural development*, 40(6), pp. 483-491.

Rubio-Codina, M., and Grantham-McGregor, S. (2019) Evolution of the wealth gap in child development and mediating pathways: evidence from a longitudinal study in Bogota, Colombia. *Developmental science*, e12810.

Ruggeri L., Saith, C.R., and Stewart, F. (2003) Does it matter that we don't agree on the definition of poverty? A comparison of four approaches. QEH Working Paper Series 107. University of Oxford, Oxford.

Rutkowski, L., and Rutkowski, D. (2016) A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher*, 45(4), pp. 252-257

Rutstein, S.O. and Johnson, K. (2004) The DHS Wealth Index. DHS Comparative Reports No. 6. Calverton, Maryland: ORC Macro.

Rutter, M., Dunn, J., Plomin, R., Simonoff, E., Pickles, A,. and Maughan, B. (1997) Integrating nature and nurture: Implications of person–environment correlations and interactions for developmental psychopathology Development and Psychopathology, 9, pp. 335–364

Ryan J.P. (1983) Introduction to latent trait analysis and item response theory. In W.E. Hathaway (Ed.), Testing in the Schools. New Directions for Testing and Measurement, 19, San Francisco: Jossey-Bass.

Sammons, P. (2014) Influences on students' GCSE attainment and progress at age 16: Effective Pre-School, Primary & Secondary Education Project (EPPSE). UK Department for Education.

Sánchez, A. and Dercon, S. (2011) Long-term implications of undernutrition on non-cognitive skills: evidence from four developing countries. Young Lives working paper 72. University of Oxford, Oxford.

Schott, W. B., Crookston, B. T., Lundeen, E. A., Stein, A. D., and Behrman, J. R. (2013) Periods of child growth up to age 8 years in Ethiopia, India, Peru and Vietnam: Key distal household and community factors. *Social Science Medicine*, 1982 (97).

Schunk, D. H., Meece, J. R., and Pintrich, P. R. (2014) *Motivation in education: Theory, research, and applications*. Harlow, Essex: Pearson.

Scriven, M.J. (1974) The Logic of the Aptitude - Achievement Distinction. In D.R. Green (Ed.) The Aptitude-Achievement Distinction, Proceedings (pp. 326-34). CTB/ McGraw-Hill.

Scott, J. A., Hoover, M., Flinspach, S. L., and Vevea, J. L. (2008) A multiple-level vocabulary assessment tool: measuring word knowledge based on grade-level materials. In Y. Kim, V. Risko, D. Compton, D. Dickinson, M. Hurdley, R. Jomenez, D. Rove (Eds.), *57th annual yearbook of the National Reading Conference* (pp. 325-340). Oak Creek, WI: National Reading Conference.

Seager, J. and de Wet, T. (2003) Establishing large panel studies in developing countries: the importance of the Young Lives pilot phase, working paper 9, Oxford: Young Lives.

Schady, N., Behrman, J., Araujo, M. C., Azuero, R., Bernal, R., Bravo, D., …Vakis, R. (2015) Wealth gradients in early childhood cognitive development in five Latin American countries. *Journal of Human Resources*, 50(2), pp. 446–463.

Sen, A. (1983) Poor, Relatively Speaking, Oxford Economic Papers, 35.

Sen, A. (1985a) *Commodities and Capabilities*. Amsterdam: North-Holland.

Sen, A. (1985b), Well-being, Agency and Freedom: The Dewey Lectures 1984. *Journal of Philosophy*, 82 (4), pp. 169–221.

Sen, A. (2001) Development as Freedom. Oxford: Oxford University Press.

Shonkoff, J. P. and Phillips, D. A. (Eds.). (2000). *From neurons to neighborhoods: The science of early childhood development*. National Academies Press.

Shonkoff, J. P., Boyce, W. T. and McEwen, B. S. (2009) Neuroscience, Molecular Biology, and the Childhood Roots of Health Disparities: Building a New Framework for Health Promotion and Disease Prevention. *Journal of the American Medical Association*, 301 (21), pp. 2252–59.

Shonkoff, J. P., Garner, A. S., Siegel, B. S., Dobbins, M. I., Earls, M. F., McGuinn, L., Pascoe, J., Wood, D.L. and Committee on Early Childhood, Adoption, and Dependent Care. (2012) The lifelong effects of early childhood adversity and toxic stress. *Pediatrics*, 129(1), pp. e232-e246.

Shonkoff, J.P. and Meisels, S.J. (2000) Handbook of Early Childhood Intervention (2nd ed.) J. P Shonkoff & S. J. Meisels (Eds.), Cambridge, MA: Cambridge University Press.

Siddiqi, A., Kawachi, I., Berkman, L., Subramanian, S.V., and Hertzman, C. (2007) Variation of socioeconomic gradients in children's developmental health across advanced capitalist societies: analysis of 22 OECD nations. *International Journal of Health Services*, 37 (1), pp. 63–87.

Sigman, M., McDonald, M. A., Neumann, C., and Bwibo, N. (1991) Prediction of cognitive competence in Kenyan children from toddler nutrition, family characteristics and abilities. *Journal of Child Psychology and Psychiatry*, *32*(2), pp. 307-320.

Sigman, M., Neumann, C., Jansen, A. A. and Bwibo, N. (1989) Cognitive abilities of Kenyan children in relation to nutrition, family characteristics, and education. *Child Development* 60(6), pp. 1463-1474.

Singer, J.D. and Willet, J. B. (2003) Applied longitudinal data analysis: modeling change and event occurrance. Oxford: Oxford University Press.

Siraj-Blatchford, I. (2010). 'Learning in the home and at school: how working class children 'succeed against the odds'', *British Educational Research Journal*, 36(3): 463-482.

Sireci, S. G. (2011) Evaluating test and survey items for bias across languages and cultures. In D. Matsumoto and F. J. R. van de Vijver. (Eds.), *Cross-cultural research methods in psychology* (pp. 216–243). New York: Cambridge University Press.

Sirin, S.R. (2005) Socioeconomic status and academic achievement: a meta-analytic review of research. *Review of Educational Research* 75(3), pp. 417–53.

Skinner E.A. (1995) *Perceived Control, Motivation, and Coping*. Thousand Oaks, CA: Sage.

Skinner, E. A., Wellborn, J. G., and Connell, J. P. (1990) What it takes to do well in school and whether I've got it: the role of perceived control in children's engagement and school achievement. *Journal of Educational Psychology*, 82, pp. 22–32.

Skinner, E. A., Zimmer-Gembeck, M. J. and Connell, J. P. (1998) Individual differences and the development of perceived control. *Monographs of the Society for Research in Child Development*, 63(2–3), v–220.

Smith, E. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, 3, 205-231.

Smith, E. (2005). Effect of item redundancy on Rasch item and person estimates. *Journal of Applied Measurement*, 6, pp. 147-163.

Smith, J., Brooks-Gunn, J., and Klebanov, P. (1997) Consequences of living in poverty for young children's cognitive and verbal ability and early school achievement. In G. Duncan and J. Brooks-Gunn (Eds.), *Consequences of growing up poor* (pp. 132–189). New York: Russell Sage Foundation.

Snow, C. E. and Kim, Y. S. (2007) Large problem spaces: The challenge of vocabulary for English language learners. In R. K. Wagner, A. E. Muse, and K. R. Tannenbaum (Eds.), *Vocabulary acquisition: Implications for reading comprehension* (pp. 123–139). Guilford Press.

Sperry, D. E., Sperry, L. L., and Miller, P. J. (2019) Reexamining the verbal environments of children from different socioeconomic backgrounds. *Child Development*, *90*(4), pp. 1303-1318.

Stahl, K. A. and Bravo, M. A. (2010). Contemporary classroom vocabulary assessment for content areas. *The Reading Teacher*, 63(7), pp. 566-578.

Stahl, S. A., and Nagy, W. E. (2007) Teaching word meanings. London, UK.: Routledge.

Stahl, S. A., and Fairbanks, M.M. (1986) The effects of vocabulary instruction: A model-based meta-analysis. *Review of educational research*, 56(1), pp. 72-110.

Stanovich, K. E. (1986) Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Journal of education,* 189 (1-2), pp. 23-55.

StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP.

Stein A.D., Behrman J.R., DiGirolamo A., Grajeda R., Martorell R., Quisumbing A., and Ramakrishnan U. (2005) Schooling, educational achievement, and cognitive functioning among young Guatemalan adults. *Food and Nutrition Bulletin*, 26 (Suppl 1), pp. S46–54.

Steinmayr, R., Meibner, A., Weidinger, A. F., and Wirthwein, L. (2014) Academic achievement. In L. H. Meyer (Ed.), *Oxford Bibliographies Online: Education.* New York: Oxford University Press.

Stenner, A. J., Fisher Jr, W. P., Stone, M., and Burdick, D. (2013) Causal Rasch models. *Frontiers in psychology*, *4*, 536.

Stewart, F. (2000) Crisis prevention: tackling horizontal inequalities. *Oxford Development Studies* 28, 245–262.

Stockman, I.J. (2000) The new peabody picture vocabulary test-III: An illusion of unbiased assessment?. *Language, speech, and hearing services in schools*, *31*(4), pp. 340-353.

Strand, S. (2011) The limits of social class in explaining ethnic gaps in educational attainment. *British Educational Research Journal*, 37(2), pp.197–229.

Super, C. M., and Harkness, S. (1997) The cultural structuring of child development. In J. W. Berry, P. Dasen, and T. S. Saraswathi (Eds.), *Handbook of cross-cultural psychology: Volume 2. Basic processes and human development* (pp. 1–39). Boston: Allyn & Bacon.

Super, C. M., and Harkness, S. (1999) The environment as culture in developmental research. In T. Wachs and S. Friedman (Eds.), Measurement of the environment in developmental research (pp. 279–323). Washington, D.C.: American Psychological Association.

Sylva, K. (2014) Students' educational and developmental outcomes at age 16 Effective Pre-school, Primary and Secondary Education (EPPSE 3-16) Project. UK Department for Education.

Sylva, K., Siraj-Blatchford, I. and Taggart, B. (2003) Assessing Quality in the Early Years: Early Childhood Environment Rating Scale Extension (ECERS-E): Four Curricular Subscales. Stoke on Trent: Trentham Books.

Tannenbaum, K. R., Torgesen, J. K., and Wagner, R. K. (2006) Relationships between word knowledge and reading comprehension in third-grade children. *Scientific Studies in Reading*, 10, pp. 381–398.

Teachman, J., Paasch, K.M., Day, R., and Carver, K.P. (1997) Poverty during adolescence and subsequent educational attainment. In G. Duncan and J. Brooks-Gunn (Eds), Consequences of growing up poor (pp. 382-418). New York: Russell Sage Foundation.

Tennant, A. and Conaghan, P. G. (2007) The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis & Rheumatism*, 57, pp. 1358–1362.

Tennant, A., Penta, M., Tesio, L., Grimby, G., Thonnard, J.L., Slade, A., Lawton, G., Simone, A., Carter, J. Lundgren-Nilsson, A., Tripolski, M., Ring, H., Bierieng-Sorensen, F., Marincek, C., Burger, H. and Phillips, S. (2004) Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. *Medical Care*, 42 (Suppl 1), pp. 37–48.

Tesfay, N. (2017) Contextual issues in the assessment of children's learning, briefing note. Oxford: Oxford Policy Management.

Tesfay, N., and Malmberg, L. E. (2014) Horizontal inequalities in children's educational outcomes in Ethiopia. *International Journal of Educational Development*, *39*, pp. 110-120.

Thurstone, L.L. (1931) The measurement of social attitudes. Journal Abnormal and Social Psychology, 26, pp. 249– 69.

Thurstone, L.L. (1946) Note on a reanalysis of Davis' reading tests. *Psychometrika*, 11(3), pp. 185-188.

Thurstone, L.L. and Chave, E. J. Theory of Attitude Measurement. In L.L. Thurstone and E. J. Chave (Eds.), *The measurement of attitude: a psychophysical method and some experiments with a scale for measuring attitude toward the Church* (pp. 1-21). Chicago: University of Chicago.

Townsend, P. (1979) *Poverty in the United Kingdom*. London: Allen Lane and Penguin Books.

Townsend, P. (1987) Deprivation. Journal of Social Policy, 16(2) pp. 125-146.

Townsend, P., Phillimore, P. and Beattie, A. (1988) Health and Deprivation: Inequality and the North. Routledge: London.

Traub, R.E. (1983) A priori considerations in choosing an item response model. In R.K. Hambleton (Ed.) Applications of item response theory (pp. 57-70). Educational Research Institute of British Columbia.

Traub, R.E. and Rowley, G.L. (1991) NCME instructional module on understanding reliability.

Trouton, A., Spinath, F. M., and Plomin, R. (2002) Twins early development study (TEDS): a multivariate, longitudinal genetic investigation of language, cognition and behavior problems in childhood. *Twin Research*, *5* (05), pp. 444-448.

Tucker-Drob, E. M., Briley, D. A., and Harden, K. P. (2013). Genetic and environmental influences on cognition across development and context. *Current directions in psychological science*, *22* (5), pp. 349-355.

Turecki, G., and Meaney, M. J. (2016) Effects of the social environment and stress on glucocorticoid receptor gene methylation: a systematic review. *Biological psychiatry*, 79(2), pp. 87-96.

Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B., and Gottesman, I. I. (2003) Socioeconomic status modifies heritability of IQ in young children. *Psychological science*, *14* (6), pp. 623-628.

Ukrainetz, T.A. and Duncan, D.S. (2000) From Old to New: Examining Score Increases on the Peabody Picture Vocabulary Test–III, *Language, Speech, and Hearing Services in Schools* 31, pp. 336-9.

Ukrainetz, T.A. and Bloomquist, C. (2002) The Criterion Validity of Four Vocabulary Tests Compared with a Language Sample, *Child Language Teaching and Therapy* 18, pp.59-79.

UNICEF (2008) Advantage or Paradox? The challenge for children and young people of growing up urban. New York, NY: UNICEF.

United Nations (1995) Report of the World Summit for Social Development. New York, NY: United Nations.

Van de Vijver, F.J.R. and Hambleton, R.K. (1996) Translating Tests: Some practical guidelines. *European Psychologist* 1(2), pp. 89-99.

Van de Vijver, F. J. R., and Leung, K. (1997). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, and J. Pandey (Eds.), *Handbook of cross-cultural psychology* (2nd ed., vol. 1, pp. 257-300). Boston: Allyn & Bacon.

Van de Vijver, F. and Leung, K. (2011) Equivalence and bias: A review of concepts, models, and data analytic procedures. In D. Matsumoto and F. Van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 17-45). Cambridge University Press, New York, NY.

Van de Vijver, F. J. R., and Poortinga, Y. H. (1997) Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, 13, pp. 29-37.

Van de Vijver, F. and Tanzer, N.K. (2004) Bias and equivalence in cross-cultural assessment: an overview. *Revue européenne de psychologie appliquée* 54, pp.119–135.

Veit-Wilson, J. (1986a), Paradigms of Poverty: A Rehabilitation of B.S. Rowntree. *Journal of Social Policy*. 15(1), pp.69-99.

Veit-Wilson, J. (1986b). Paradigms of poverty: a reply to Peter Townsend and Hugh McLachlan. *Journal of Social Policy* 15(4), pp. 503-507.

Victora, C. G., Barros, F. C., Lima, R. C., Behague, D. P., Gonçalves, H., Horta, B. L., Gigante, D. P., and Vaughan, J. P. (2003) The Pelotas birth cohort study, Rio Grande do Sul, Brazil, 1982-2001.*Cadernos de Saúde Pública*, 19(5), pp. 1241-1256.

Victora, C., Adair, L., Fall, C., Hallal, P. C., Martorell, R., Richter, L., and Sachdev, H. S. (2008) Maternal and child undernutrition: Consequences for adult health and human capital. *The Lancet*, 371(9609), pp. 340–357.

Victora, C., de Onis, M., Hallal, P. C., Blössner, M., and Shrimpton, R. (2010) Worldwide timing of growth faltering: Revisiting implications for interventions. *Pediatrics*, 125(3), e473–e480.

Walford, G., Tucker, E., and Viswanathan, M. (2010) *The SAGE handbook of measurement*. London: Sage Publications.

Walker, S. P., Chang, S. M., Powell, C. A., and Grantham-McGregor, S. M. (2005) Effects of early childhood psychosocial stimulation and nutritional supplementation on cognition and education in growth-stunted Jamaican children: prospective cohort study. *The lancet*, *366* (9499), pp. 1804-1807.

Walker, S.P. and Grantham-McGregor, S.M. (1990) Growth and development of West Indian children. Part 2: development. *West indian medical journal*, *39*(1), pp.12-9.

Walker, S.P., Wachs, T.D., Gardner, J.M., Lozoff, B., Wasserman, G.A., Pollitt, E., and Carter, J.A., 2007. Child development in developing countries 2: risk factors for adverse outcomes in developing countries. *The Lancet*, 369, pp. 145-157.

Werner H, Kaplan E. (1950) Development of word meaning through verbal context: An experimental study. *Journal of Psychology*, 29, pp. 251–257.

White, K.R. (1982) The relation between socioeconomic status and academic achievement. *Psychological Bulletin,* 91(3), pp. 461–81.

White, B. L., Kaban, B.T., and Attanucci, J.S. (1979) The Origins of Human Competence: The Final Report of the Harvard Preschool Project. Lexington Books.

Whiting, B. B. (1980) Culture and social behavior: A model for the development of social behavior. *Journal of the Society for Psychological Anthropology*, 8, pp. 95-115.

WIDER (2007) *World International Inequality Database (WIID2).* Helsinki, Finland: World Institute for Development Economics Research.

Williams, T. R. (1972) Introduction to Socialization: Human Culture Transmitted. St. Louis: Mosby.

Williams, K. T. and Wang, J. J. (1997) *Technical references to the Peabody Picture Vocabulary Test-(PPVT-III)*. American Guidance Service.

Woldehanna, T. (2016) Inequality, preschool education and cognitive development in Ethiopia: Implication for public investment in pre-primary education. *International Journal of Behavioral Development*, *40*(6), pp. 509-516.

Woldehanna, T., and Gebremedhin, L. (2012) The Effects of Pre-school Attendance on the Cognitive Development of Urban Children Aged 5 and 8 Years: Evidence from Ethiopia, Young Lives Working Paper 89.

Woldehanna, T., Gudisa, R., Tafere, Y. and Pankhurst, A. (2011) *Understanding changes in the lives of poor children: initial findings from Ethiopia round 3 survey*. Oxford, UK: Young Lives.

Woldehanna, T., Mekonnen., A., and Alemu, T. (2008) Young Lives: Ethiopia Round 2 Survey, Country Report, Young Lives: Oxford.

World Bank Group (2020) Early childhood development https://www.worldbank.org/en/topic/earlychildhooddevelopment [Accessed on 09 March 2020]

World Bank Group (2018) Poverty and shared prosperity: Piecing together the poverty puzzle. Washington D.C.: World Bank Group.

World Bank Group (2015) Ethiopia Poverty Assessment 2014. Washington, DC.: World Bank.

World Health Organisation (1986) Use and interpretation of anthropometric indicators of nutritional status. *Bulletin of the World Health Organization*, *64*(6), pp. 929–941.

World Health Organisation (1995) Physical status: the use and interpretation of anthropometry. *WHO technical report series*, 854 (9). Geneva, Switzerland: WHO.

World Health Organization (1997) WHO global database on child growth and malnutrition (No. WHO/NUT/97.4). Geneva: World Health Organization.

World Health Organization (2010). Indicators for assessing infant and young child feeding practices: part 2: measurement. Washington D.C.: WHO.

Wright BD (1996) Time 1 to time 2 comparison. Rasch Measurement Transactions 10: 478–479. http://www.rasch.org/rmt/rmt101f.htm

Yach, D., Cameron, N., Padayachee, L., Wagstaff, L., Richter, L., and Fonn, S. (1991) Birth to ten: Child health in South Africa in the 1990s. Rationale and methods of a birth cohort study. *Paediatric Perinatal Epidemiology*, 5, pp. 11–33.

Young Lives (2002): Justification of the content of the household questionnaires. Oxford: UK: Young Lives.

Young Lives (2009). Ethiopia fieldworker manual child questionnaire. Oxford, UK: Young Lives.

Young Lives (2014a). Qualitative Longitudinal Research. Available from: http://www.younglives.org.uk/what-we-do/longitudinal-qualitative-research [Accessed 14 April 2015]

Young Lives (2014b). School Survey. Available from: http://www.younglives.org.uk/what-we-do/school-survey [Accessed 14 April 2015]

Zhang, Z., Zyphur, M. J. and Preacher, K. J. (2009) Testing multilevel mediation using hierarchical linear models: Problems and solutions. *Organizational Research Methods*, *12*(4), pp. 695-719.

Zenisky, A. L., Hambleton, R. K. and Sireci, S. G. (2002) Identification and evaluation of local item dependencies in the medical college admissions test. *Journal of Educational Measurement*, 39, pp. 291-309.

Zimmerman, B. J. (1990) Self-regulated learning and academic achievement: An overview. *Educational psychologist*, *25*(1), pp. 3-17.

Appendix 1. Excerpt of the multidisciplinary literature mapping

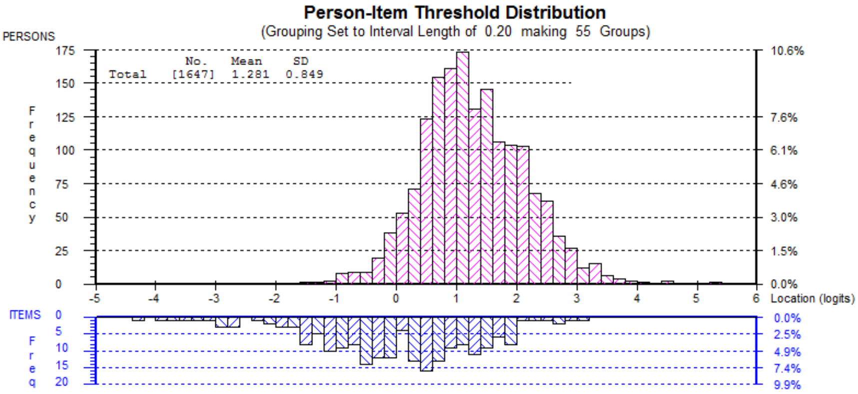
STUDY	OBJECTIVES / RESEARCH DESIGN	NOTES
Strategic thrust: education for development in the post MDG/EFA era		
EFA Global monitoring report 2013	Teaching and learning for human development and poverty reduction	In the post-MDG era, education will not only be about getting children into school, but about how children are actually developing throughout childhood, and what they are actually learning
World Developme nt Report 2013	Building and measuring non-cognitive skill development and the factors that shape developmental processes	In the post-MDG era, emphasis will be placed on building skills for development, jobs and poverty reduction, including what economists have termed "soft skills" or "non-cognitive skills" (problem-solving ability, self-esteem, optimism, agency…)
EFA Global monitoring report 2012	Building and measuring skill development across a range of domains (foundational, transferrable and vocational/technical)	Emphasis on skills for jobs and participation in society
World Bank, No Small Matter 2012	Importance of early childhood development	Emphasis on preventing and offsetting lost developmental potential
Save the children education in Africa strategy	Measuring schooling participation beyond enrolment (progression) Measuring literacy and numeracy (what children are actually learning) Measuring the effect of poor early cognitive and socio-emotional development on later schooling;	Strategic message: Ensuring that all girls and boys are achieving good learning outcomes by the age of 12 with gaps between the poorest and the richest significantly reduced
Brookings Institute – global compact on learning strategic document	Measuring cognitive development, literacy and numeracy (what children are actually learning), and the factors that shape developmental processes (child, family, neighborhood and school); Measuring the effect of poor early cognitive and socio-emotional development on later school progress and development;	ECCE Learning and metrics in primary school Post primary opportunities Also see global compact on learning 'learning metrics taskforce': http://www.globalcompactonlearning.org/global-learning-metrics/ The task force will propose global learning competencies and measures at the early childhood, primary, and post-primary levels. (Note that for this exercise, "learning" is not just literacy and numeracy but is conceived more broadly.)
Learning Metrics Taskforce (LMTF)	Proposes competencies of measured learning outcomes throughout childhood	Final document to be released in September 2013 Proposed Competencies for Learning Outcomes: Early Childhood, Primary, and Post-Primary:
ODI Post 2015 education MDGs	Presents current debates surrounding learning competencies that children should acquire at achool	http://www.odi.org.uk/sites/odi.org.uk/files/odi-assets/publications-opinion-files/7776.pdf p.4 Positive developments towards EFA goals and section 5 p.5 Challenges towards meeting EFA goals – universal enrolment, conflict areas, completion,

STUDY	OBJECTIVES / RESEARCH DESIGN	NOTES
		adult literacy, gender, inequality, quality, teacher training p.8 current debate around learning – (i) mastering the curricula and (ii) statistical measures of achievement and progress (inadequacy of enrolment and completion measures of learning)
Education for All Blog - SCUK	Latest debates are presented here	http://www.educationforallblog.org/education-results/would-you-send-your-child-to-school-if-they-had-a-50-chance-of-learning-nothing Also see: http://www.globaleducationfirst.org/
Normative frameworks, theories/models of the role of education		
Human capital theory		
World Development Report, 2013	Emphasis on the cumulative nature of human capital formation and the importance of childhood (and early childhood) for skill development in the longer term	Human capital, the most conventional theory of education, argues that the role of an education system is to provide individual with the skills and abilities required to become productive members of a workforce and in turn, contribute to the generation of economic growth (Becker, 19XX)
Becker, etc.	The role of education within a human capital framework	Historical/theoretical foundations
Robeyns, 2011	The role of education within a human capital framework	Comparative analysis of the theoretical model of human capital and the normative human capabilities and human rights frameworks
Jeffreys, 2008		*Degrees of Freedom* A critique of human capital theory
Educational opportunity theory and human capabilities		
Sen and Dreze, 1995; Sen, 1999	The role of education within a human capabilities framework (educational opportunity theory)	Historical/normative foundations
Robeyns, 2011	The role of education within a human capabilities framework	Comparative analysis of the theoretical model of human capital and the normative human capabilities and human rights frameworks
McCowan, 2011	The role of education within a human capabilities and human rights framework	
International human rights framework		
United Nations	Universal declaration of human rights, Convention on the Rights of the Child, Education for All	Historical/normative foundations
Robeyns, 2011	The role of education within a human capabilities framework	Comparative analysis of the theoretical model of human capital and the normative human capabilities and human rights frameworks
McCowan, 2011	The role of education within a human capabilities and human rights framework	

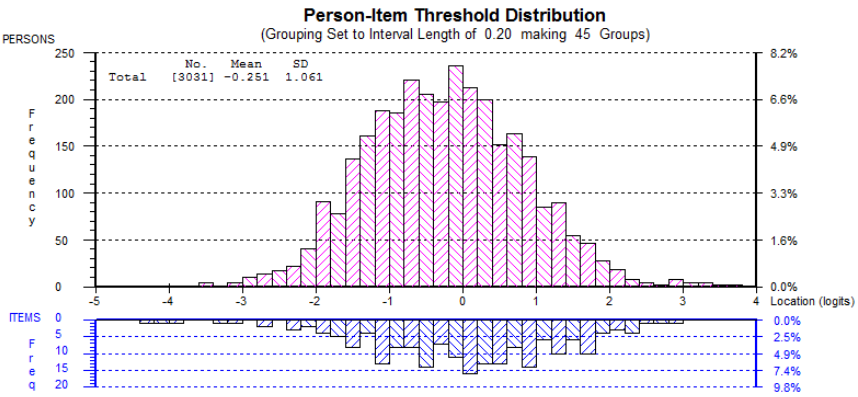
Appendix 2. Summary analysis of the full scale of 204 items

Although the findings of the psychometric analysis will not contribute to the refinement of the Young Lives PPVT scale, it was decided that the responses to the full scale of 204 items should also be analysed. The full scale is available for the younger cohort data at ages five and eight years, and at 12 and 15 years for the older cohort children. For the third administration, the test was not given to the older cohort who were 19-years old at the time, and only a subset of 55 items were retained for the younger cohort who were aged 12 at this point. This section provides a short summary of the findings.

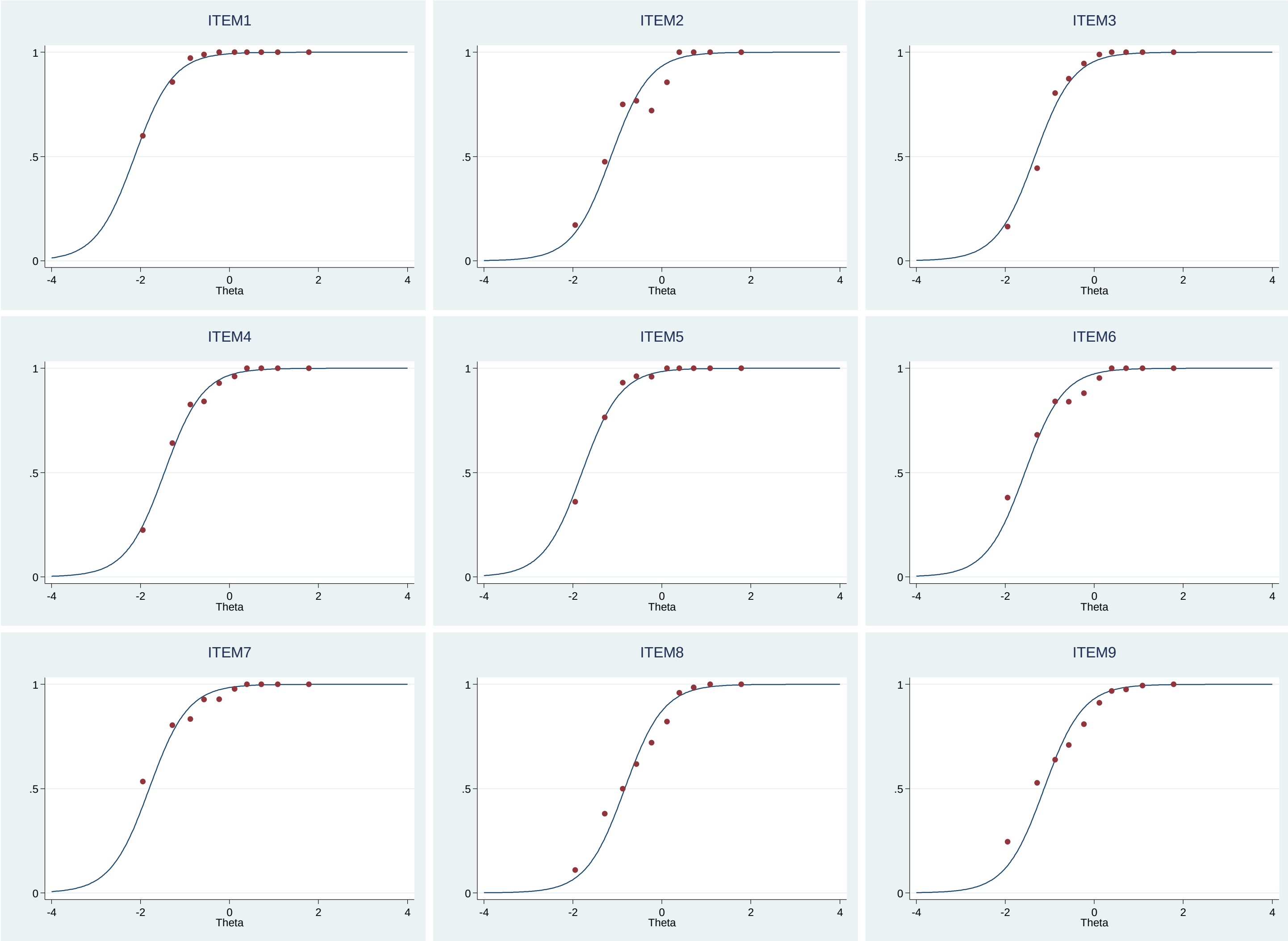
Test performance among the older cohort children showed good overall data-model fit. However, by age 15 a substantial ceiling effect was observed. The distribution of persons relative to items for the pooled sample is shown below. The test is too easy for the older cohort children. This raises questions about the construct validity of the instrument, as the PPVT is intended to be used from the age of two onward. Moreover, a large proportion of later items were located towards the start of the continuum, suggesting that the rank ordering of items was not consistent with the test order (not shown here). The most important finding relates to the presence of response dependency, which is widespread. The finding of local dependence is an important reliability and validity issue in both the Oromifa and Tigrinya versions of the test, with a large proportion of items showing very high residual correlations above 0.70. There are ways of handling deviating items, including removing them. However, due to the extent of dependency eliminating items without substantively understanding the reason for their misfit would further undermine scale validity (Andrich, 2004).

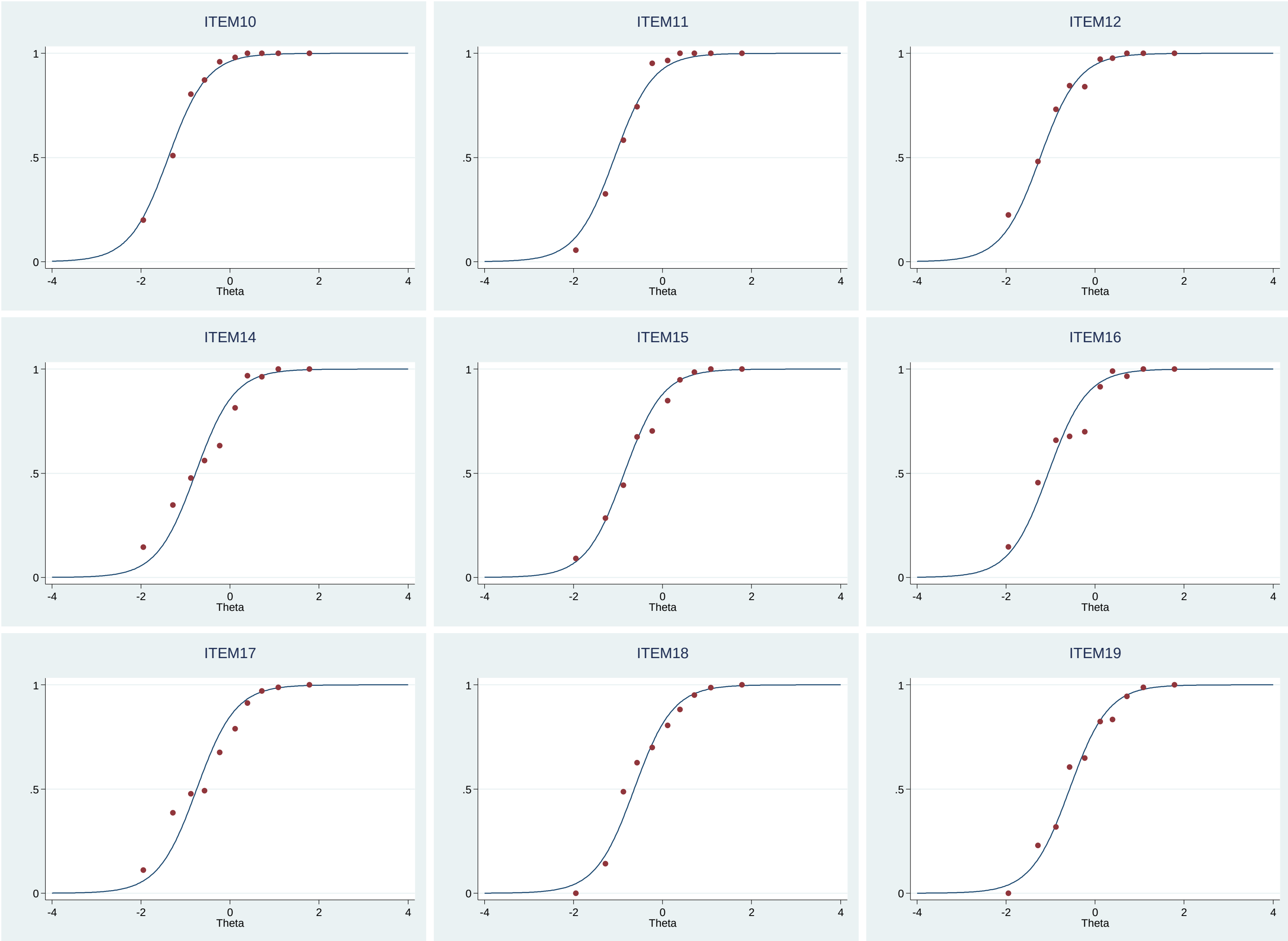


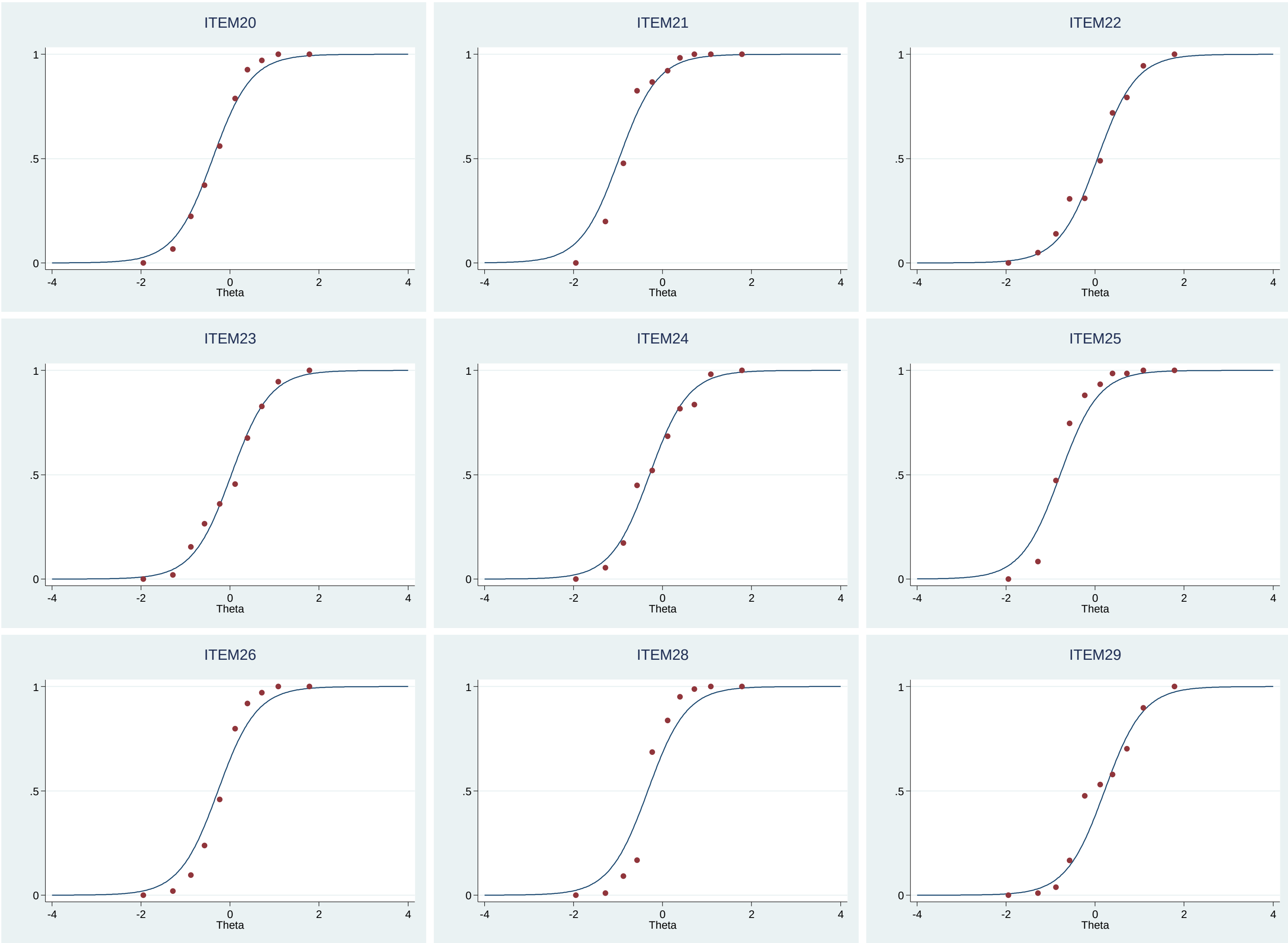
The test worked much better for the younger cohort, in terms of overall fit and item targeting, as shown below. Further, the difficulty ordering of the items is more consistent with the test design (not shown here). Again, the assumption of local independence does not hold. The problem is most evident among the Oromifa and Tigrinya item responses. Further investigation of the raw data would be needed to understand the pattern of dependence, its effect on parameter estimattion and subsequent results interpretation. Alternatively, the high level of dependency may warrant the removal of several items, to improve the validity of the scale, noting that item elimination would need to be justified statistically and based on substantive knowledge of the construct.

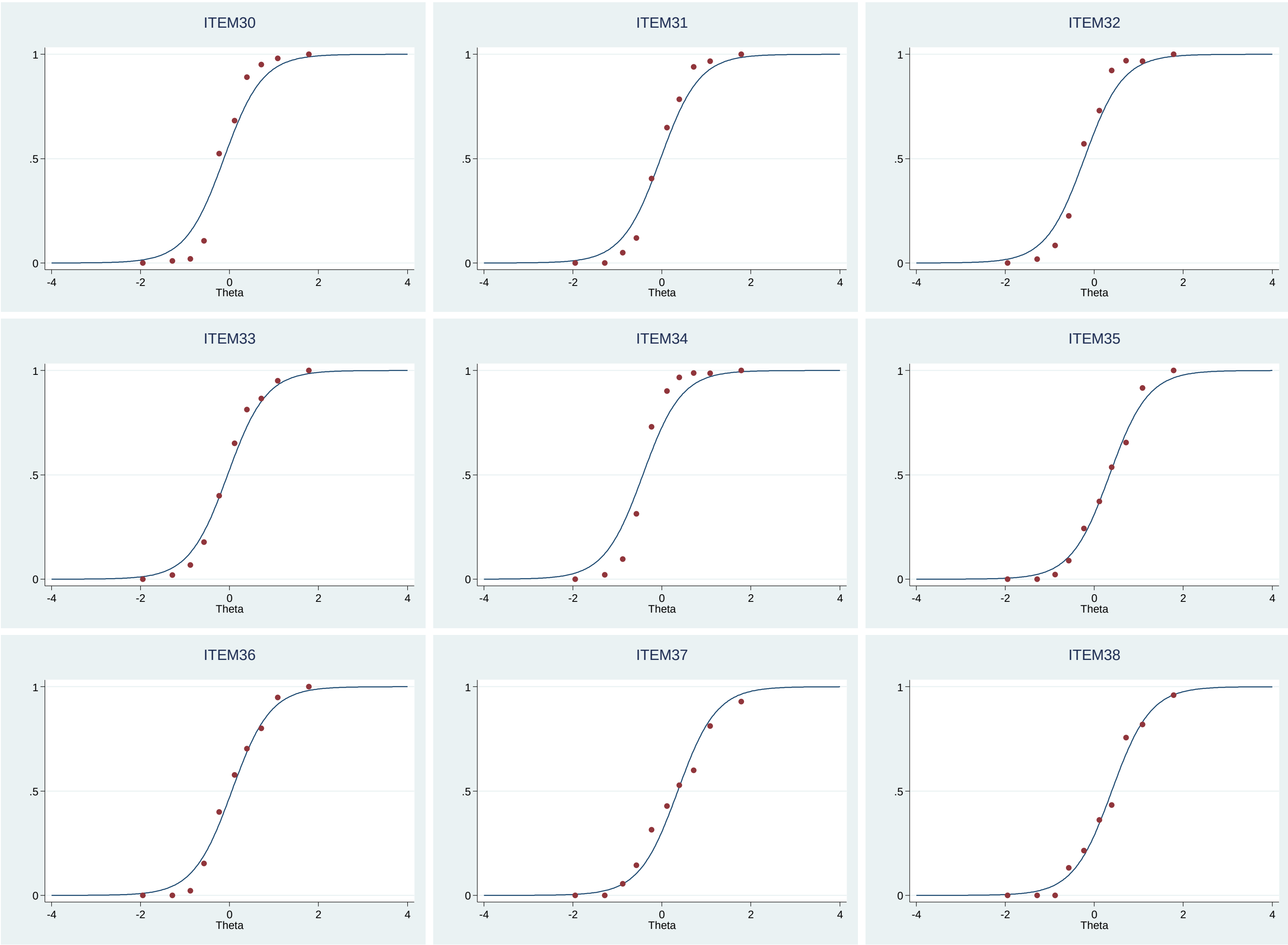


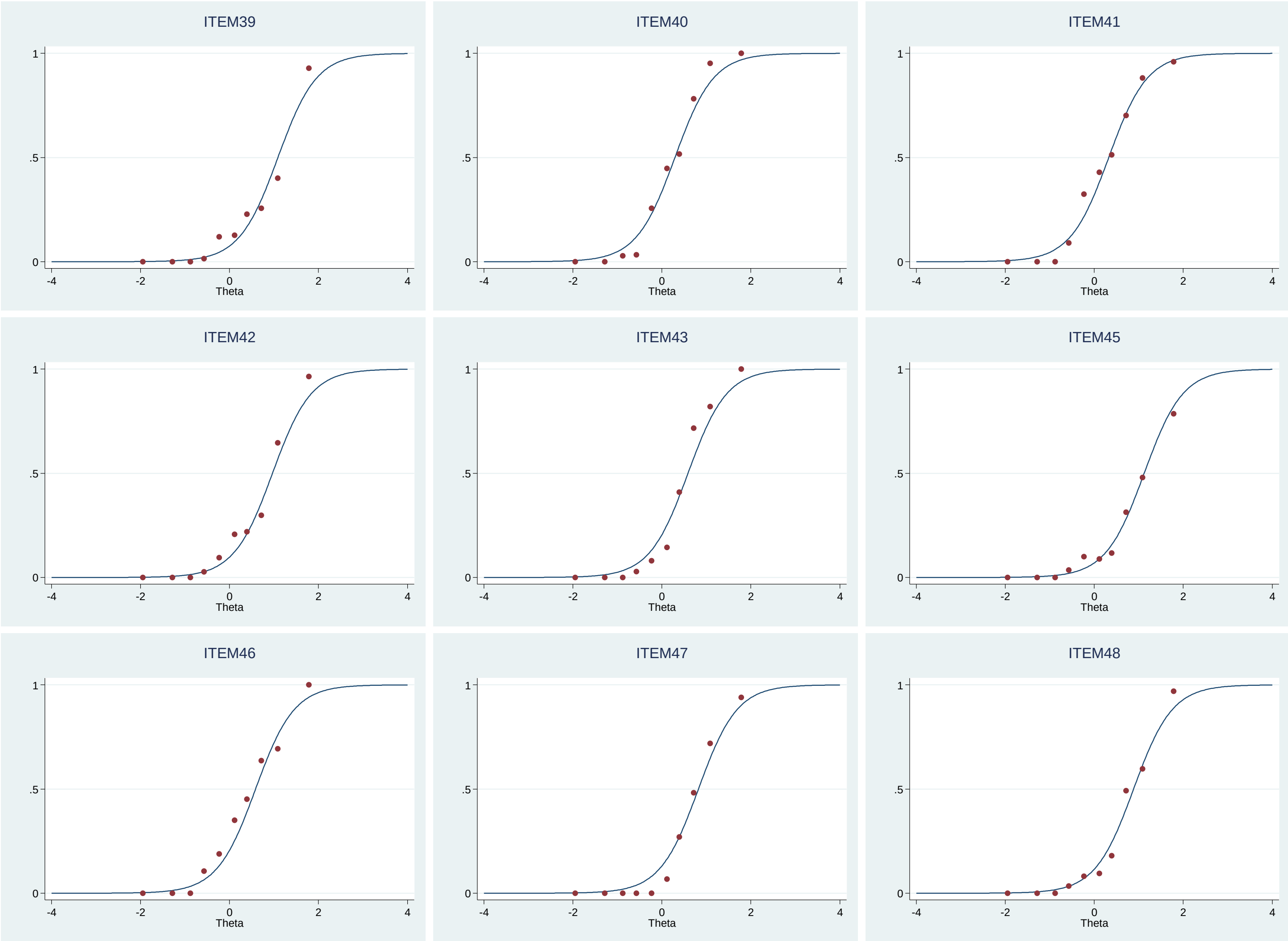
Appendix 3. Item Characteristic Curves for the PPVT (Amharic)

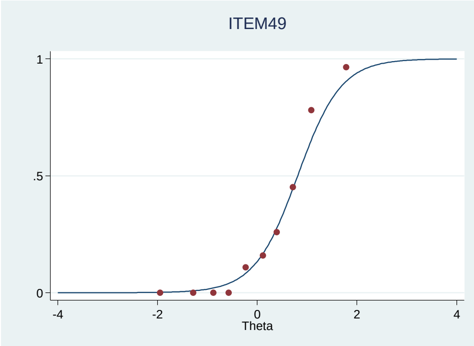


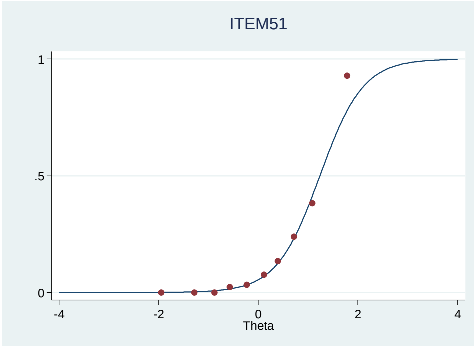

ITEM10
ITEM11
ITEM12
ITEM14
ITEM15
ITEM16
ITEM17
ITEM18
ITEM19
Theta
1
.5
0
-4
-2
0
2
4

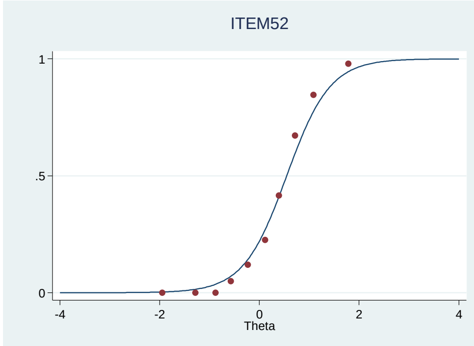

ITEM20
ITEM21
ITEM22
ITEM23
ITEM24
ITEM25
ITEM26
ITEM28
ITEM29
1
.5
0
-4
-2
0
2
4
Theta

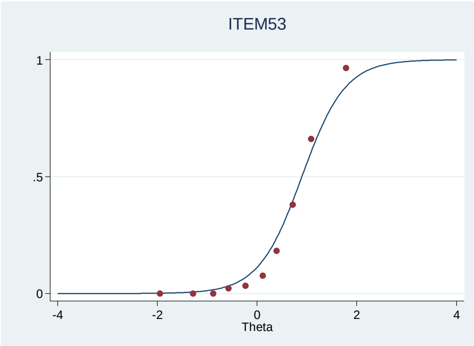

ITEM30
ITEM31
ITEM32
ITEM33
ITEM34
ITEM35
ITEM36
ITEM37
ITEM38
Theta


ITEM39
ITEM40
ITEM41
ITEM42
ITEM43
ITEM45
ITEM46
ITEM47
ITEM48
Theta
0
.5
1
-4
-2
0
2
4


ITEM49
1
.5
0
-4
-2
0
2
4
Theta


ITEM51
1
.5
0
-4
-2
0
2
4
Theta


ITEM52
1
.5
0
-4
-2
0
2
4
Theta


ITEM53
1
.5
0
-4
-2
0
2
4
Theta