Early Childhood Research Quarterly 41 (2017) 21–36

Contents lists available at ScienceDirect

Early Childhood Research Quarterly

journal homepage: www.elsevier.com/locate/ecresq

Measuring school readiness globally: Assessing the construct validity and measurement invariance of the International Development and Early Learning Assessment (IDELA) in Ethiopia

Sharon Wolf[a,b,*], Peter Halpin[b], Hirokazu Yoshikawa[b], Amy Jo Dowd[c], Lauren Pisani[c], Ivelina Borisova[d]

[a] Graduate School of Education, University of Pennsylvania, 3700 Walnut Street, Philadelphia, PA 19104, USA
[b] Global TIES for Children, New York University, USA
[c] Save the Children, USA
[d] UNICEF, USA

ARTICLE INFO

Keywords:
School readiness
Factor analysis
Measurement invariance
Sub-Saharan Africa
Ethiopia

ABSTRACT

The post 2015 context for international development has led to a demand for assessments that measure multiple dimensions of children's school readiness and are feasibly administered in low-resource settings. The present study assesses the construct validity of the International Development and Early Learning Assessment (IDELA) developed by Save the Children using data from a sample of children (~5 years of age; $N = 682$) from rural Ethiopia. The study (a) uses exploratory and confirmatory bi-factor analyses to assess the internal structure of the assessment with respect to four hypothesized domains of school-readiness (Early Numeracy, Early Literacy, Social-Emotional development, and Motor development); (b) uses latent regression to examine concurrent validity of the domains against a limited set of child and family characteristics; and (c) establishes measurement invariance across three focal comparisons (children enrolled in center-based care versus home-based care; girls versus boys; and treatment status in a cluster randomized controlled trial of a center-based program). The results support the conclusion that the IDELA is useful for making inferences about children's school readiness. Implications for future use of the IDELA and similar instruments are discussed.

1. Introduction

Early learning skills, or "school readiness", are crucial for children's transition and adaptation to school (e.g., Blair & Razza, 2007; Cueto et al., 2016; McClelland, Morrison, & Holmes, 2000). In light of growing evidence, governments worldwide have acknowledged that the skills young children bring to the start of school are a major national issue. In particular, the ratification of the 2015–2030 Sustainable Development Goals (SDGs; United Nations, 2015) has signified an increased commitment on the part of governments to improve young children's skills and knowledge to increase their success in early primary grades and beyond. The research base and policy context of current efforts in international development have brought to the forefront the importance of developing assessments of early childhood development (ECD) and school readiness that are feasible to administer, conceptually and psychometrically validated across contexts, and aligned with national monitoring systems (e.g., Bartlett, Dowd, & Jonason, 2015; Yoshikawa & ECDAN Data Task Force, 2017).

School readiness and its measurement have received quite a bit of attention in the developmental and educational literature in high-income countries, but less in low- and middle-income countries (LMICs). This is partly due to a lack of validated measures and available data. Direct assessments that capture multiple domains of ECD and school readiness skills, are feasible to administer, and can be used and compared across regions, countries, and contexts are needed (Chavan & Yoshikawa, 2013). In addition, assessments that are sensitive enough to be used for program evaluation could further the efforts of organizations and governments aiming to improve early childhood outcomes by allowing them to assess and compare program impacts.

Such assessments may be particularly important in informing intervention efforts in Sub-Saharan Africa, where regional estimates indicate that the largest numbers of young children are not reaching their developmental potential due to stunting and living in poverty (Black et al., 2017). As a region, Sub-Saharan Africa has the largest number and proportion of 3- and 4-year old children (29.4 million, or 44%) compared to any other region with low performance in terms of

* Corresponding author at: Graduate School of Education, University of Pennsylvania, 3700 Walnut Street, Philadelphia, PA 19104, USA.
E-mail address: wolfs@upenn.edu (S. Wolf).

http://dx.doi.org/10.1016/j.ecresq.2017.05.001
Received 22 June 2016; Received in revised form 5 April 2017; Accepted 8 May 2017
Available online 06 June 2017

cognitive and socioemotional development (McCoy, Peet, et al., 2016). Multi-faceted, easily administered measures would allow governments to have a more nuanced understanding of if and how their investments are translating into improved child development.

1.1. Assessing school readiness requires a multi-dimensional view of learning and development

School readiness can be defined broadly as an outcome of the early years that covers multiple dimensions of development, including early academic and behavioral skills, Social-Emotional development, and aspects of physical health including Motor development (Snow & Van Hemel, 2008; UNESCO, 2013). While most existing measures are multi-domain in nature, there are surprisingly few published empirical analyses investigating a multi-domain structure of ECD and school readiness. One study assessed the measurement properties of a composite measure of the Bracken School Readiness Assessment in a U.S. sample of kindergarten children by examining the measure's predictive and discriminant validity (Panter & Bracken, 2009). However, competing potential structures of a single overall factor vs. factors representing each of the sub-domains (A, B, C, and D of the Bracken) were not tested. The Early Development Instrument (EDI) was analyzed on a sample of Canadian children using exploratory factor analytic methods, and found to measure six distinct dimensions of development including physical health and well-being, social competence, emotional maturity, language and cognitive development, communication skills and general knowledge domains (Janus & Offord, 2007). Such studies are rare in general and even less common in LMICs. One effort to assess a contextually relevant scale, the East Asia Pacific–Early Child Development Scales (Rao et al., 2014), stands out. By assessing content validity, internal reliability, and item discrimination in six Asian countries, the results of this study revealed seven developmental domains relevant to ECD in East Asia and the Pacific, including: approaches to learning, cognitive development, cultural knowledge and participation, language and emergent literacy, Motor development, health, hygiene, and safety, and Social-Emotional development.

Domains of school readiness are also comprised of various essential skills. For example, meta-analyses have shown that Early Literacy is comprised of five essential components that are necessary but not sufficient for oral reading fluency, including oral language, vocabulary, phonemic awareness, print awareness, and letter knowledge (Snow, 2006). These constructs are often examined as a composite representing Early Literacy (e.g., Hood, Conlon, & Andrews, 2008), though some studies have examined skills individually (e.g., Sénéchal & LeFevre, 2002). Similarly, Early Numeracy consists of components including number concepts and quantities, number relationship and operations, geometry and spatial sense, patterns, and measurement and comparison (Office of Head Start, 2010). These are commonly combined to represent children's Early Numeracy ability (e.g., Anders et al., 2012).

The domain of Social-Emotional development has also been differentiated into individual skill areas, including recognizing and managing emotions, appreciating the perspectives of others, establishing positive goals, making responsible decisions, and handling interpersonal situations effectively (CASEL, 2012). Finally, motor skills are often considered in two categories – gross motor and fine motor, with the latter shown to be more important in the transition to school (i.e., Cameron et al., 2012). Most assessments have not considered measurement implications of the hypothesized sub-construct structures within broader developmental domains.

1.2. The International Development and Early Learning Assessment (IDELA)

The IDELA was developed in 2011 by Save the Children as a holistic, rigorous, open source assessment that is feasible and easily adapted to different national and cultural contexts. The initial set of items were inspired by and conceptually adapted from existing assessments such as the Denver, the Ages and Stages Questionnaire, the Bayley Scales of Child Development, and the Early Development Instrument, among other assessments. The version of the assessment utilized in this research consisted of 101 items administered through 24 subtasks designed to measure a total of 4 domains of child development: Emergent Numeracy, Emergent Literacy, Gross and Fine Motor Skills, and Social-Emotional Learning. *Items* refer to the individual responses, usually scored as correct/incorrect or yes/no (e.g., “Can you tell me how old you are?” “Can you show me the smallest circle?”). *Subtasks* refer to groupings of one or more items based on similar stimulus materials or content (e.g., the total number of letters a child can identify) and represent skills within each domain. A description of the subtasks and items is included in Section 2.3. The full assessment is available from Save the Children upon request. The adaptation of the tool to different countries includes sourcing locally appropriate materials, such as picture cards, small items for counting, and children's books, and including local educational staff to inform the content of some of the subtasks.

The intended purpose of the IDELA is to support program improvement across Save the Children's and partners’ numerous country sites, to increase accountability among ECD initiatives globally with direct assessments of ECD outcomes, and to offer comparable data about children's learning and development across countries and programs that can help bring successful ECD programs to scale. IDELA data is intended for use at aggregate levels (i.e., for impact evaluation comparisons of groups; potentially for national monitoring) and not for screening individual children for developmental delays (see ECD Action Network, 2017). The development of the assessment is described in a technical working paper (Pisani, Borisova, & Dowd, 2015).

1.3. Ethiopian context and early childhood care and education policy

We examine the IDELA measure on a sample of children from Ethiopia, a country in the horn of Africa with a population of close to 100 million. As a region, Sub-Saharan Africa has the largest proportion of young children not meeting their developmental potential (Black et al., 2017) nor basic developmental milestones (McCoy, Peet, et al., 2016). On the Human Development Index, a composite statistic of life expectancy, education, and income per capita indicators used to rank countries by overall human development and conducted by the UNDP, Ethiopia ranks in the bottom tier, number 174 out of 188 countries (UNDP, 2015). While country- and region-level estimates of the state of early childhood development in Ethiopia do not currently exist, the Human Development Index is strongly correlated to country-level estimates of early childhood development, both in terms of stunting rates ($r = .72$) and cognitive and Social-Emotional development ($r = -.84$) (McCoy, Peet, et al., 2016). The Oromia region in Ethiopia, where the data for this study come from, is one of the nine ethnically based regional states of Ethiopia and is the largest state in population and area. The region includes the nation's capital, Addis Ababa, but just 10.5% of its population live in urbanized areas (Ethiopian Government, 2016). In rural areas in Oromia, from which the sample in the present study are drawn, agriculture is the main source of livelihood and the majority of the households are poor. In 2009, 55% of families in Oromia reported falling below the consumption poverty line. This estimate is lower than the 62–79% in other rural regions in the country (Tafere & Woldehanna, 2012).

Ethiopian children officially enter grade 1 at age seven. Despite considerable progress in primary school enrollment, nearly one-quarter of 8-year olds are not enrolled in school in Ethiopia (Woldehanna, Tafere, Pankhurst, & Gudisa, 2011). Gaps in enrollment are apparent based on family and child characteristics. Specifically, children from non-poor families (measured by household assets and consumption poverty), and children with more educated parents have higher school enrollment rates than their counterparts (Woldehanna et al., 2011). Despite gender differences in enrollment, with males being more likely to be enrolled in primary school, the net attendance rate in primary school is not different for boys and girls (UNICEF, 2013). It is important to understand whether similar gaps are present in key early childhood outcomes, and if direct assessments are sensitive enough to identify such gaps. If gaps were similar in early childhood, we would expect parents' education level and household socioeconomic status to positively predict ECD outcomes. In line with research on formal early childhood education (Yoshikawa et al., 2013), and on children's developmental growth, we would expect children's enrollment in early childhood care and education (ECCE) and age to positively predict ECD outcomes. Furthermore, given lower enrollment rates of girls, we may also expect that girls would have lower enrollment in ECCE and thus lower ECD outcomes. However, given the lower likelihood of families relying on young children to work, it is possible that gender gaps in enrollment and outcomes may not yet be present in early childhood.

As part of its strategy to improve educational outcomes for all children, the government has begun investing more attention in early childhood care and education and developed a national Early Childhood Care and Education strategy in 2010. Increasing access to and equity in ECCE service provision and improving ECCE service quality were central in this policy. This policy has four basic areas for the delivery of services: parental education; health and early stimulation (pre-natal to 3 years); pre-schools/kindergartens (4–6 years); and community-based non-formal school readiness programs. Currently, Ethiopia is implementing ECCE in all schools in its Education Sector Development, which has led to an increase in the gross enrollment rate of pre-school children from 5.3% in 2010–11 to 34.0% in the 2013–2014 academic year (Ethiopian Ministry of Education, 2012). Though this government program is very ambitious, pre-school education is marred by many challenges, such as the lack of trained and independent facilitators/teachers; the unavailability of curriculum and guidelines; a lack of adequate center facilities, developmentally appropriate learning materials, and playgrounds; and lack of incentives/salary for teachers assigned for this program, among others (Dowd, Borisova, Amente, & Yenew, 2016). And children living in rural areas are far less likely to have access to formal ECCE compared to their urban counterparts (Woldehanna et al., 2011).

1.4. The current study

The present study examines the measurement properties of the IDELA in a sample of Ethiopian children (age ~5 years). We examine the conceptual and empirical structure of the IDELA, as well as it's utility for making comparisons across subgroups. A previous technical report addressed the internal consistency reliability of domains, inter-rater reliability of the tool, test–retest reliability, and construct validity of subscales of the IDELA as well as the overall instrument (Pisani et al., 2015). This study contributes to this research by providing a more detailed investigation of the internal structure of the IDELA in Ethiopia. The research questions addressed are as follows:

1. Are the items of the IDELA that are currently used to measure each of the domains – Early Literacy, Early Numeracy, Social-Emotional, and Motor – consistent with the hypothesis of a single domain-level construct or factor? This question addresses whether the individual domains measured by IDELA are in fact each unidimensional.
2. Are items on the same subtask related to one another, after controlling for the domain-level factor? If so, this would suggest the need for a "testlet" or bi-factor model that accounts for the subtask structure of the IDELA. (See e.g., Rijmen, 2010, for discussion of the relationships among testlet, bi-factor, and hierarchical models.)
3. Do any items measure an IDELA domain other than their intended domain? This addresses the specificity of the items as indicators of their target construct.
4. How are the IDELA domains related to one another? In particular, is there evidence for four distinct domains? If there are four distinct domains, are the correlations among the domains compatible with the hypothesis of a single, hierarchical factor corresponding to school-readiness, or are the domains related to one another in some other way?
5. Do key child and family variables, known to predict educational outcomes for primary school children in Ethiopia and early childhood development in LMICs, predict children's IDELA scores? This provides an initial, albeit limited, look at the concurrent validity of the assessment.
6. Finally, is the factor structure of the IDELA invariant with respect to (a) pre-existing subgroups of interest to international development researchers (i.e., enrollment in formal early childhood care and gender) and (b) experimentally induced subgroups used for program evaluation and impact analysis? Measurement invariance is an important step to establishing whether the IDELA can be used to make meaningful comparisons among groups of interest. In the case of program evaluation, we are additionally concerned with the sensitivity of the IDELA to potential treatment effects (e.g., effect sizes on the domains).

2. Method

2.1. Participants

In 2012, Save the Children began piloting an intervention in Ethiopia aimed at supporting critical Emergent Literacy and Maths (ELM) skills in its preschool programs. The goal was to strengthen Early Childhood Care and Education in the regions of Tigray and Oromia. There are two components of the ELM program: The center-based program (ELM Center) focuses on improving instruction in existing classroom-based programs, and the home-based program (ELM Home) provides guidance and support to parents about how to improve stimulation and early learning. The sample for this study comes from the baseline and endline data collected as part of an intervention evaluation of the ELM program in 4 districts of West Showa, Oromia. At the baseline wave, conducted in November 2014, the child assessment and caregiver survey covered 682 children and the same number of parents from 36 villages. Nine villages received the ELM Center intervention, 9 villages received both the ELM Center and Home interventions, 9 villages received the ELM Home program, and 9 villages received the government supported "O" class program (i.e., pre-primary education for children who do not have access to a kindergarten, taught by teachers from their local primary schools). For the purposes of this analysis the children receiving the ELM Center intervention and the ELM Center and Home interventions are combined into one group as they all represent children receiving a quality center-based program (now referred to as the ELM Center group).

At baseline, children's average age was 5.9 years (SD = .40, range 4–7), 52% were female, and about three-quarters of the sample (N = 519; 76.1%) was enrolled in a center-based early childhood education center (170 in the government O class group, 349 in ELM Center group). At endline (N = 625), children's average age was 6.2 years (SD = .40, range 5–8), 52% female, and three-quarters of the sample (N = 465; 74.4%) were in one of the three experimental conditions described above. Notably, nearly all (91.6%) of the sample was retained between baseline and endline. Table 1 for descriptive statistics on the sample.

2.2. Procedures

In order to collect child and caregiver data, 18 people from the local community were hired to visit all 36 villages over the course of three weeks. The data collectors were trained on the IDELA child tool and caregiver survey for three days. The training was facilitated by the Save the Children team from the West Showa Field Office. The data collectors were trained on the assessments for three days, practiced

Table 1
Demographic statistics for the sample at baseline.

	M or %	SD	Correlation coefficient					
			1	2	3	4	5	6
Child characteristics								
1. Female	52.6		–					
2. Age	5.9	0.4	−0.055	–				
3. Enrolled in ECCE	76.1		−0.063	0.006	–			
Household characteristics								
4. Mother has primary school education	27.5		−0.030	−0.040	0.037	–		
5. Father has primary school education	60.1		−0.030	**−0.120**	**0.102**	**0.246**	–	
6. Total assets (out of 13)	8.3	1.8	−0.027	−0.001	**0.080**	**0.141**	**0.223**	–
7. Language spoken at home								
Afaan Oromo	92.8		0.032	−0.032	**0.230**	−0.071	−0.049	0.012
Amaharic	1.5		−0.031	0.063	**−0.161**	**0.144**	0.075	−0.007
Guraghe	5.7		−0.019	0.003	**−0.173**	0.004	0.015	−0.010

Note: Bold indicates correlation is statistically significant at $p < .05$.

administering the tools with each other and finally practiced administering the assessments in pilot schools that were not included in any of the intervention groups. Data collection was completed using Android tablets on Tangerine software. Data was overseen by a Save the Children Measurement, Evaluation, and Learning Officer.

At baseline, 20 children were randomly chosen from each village with a priority on 5 and 6-year-olds and gender balance (i.e., 10 girls and 10 boys per village where possible). The same children were assessed again at endline. Before the assessment began all caregivers were asked for consent to participate. In addition, all children were asked for their assent to participate before each IDELA administration and were allowed to stop the assessment at any time without penalty. The assessment was predominantly administered at ECD centers in the study area. Enumerators took children out of their classroom, administered the assessment in a quiet location within the school area, and then returned the child to class. One enumerator worked with one child for all assessments and children's responses were scored using Android tablets with Tangerine software. If a child completed at least three assessment items, the case was considered a valid assessment even if the child chose to stop participating later in the assessment. If a child agreed to participate but did not complete three items before stopping or stalling, the assessment was terminated and another child was randomly chosen. Each IDELA assessment took approximately 35 min per child.

2.3. Measures

2.3.1. International Development and Early Learning Assessment

All children were assessed using the IDELA assessment, which was translated into Afaan Oromo, the mother tongue of the participants, and adapted using a process of review and field testing by the West Showa Field Office in collaboration with the tool developers from Save the Children. Below we summarize the content of each domain and describe some example items. Note that the number of items per subtask is specific to the version of the IDELA used for data collection in the present study. Minor additions and omissions of specific items have occurred in later versions of the IDELA, but these have retained the overall subtask and domain structure of the instrument described here. Most items are scored as "correct/incorrect", but a few are scored as ordered-categorical, with higher numbers denoted better performance on the item. Items with more than 2 response categories are indicated below.

2.3.1.1. Motor development. The Motor domain of the IDELA consists of 10 items, grouped into 4 subtasks: Copying a Shape (1 item), Drawing a Human Figure (7 items), Folding a Piece of Paper (1 item; 4 categories), and Hopping on One Foot (1 item; 11 categories). These items are intended to assess both Gross and Fine Motor skills. As an example of an item related to gross motor skills, Hopping on One Foot requires the child to stand on one foot and hop forward. The assessor counts the number of steps hopped by the child without putting down the other foot (up to 10). As an example of an item measuring fine motor skills, Folding a Piece of Paper requires the child to follow a four-step example of the assessor folding a piece of paper. Each step is scored correctly if the child closely replicates the various folds at each step (within 1 cm).

2.3.1.2. Social-Emotional development. The Social-Emotional domain consists of 14 items grouped into 5 subtasks: Emotional Awareness (2 items), Perspective Taking (3 items), Conflict Solving (2 items), Personal Awareness (6 items), and Names of Friends (1 item; 11 categories). For example, each Conflict Solving item asks the child to decide what he/she would do if he/she were playing with a toy and another child wanted to play with the same toy. "Correct" answers, as agreed upon with local staff, included talking to the child, taking turns, and sharing.

2.3.1.3. Early Literacy[1]. The domain of Early Literacy consists of 38 items grouped into 6 subtasks: Print Awareness (3 items), Letter Identification (20 items), Phonological Awareness (6 items), Oral Comprehension (5 items), Emergent Writing (2 items; 1 item with 4 categories), and Expressive Vocabulary (2 items; 11 categories each). All text-based items used the Latin alphabet. Notably, over half of the items on the Early Literacy domain comprised a single subtask – Letter Identification. This task requires the child to point to letters as they are read out by the assessor. The assessor starts with 10 high frequency letters, and then moves onto 10 lower frequency letters. Importantly, a skip pattern was utilized on this subtask, which led to the vast majority of children (78%) in the baseline sample being administered only the first 10 items. The overall percentage of children seeing only the first 10 items was lower at endline (51%), but remained high for the control group (80%). Due to the low response rate for the latter 10 items on the Letter Identification subtask, these items were not used in the present analysis.

As a second example, one of the Print Awareness items involved asking children to help the assessor open a book so that they could read a story together. The book is handed to child upside down, with the cover facing up. The item is scored as correct if the child orients the book correctly and opens the front cover.

2.3.1.4. Early Numeracy. The domain of Early Numeracy consists of 38

[1] The items on Early Literacy domain used in the endline sample differ from that described here. Only three Phonological Awareness items were administered instead of 6; and only 1 item was administered for Emergent Writing instead of 2.

items grouped into 8 subtasks: Comparison by Size and Length (4 items); Sorting and Classification (2 items); Shape Identification (5 items); Number Identification (20 items), Counting (3 items), Addition and Subtraction (3 items); and Puzzle Completion (1 item; 4 categories). The Number Identification used 20 Arabic numerals in a manner analogous to the Letter Identification subtask, with the numbers 1–10 administered first, followed by the numbers 11–20. The non-response rates were also similar to the Letter Identification items, with 75% of children being administered only the first 10 items at baseline, and 41% of children (70% of control group) at endline. Due to the low response rate for the last 10 items on the Number Identification subtask, these items were not used in the present analysis.

As an additional example, the Shape Identification subtask involved showing the child a picture with six geometric shapes and asking the child to identify a subset of the shapes named by the assessor. Each response was considered correct if the child pointed to the correct shape.

2.3.2. Caregiver survey measures

All caregivers were assessed with the IDELA caregiver survey. At endline, caregivers were 58% mothers, 32% fathers, 4% grandparent, 3% brother/sister, and 3% other. The items used in this study include *mother's education* (whether the mother has at least a primary education; 27.5%); *father's education* (whether the father has at least a primary education; 60.1%); *enrollment in formal early childhood care and education* (76.1%), and a *household asset index* (measured on a scale from 1 to 15; $M = 8.3$, $SD = 1.8$). Assets included the following (adapted from the MICS; UNICEF, 2014): bedroom; kitchen; living room; washroom; inside toilet; outdoor toilet; radio; television; refrigerator; bicycle; motorcycle; mobile phone; electricity; land for crops; and livestock, family animals, or poultry.

2.4. Analytic plan

Three sets of analyses were conducted. Research questions 1 through 4 were addressed using exploratory and confirmatory factor analysis. Research question 5 was addressed using latent regression. Finally, measurement invariance analyses were undertaken to address question 6. All analyses used the baseline sample, except for the third and final measurement invariance analysis, which considered treatment group status at endline. The item-level data were modeled as binary/ordered-categorical with a probit-link function, except where small sample sizes in the measurement invariance analyses led to empty cells for the ordered-categorical variables in one of the subgroups. In these cases (noted below), we treated the ordered-categorical variables as continuous and normally distributed. All analyses were conducted in Mplus (Muthén & Muthén, 2014), using the Weighted Least Squares estimator with cluster-robust chi-square statistics and standard errors used to correct for nesting of students within communities. Chi-square difference tests of nested models used Mplus's diff-test module.

2.4.1. Factor analyses

We randomly divided the $N = 682$ baseline observations into two subsamples: an exploratory sample and a confirmatory sample. The purpose of the exploratory sample was to allow for multiple variations on initial models to be fitted in order to arrive at a "proposed model" for each domain, as well as for the overall IDELA. The purpose of the confirmatory sample was to ensure that the proposed models demonstrated out-of-sample generalizability. A minimum sample size of $N = 454$ for the confirmatory sample was determined by conducting power analyses using a bi-factor model for each domain (see Appendix A). The remaining 228 observations were used for exploration.

Research questions 1 and 2 were addressed by analyzing each domain separately. The exploratory analysis focused on evaluating the fit of the hypothesized unidimensional model within each domain, and whether any departures from the hypothesized model were compatible with the pre-existing subtask structure of the IDELA. After establishing suitable models for the domains, we then addressed item specificity and the relationship among the domains (research questions 3 and 4) by combining the within-domain models into an overall model.

2.4.1.1. Exploratory analyses within domains. For each domain of the IDELA, we conducted a series of analyses using the exploratory sample. First we fit a unidimensional factor model to the domain items, without modeling the subtask structure. We examined targeted misspecification indices ("modification indices"; see Sörbom, 1989) for the residual correlations among items, in order to identify whether any deviations from the single-factor model were consistent with the subtask structure of the IDELA (i.e., whether larger residual correlations were associated with items on the same subtask, rather than items on different subtasks). Second, we conducted an exploratory factor analysis (EFA) using the bi-factor rotation (Jennrich & Bentler, 2012). We ran a series of EFA models, varying the number of residual factors from 1 to the maximum number computable for each subtask to evaluate whether the factor pattern of the residual factors was consistent with the subtask structure of the IDELA (i.e., whether items on the same subtask loaded on the same residual factor). The EFA provided evidence about whether any additional factors were consistent with the pre-existing subtask structure of the IDELA.

Based on evaluation of the previous two analyses, we arrived at a proposed model for each domain. Our third analysis with the exploratory sample assessed the goodness of fit of these proposed models using conventional standards for CFA (Hu & Bentler, 1999).

2.4.1.2. Exploratory analysis across domains. We next analyzed all items simultaneously by combining the proposed models from the within-domain analyses. A total of three models were fit to the full IDELA assessment, again using the exploratory sample. These three models differed in terms of how the correlations among the domain factors were modeled.

The first model combined the domain-level models without placing any restrictions on the correlation matrix of the domain factors. We addressed whether any items loaded on more than one domain (i.e., item specificity) by examining overall model fit and targeted model misspecification indices. This first model provided a reference model for chi-square difference testing of the two following models. We therefore refer to it as the "Unconstrained Model."

The second model replaced the four domain-level factors with a single factor to test whether the four domains were really providing unique information, or whether IDELA *only* measures a single overarching construct. We refer to this as the "Unidimensional Model." The third model was a hierarchical factor model in which the correlations among the four domain factors were modeled using a higher-order unidimensional factor model. This model tested the assumption that the four IDELA domains were related to one another via a single overarching construct. We refer to this as the "Hierarchical Model."

The fit of the Unidimensional Model and the Hierarchical Model was assessed using chi-square difference testing against the Unconstrained Model. This approach provides a more rigorous test of the Hierarchical Model and the Unidimensional Model than using goodness of fit indices alone (see e.g., Millsap, 2012). Power analysis for this test is reported in Appendix A.

2.4.1.3. Confirmatory analyses. The final analytic step in our factor analyses involved assessing the out-of-sample generalizability of the models developed for the within- and across-domain analyses, using the confirmatory sample.

2.4.2. Latent regression analysis

We considered how a small set of parent and child characteristics were associated with each of the domain factor scores. This purpose of this analysis was to provide an initial consideration of the concurrent

Table 2
Summary of goodness of fit for proposed models at the domain level: exploratory and confirmatory sample results.

Domain	χ^2 (df)	RMSEA (90% CI)	TLI
Exploratory sample			
Gross and Fine Motor	15.53 (15)	.013 (.000, .064)	.999
Early Literacy	386.53 (325)	.029 (.015, .039)	.986
Early Numeracy	363.13 (325)	.023 (.000, .035)	.995
Social-Emotional	111.02 (67)	.054 (.035, .071)	.952
Social-Emotional*	59.93 (46)	.034 (.000, .060)	.983
Confirmatory sample			
Gross and Fine Motor	11.77 (9)	.026 (.000, .062)	.997
Early Literacy	485.14 (325)	.033 (.027, .039)	.987
Early Numeracy	381.13 (325)	.020 (.009, .027)	.996
Social-Emotional	59.39 (46)	.025 (.015, .042)	.991

Note: χ^2 (df) denotes the chi-square test of model fit and its degrees of freedom. RMSEA denotes the root mean square error of approximation and (90% CI) its 90% confidence interval. TLI denotes the Tucker Lewis Index.

* The Social-Emotional domain had two items that exhibited sizable residual correlations with items on other subtasks. Results here represent model fit after removing these two items. All subsequent models for the Social-Emotional domain omit these two items.

validity of the IDELA, by examining whether parent and child characteristics that are known to be related to children's early primary school enrollment in Ethiopia and are often of interest to the international development research community, are, in fact, related to the IDELA factors in the present sample of children. Because the present sample was not designed to provide concurrent validity evidence about the IDELA, our efforts were necessarily opportunistic, and certainly fall short of a full investigation of validity. However, we provide some initial indication of the relationships between the IDELA and the following, limited, list of variables available for analysis: child age, child gender, whether the child is enrolled in an ECCE program; mother's and father's level of education; and an index of household assets as a proxy for socio-economic status.

We used conventional latent regression analysis to simultaneously regress the IDELA domains on these variables, using the full sample at baseline, as well as the above noted corrections to standard errors for nesting of children within communities.

2.4.3. Measurement invariance analysis

We investigated measurement invariance of the IDELA over three pairs of subgroups: (1) children enrolled in an early care and education center (ECCE; $N = 519$) versus home care ($N = 163$); (2) boys ($N = 323$) and girls ($N = 359$); and (3) ELM Center treatment group ($N = 465$) versus control group ($N = 160$). The first two analyses used the baseline data and provide a more thorough examination of group differences reported in the latent regressions. The third analysis was conducted using endline data to provide initial indication of the suitability of the IDELA as an outcome measure in program evaluation. We examined whether the measurement properties of the IDELA were invariant over treatment conditions, and well as whether the domain-level factors were sensitive to treatment effects.

We first evaluated the configural model (whether number of factors and general pattern of factor loadings is the same across groups). We then evaluate the metric (whether the factor loadings are the same across groups) and scalar models (whether the intercepts or thresholds of the items are equivalent across groups) using chi-square difference testing against the configural model. For more details on the different types of measurement invariance, see Millsap (2012). Each of the measurement invariance analyses were quite highly powered (see Appendix A). However, it is important to note that two of the comparisons involved subgroups with sample sizes that are quite small by psychometric standards (the home care group at baseline; and the control group at endline). In particular, the low frequency of correct responses to more difficult items meant that the correlation matrix among the items was not estimated very precisely in these subgroups. Consequently, the results of these two comparisons must be regarded as preliminary and tentative.

3. Results

Several items were removed from the analysis presented due to low response rates resulting from stopping rules, as well as "Heywood cases," which occur when an item has a negative residual variance (see, Bartholomew, Knott, & Moustaki, 2011, sec. 3.12). A summary of items removed is included in Appendix B.

3.1. Factor analyses

3.1.1. Exploratory factor analysis within domains (RQ1, RQ2)

A unidimensional model did not fit any of the IDELA domains particularly well (e.g., RMSEA values in the range [0.058, 0.082]). However, visual inspection of the modification indices from the unidimensional models indicated a clear pattern in every domain; there were many large residual correlations among the items on the same subtask. Aside from two items on the Social-emotional domain (see below for description), there were no large residual correlations across subtasks on any domain. Visual inspection of the factor pattern in the exploratory bi-factor analysis also supported the conclusions that residual factors corresponded to items on the same subtask. Additionally, when the number of residual factors were increased, the items retained relatively large loadings on the general factor.

Based on these evaluations of solutions for the unidimensional model and EFA, we concluded that there was reasonable empirical evidence to proceed with a bi-factor model within each domain, using the residual factors to model the individual subtasks. Subtasks comprised by a single item by definition did not receive a residual factor. For subtasks comprised by two item or three items, it is statistically equivalent to treat the subtasks either in terms of residual correlations or a residual factor (i.e., the same number of parameters are used in both approaches). We chose to use a residual factor in these cases, to avoid possible confusion about the interpretation of the factors versus residual correlations. For the two-item subtasks, both factor loadings were fixed to one to identify the variance of the residual factor.

Table 2 summarizes the goodness of fit of the bi-factor models in each of the domains. The fit for Early Literacy, Early Math, and Motor domains was very good; but the fit for the Social-Emotional domain was marginal. As noted above, the Social-Emotional domain had two items that exhibited sizable residual correlations with items on other subtasks. One item was on the Personal Awareness subtask (Does the child know the name of the country he/she lives in?), which correlated with items on the Perspective Taking, Emotional Awareness, and Conflict Solving subtasks. This item was more difficult than the other Personal Awareness items and was not strongly related to its residual factor. The second item was the on the Conflict Solving subtask (Does the child provide an example about how to resolve conflict with peers?). While this item correlated very highly (.868) with the other item on the Conflict subtask, it also had large residual correlations with the items on the Perspective Taking subtask. After omitting both of these items, the remaining items showed very good fit to the bi-factor model (see the final row of Table 2). We therefore omitted these two items from all subsequent analyses.

3.1.2. Exploratory factor analysis across domains (RQ3, RQ4)

Using the domain-specific models described above, the next step involved fitting all of the items in the same model. The goodness of fit for the three overall models is reported in Table 3.

3.1.2.1. Unconstrained Model. The Unconstrained Model for all items showed good model fit. However, examination of modification indices showed that the residual factors for the Letter Identification and Number Identification scales were strongly correlated ($r = .655$).

Table 3
Summary of goodness of fit for the overall IDELA model: exploratory and confirmatory sample results.

Model	χ^2 (df)	RMSEA (90% CI)	TLI	χ^2-diff (df)	*p*-Value
Exploratory sample					
Unconstrained	2775.24 (2633)	.015 (.006, .021)	.985	NA	NA
Unidimensional	2795.87 (2639)	.016 (.008, .022)	.983	33.81 (6)	< .001
Hierarchical	2782.64 (2635)	.016 (.006, .021)	.984	11.68 (2)	.003
Confirmatory sample					
Unconstrained	2869.24 (2633)	.014 (.010, .018)	.987	NA	NA
Unidimensional	2881.89 (2639)	.014 (.010, .018)	.986	23.11 (6)	< .001
Hierarchical	2872.64 (2635)	.014 (.001, .018)	.987	6.15 (2)	.046

Note: χ^2 (df) denotes the chi-square test of model fit and its degrees of freedom. RMSEA denotes the root mean square error of approximation and (90% CI) its 90% confidence interval. TLI denotes the Tucker Lewis Index. χ^2-diff (df) denotes the chi-square difference test and is df; *p*-value is reported in the last column.

These two subtasks share a similar format and both involve print awareness (i.e., pointing to a spoken letter or number); therefore, the correlation between the residual factors was theoretically plausible and we added it to the overall model. We also found that one item on the Motor domain (Hopping on One Foot; see above for description) was correlated with items on both the Early Literacy and Early Numeracy domains. The hopping item was quite easy for most students (over 85% of children obtained the maximum score of 10), and this item had a relatively weak loading on the Motor domain factor (.434). The residual correlation was apparently due to the fact that children who could not hop the full 10 times were also more likely to do poorly on many of the literacy and numeracy items. Based on these considerations, we removed the hopping item from further analyses. Note that goodness of fit reported in Table 3 is for the model with these two modifications included.

It was notable that the correlations among the factors were very large, ranging from r = [.747, .958]. This suggested that a simpler model might be viable.

3.1.2.2. Unidimensional Model. The second line of Table 3 shows the goodness of fit of a unidimensional factor fitted to all items, while preserving the bi-factor structure for the individual subtasks. Although the overall fit of the model is acceptable, the chi-square difference test revealed that the Unidimensional Model fit the data significantly worse than the Unconstrained Model. Thus, while the correlations among the domains were large, we rejected the hypothesis that a single-factor model provided equivalent fit to the data. This was somewhat surprising, given the high correlations among the factors in the Unstrained Model. We further explored the possibility of a reduced number of domains by collapsing the two most highly correlated domains into a single factor (Early Literacy and Numeracy, r = .958), while leaving the other two domains intact. Again, the restricted model was found to fit the data worse than the unconstrained model, although this time by a much narrower margin (χ^2 (3) = 11.015, p = .012).

3.1.2.3. Hierarchical Model. The last row of Table 3 indicates that the Hierarchical Model also fit the data well, but this model was also rejected by the chi-square test against the Unconstrained Model. The factor loadings on the higher order factor ranged from r = [.860, .971], indicating that all of the domains were very strongly related to the higher-order factor. However, the relationships among the four domains were not strongly consistent with the hypothesis of a single, higher-order factor.

3.1.3. Confirmatory factor analyses

As final step, we considered the out-of-sample generalizability of the results reported in the exploratory analyses using the confirmatory sample. As shown in Table 2, all of the domain-level models were replicated with very acceptable goodness of fit in the confirmatory sample. Parameter estimates are summarized in Figs. 1–4 .

As shown in Table 3, the confirmatory sample also led to the same conclusions about the overall model. While all models fit the data well overall, the chi-square test against the Unconstrained Model again led to a sound reject of the Unidimensional Model, and a less decisive rejection of the Hierarchical Model. Parameter estimates for the Unconstrained Model and the Hierarchical Model are reported in Figs. 5 and 6, respectively. As with the exploratory sample, the correlations among the domains were very large, r = [.803, .961]. Interpreting the squared correlation coefficients in terms of proportions of shared variance, the highest value of 92% shared variance was between the Early Literacy and Early Numeracy domains. The Social-Emotional domain had the lowest correlations with all other domains, with proportions of shared variance ranging between 64% with the Motor domain, and 82% with the Early Literacy domain.

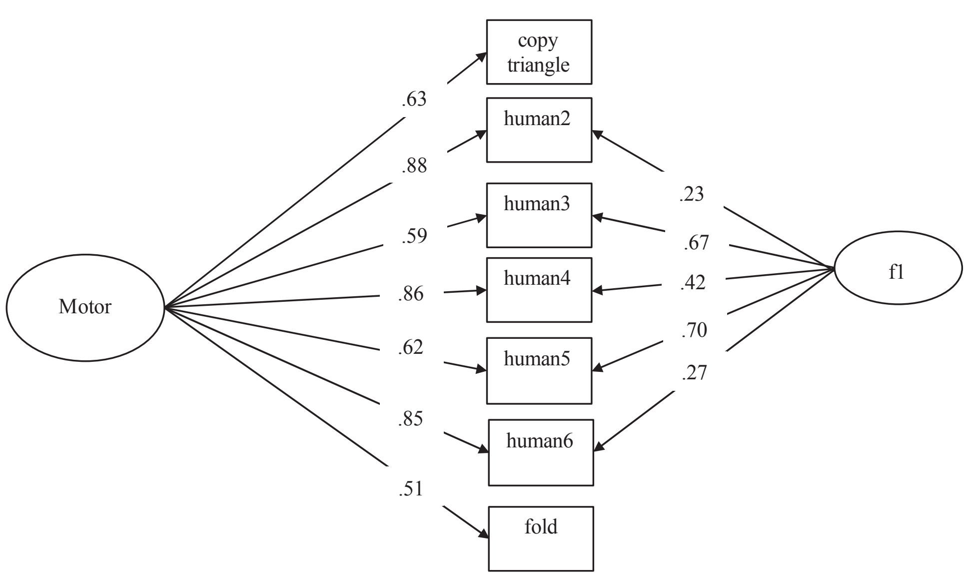

Fig. 1. Diagram of the final exploratory model of the Motor Development domain.

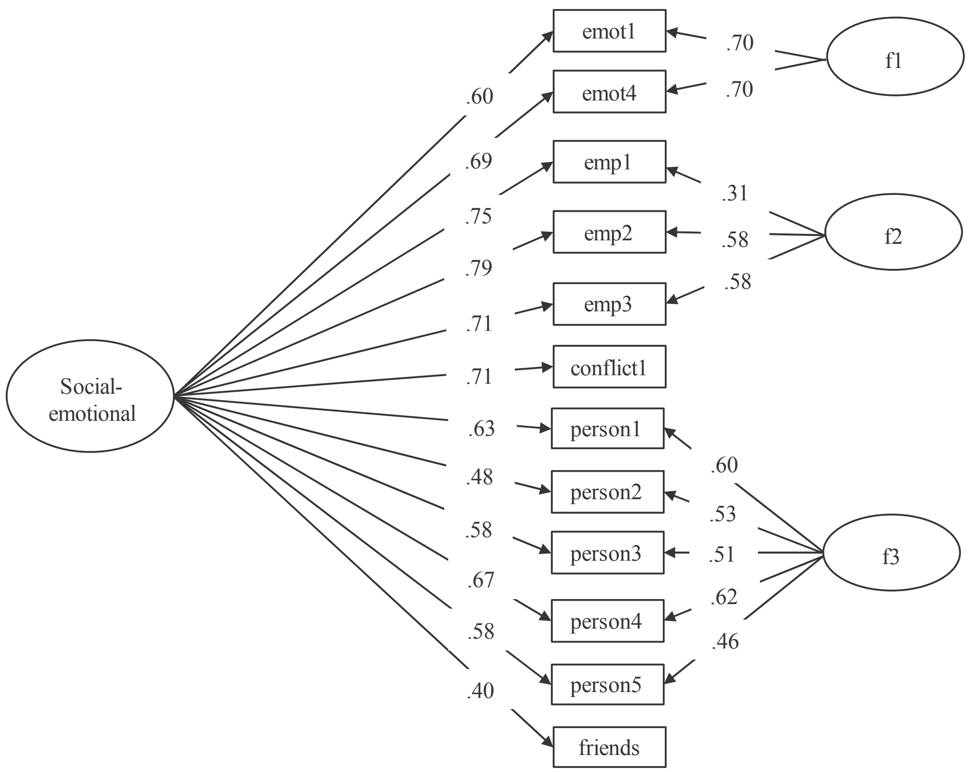

Fig. 2. Diagram of the final exploratory model of the Social-Emotional development domain.

3.2. Latent regression analysis (RQ5)

To provide an initial indication of the concurrent validity of the IDELA, we conducted latent regression analysis with the full baseline sample of N = 682 children using the Unconstrained Model. This allowed us to address how the predictor variables of interest were related to the each of the four domains. The model with the predictors had good fit (χ^2 (3059) = 3431.23, *RMSEA* = .013, *TLI* = .982) and the results are summarized in Table 4.

The consistent predictors of child outcomes were child age, enrollment in early education and fathers' education, all associated with higher scores on all four domains. Enrollment in ECCE had the largest associations by an order of magnitude (see Table 4). Household assets (a proxy for socioeconomic status) did not significantly predict any outcome.

3.3. Measurement invariance analysis (RQ6)

Measurement invariance was assessed over three pairs of subgroups: (1) enrolled in an early care and education center (ECCE; N = 519) versus home care (N = 163), at baseline; (2) boys (N = 323) and girls (N = 359), at baseline; and (3) ELM Center treatment group (N = 465) versus control group (N = 160), at endline. In each case, the Unconstrained Model depicted in Fig. 5 was fitted to each subgroup. Due to small sample sizes in the No ECCE group and in the control group, we treated ordered-categorical items (see description of the IDELA subtasks in the Section 2.3) as continuous and normally distributed for these comparisons. This was required to avoid empty cells occurring in one- and two-way tables of those items. No modifications were made to binary items. The goodness of fit for all three comparisons is shown in Table 5, and the group mean differences on the domains are summarized in Table 6.

3.3.1. Goodness of fit

For the comparison between boys and girls, the goodness of fit and chi-square tests against the configural model provided strong support for scalar invariance. In the endline sample, there was again strong evidence for scalar invariance between treatment conditions. However, for the ECCE versus home care comparison, the chi-square tests rejected the both scalar and metric invariance (however scalar invariance was not rejected in comparison to the metric model).

The modification indices for the ECCE versus home care comparison revealed two subtasks that were plausible sources of invariance: The Emergent Writing subtask on the Early Literacy Domain, and the Puzzle Completion subtask on the Early Numeracy domain. The Emergent Writing subtask was much more strongly related to the Early Literacy domain in the ECCE group, as compared to the home care group. This was plausibly due to lack of exposure to writing materials for children who only experienced home care. The Puzzle Completion subtask was much easier for children in the home care group, as compared to the other Early Numeracy subtasks. Unlike the other Early Numeracy subtasks, completing the puzzle did not require explicit knowledge of numbers or mathematical vocabulary. Allowing the measurement parameters (factor loadings and thresholds) on these two subtasks to vary over conditions, the resulting partial scalar invariance model was not rejected, when tested against the configural model (χ^2 (136) = 160.239, p = .076). We report mean differences for partial invariance model.

3.3.2. Group differences

Table 6 reports the mean difference, in standard deviation units, between each pair of groups on each of the domains. There were no gender differences on the IDELA domains. The group of children enrolled in ECCE show significantly higher scores on every factor than those in the home care condition, with differences ranging between 1.152 and 1.802 standard deviation units. In interpreting these differences, note the Early Literacy and Early Numeracy domains do not reflect differences on the Emergent Writing and Puzzle Completion subtasks, respectively. Finally, in the treatment contrast, we found that the children in the treatment conditions had significantly higher scores on three of the four domains, the exception of the Social-Emotional domain. The significant treatment effect sizes ranged from 0.423 on Early Literacy to 0.499 on the Motor domain.

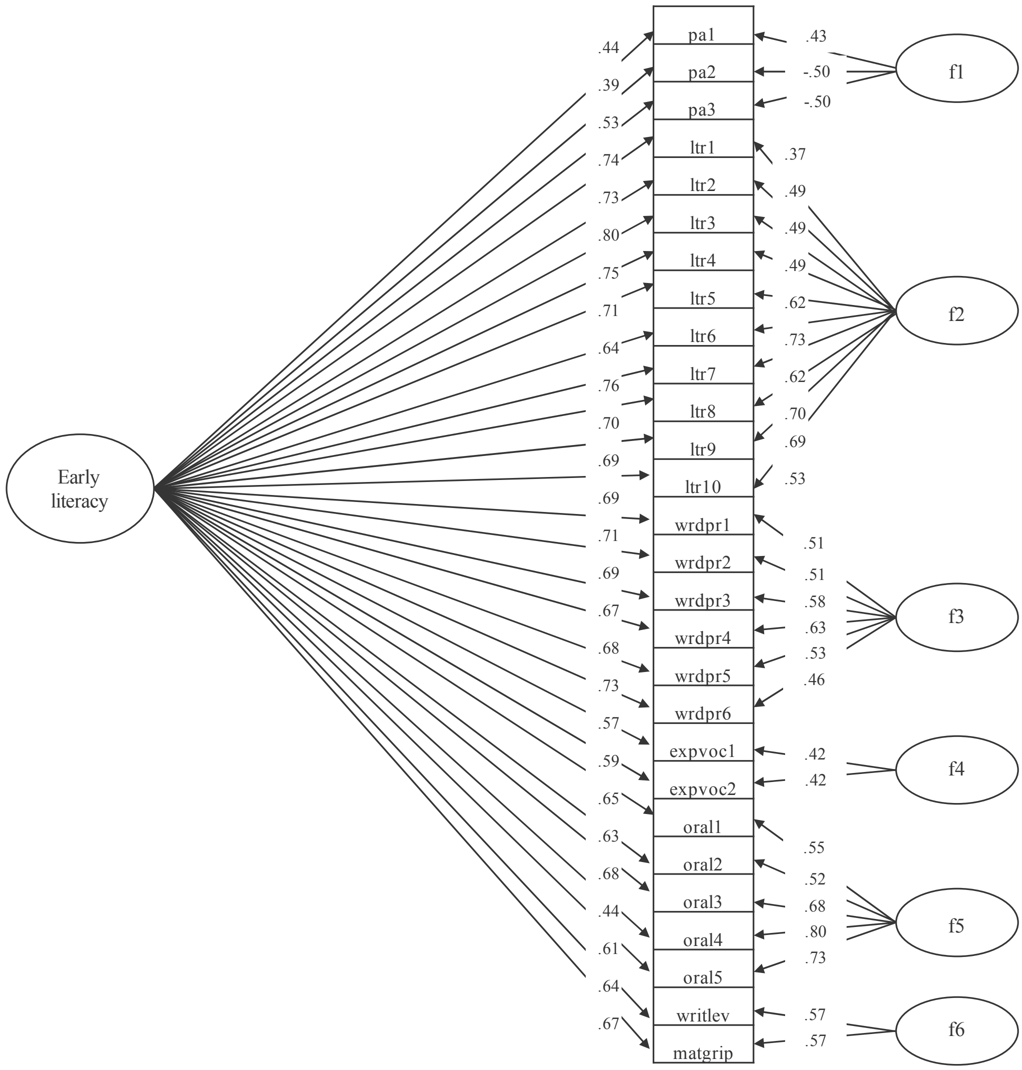

Fig. 3. Diagram of the final exploratory model of the Early Literacy domain.

4. Discussion

Assessments of school readiness are increasingly in demand in LMICs, and are critical in aiding governments to track the progress of children and the effectiveness of their programs (Yoshikawa & ECDAN Data Task Force, 2017). While several tools are currently being developed and used, few have been subject to rigorous analyses of their factor structure, measurement invariance, and validity. We presented one of the first studies of a direct early childhood assessment designed for use in LMICs to examine exploratory and confirmatory factor structures, as well as convergent validity and measurement invariance across important subgroups in Ethiopia. We found empirical evidence that in this context, the IDELA measures four distinct domains of children's development – Motor, Social-Emotional, Early Literacy, and Early Numeracy – with bi-factor models for individual constructs/sub-skills within each domain.

While the four constructs measured in the IDELA were distinct, they were also very highly correlated both to one another and to a higher-order construct measuring overall school readiness. However, replacing the four constructs with a single construct did not provide acceptable fit to the data, indicating that accounting for the unique domains is important. Thus in this sample, the IDELA was best conceptualized as four distinct developmental domains with a hierarchical factor structure for subconstructs. Early Literacy and Early Numeracy were the most highly correlated factors, and the Social-Emotional domain had the lowest correlations with the other factors (though still above 0.8). A recently developed parent-reported assessment of early childhood development (albeit for young children, 0–3) has identified similar patterns of relations among developmental domains in Tanzania, with cognitive, language, and motor items loading on to a single factor and social-emotional items loading on a separate factor (McCoy, Black, Daelmans, & Dua, 2016).

Notably, a three-level hierarchical factor with each domain loading on to a hierarchical "school readiness" factor accounting for the relationship among the four domains also fits the data very well. In addition, it appeared to be important to account for the sub-task structure (i.e., accounting for items that are administered with the same materials or questioning format, and that assess different skills within each domain). The structure reflects constructs and subconstructs of children's school readiness that have been agreed upon globally (Snow & Van Hemel, 2008; UNESCO, 2013), and have successfully been measured in both Western samples (Janus & Offord, 2007; Panter & Bracken, 2009) and in the East Asia/Pacific region (Rao et al., 2014). To our knowledge, this is the first study to identify empirical evidence of these constructs as distinct yet related domains in a sample of children in Ethiopia. Given the high correlations found among the domains, further research in other samples within Ethiopia and other countries is needed to understand the nature of the relations among developmental domains in LMICs.

We also found evidence of concurrent validity between the four developmental domains and different child and family variables. Building on previous research on gaps in enrollment in primary school

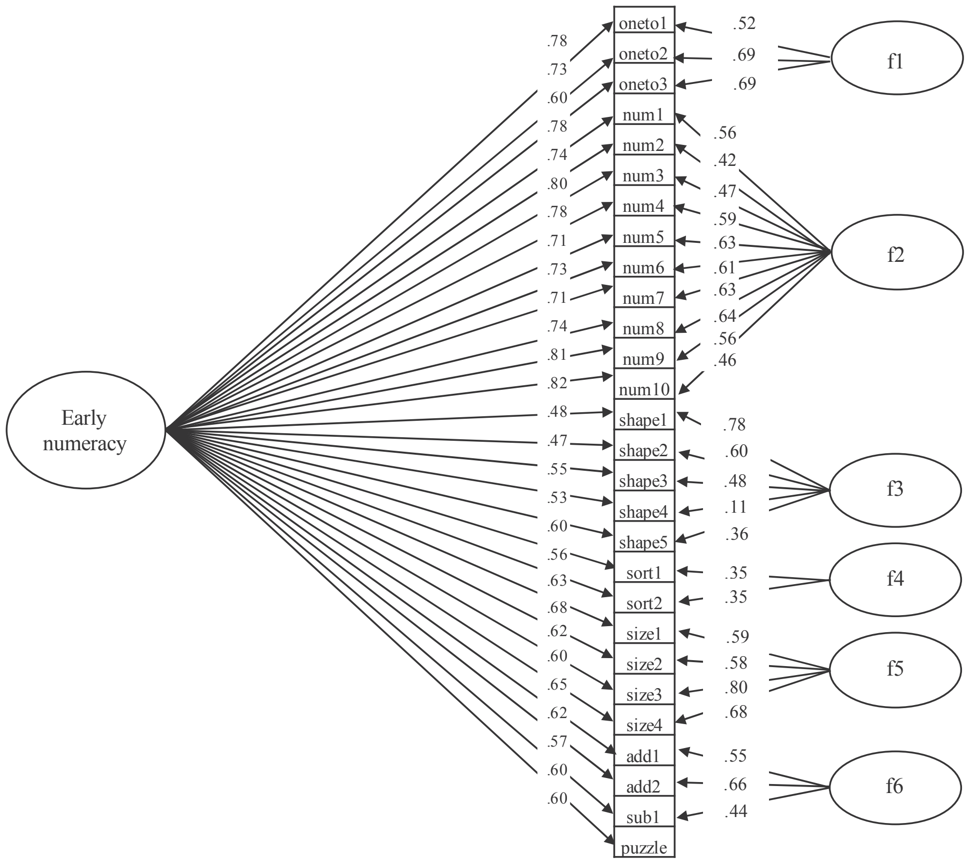

Fig. 4. Diagram of the final exploratory model of the Early Numeracy domain.

(Woldehanna et al., 2011), this analysis also contributes to the literature on early childhood development in Ethiopia specifically, and in low-income countries generally. As was found for primary school enrollment in the Oromia region of Ethiopia (Woldehanna et al., 2011), parents' education (in our case, fathers' education) predicted higher IDELA scores on all domains. Unlike primary school enrollment rates, household assets (a proxy for socio-economic status) and gender did not significantly predict outcomes in any of the four domains. It is possible that this index was not a good indicator of socio-economic status for this region, or that at this age, socio-economic status does not yet differentiate children's development. Further research is needed to understand whether the items comprising the index have relevance for household wealth in rural Ethiopia. Regarding gender, it is possible that gaps observed in primary school enrollment rates and academic outcomes identified in Ethiopia (e.g., Kassahun & Kedir, 2006) have not yet developed in early childhood. Additionally, children's age positively predicted outcomes in all domains, indicating that the tool successfully identifies developmental domains that tend to be acquired with age. Lastly, as has been shown in several other studies to date (Yoshikawa et al., 2013), enrollment in ECCE was a strong predictor, and larger in magnitude than the family factors investigated, of children's development across all domains.

Finally, we find that the IDELA factor structure was invariant to gender, early education and care setting, and treatment status subgroups, indicating the utility of the assessment in making comparisons across these subgroups. Importantly, we find invariance across treatment and control conditions after six months of implementation of the Early Literacy and Math (ELM) Intervention (Amente, Takele, Pisani, & Anis, 2015), indicating the utility of using the IDELA assessment in program evaluations. Notably, the ELM intervention consisted of three different versions of the treatment. The estimates derived in this analysis consider all of the three treatment conditions as one group for the sake of the exercise in this study. Thus, the estimates of program impacts are for the receipt of any type of treatment relative to the control condition. These results add to the findings from Save the Children's endline report (Amente et al., 2015).

We found smaller impacts on social-emotional skills compared to early academic and motor skills. The evidence to understand how different domains of development may be differentially impacted by ECE interventions in LMICs is not yet clear. One study in Indonesia found positive impacts of being offered access to ECE on children's language, cognitive, motor, and social-emotional domains, with the smallest impacts on social-emotional (Hasan, Hyson, & Chang, 2013). Another study in rural Mozambique found positive impacts of being offered access to ECE on children's cognitive development and language, but no impacts on social and emotional competence, as reported by first grade teachers (Martinez, Nadeau, & Periera, 2012). Finally, studies in Chile (Yoshikawa et al., 2015) and Ghana (Wolf, Halpin, & Yoshikawa, 2017) found that programs to improve ECE quality for children already enrolled in ECE impacted only social-emotional and behavioral skills. In this data, we find evidence that early academic and motor skills are more malleable to increased ECE access than social-emotional skills in rural Ethiopia.

Based on this analysis, we recommend several revisions to the IDELA and for analyzing IDELA data. First, we suggest the evaluation and removal of highly collinear items. Second, we recommend reducing the number of items on the letter and number identification scales to avoid the use of stop rules (i.e., skipping more difficult items if children are unable to answer simpler items on the same domain). This reduction could allow for the addition of different items to the motor domain, since many collinear items in this dataset were on this scale. Third, for researchers who may not have 35 min with each child, we recommend considering planned missingness designs (e.g., only administer one or two domains per child, randomly assigned; Little & Rhemtulla, 2013). Importantly, further revisions would likely need to be motivated by a longitudinal study to understand if the relations among items change

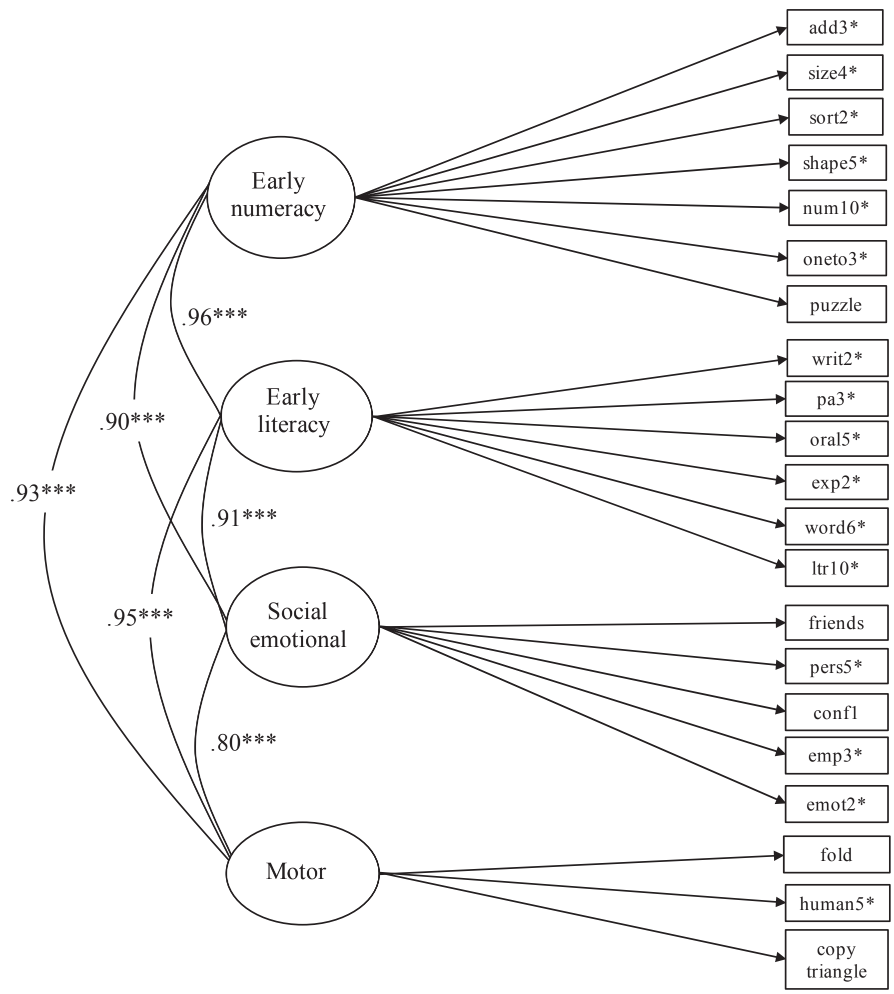

Fig. 5. Diagram of the unconstrained factor analysis across four domains. *Note*: Diagram depicts subtasks only, though item level responses were included in analysis. The number indicates the number of items in the particular subtask. * indicates that a residual factor was fit for the items in the subtask.

over time, as well as the predictive validity of the domains as currently measured.

The global need for direct early childhood assessments for national-level progress on SDG Target 4.2 will only grow in coming years. Consensus exists that caregiver reports of child development are inadequate. Multi-domain direct assessments that are feasible to administer and capturing the distinct domains of developmental included in IDELA would meet national-level reporting requirements for the target. Indeed, recent work with IDELA data from five additional countries in multiple regions of the world suggests that the four-domain factor structure consistently fits the data extremely well (Wolf et al., 2017; Wuermli et al., 2016), suggesting that IDELA is a potential candidate for national monitoring systems of early childhood development.

4.1. Limitations

This study has a number of limitations related to the research sample and the methodology that we hope will be addressed in future research on the IDELA and similar assessments. First, while the sample was representative of a single region in Ethiopia, the results cannot be assumed to generalize to the Ethiopian population at large, and, of course, cannot be assumed to generalize to other countries. All of the reported findings are therefore quite preliminary, although we hope the analyses illustrate some useful strategies for researchers addressing similar questions across a range of country contexts. Second, while the data were longitudinal, the time span was only 6–7 months, and most children received an experimental intervention during this time. Longitudinal data over a longer span of development would be useful to assess the stability of the IDELA factor structure as children age, as well as its predictive relations to child outcomes as in primary school.

In terms of methodology, the current study relied on a covariance-based estimator to handle a relatively large number of factors in the overall IDELA model (4 domains plus 16 residual factors). Recent methodological developments in Item Response Theory (e.g., Cai, 2010) allow for maximum likelihood estimation based on the full contingency table, even with large number of latent variables. However, these methods are not yet widely available in commercial software.

A second limitation is that we have not broached the subject of how to score the individual domains (e.g., using the total score within domain, using averages over subtasks). The factor structure of the IDELA is certainly quite complex, and relatively simple scoring procedures based on the raw data are not likely to perform as well as model-based scoring procedures commonly used in the psychometric literature (see e.g., Hambleton, Swaminathan, & Rogers, 1991). The current best practice for dealing with educational survey data is to use so-called "plausible values" (see Mislevy, 1991), in which multiple imputations are drawn from the posterior distribution of the factors. This approach was developed for use with the National Assessment of Education Progress in the United States, and has more recently been adopted by international assessments, including the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS). However, it can be very complex to implement in practice. Alternatively, researchers may prefer to use latent regression and other structural equation modeling techniques to simultaneously estimate the IDELA factor structure and the relation of the factors to covariates of interest, without needing to estimate domain scores for each child as an intermediate step (e.g., Bollen, 1989;

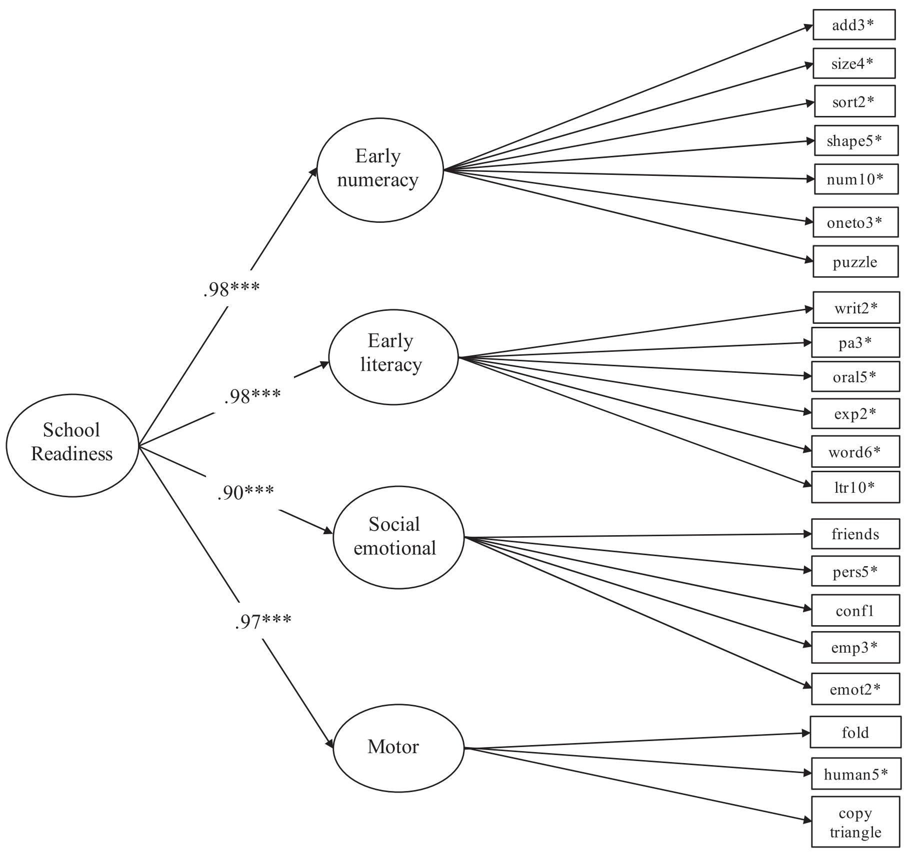

Fig. 6. Diagram of the hierarchical factor model, with all domain factors loading onto one general factor. *Notes*: Diagram depicts subtasks only, though item level responses were included in analysis. The number indicates the number of items in the particular subtask. * indicates that a residual factor was fit for the items in the subtask.

Mislevy, 1985; Skrondal & Rabe-Hesketh, 2004). In general, while the research reported here has provided some initial support for the construct validity of the IDELA, applied researchers should note that this work does not validate any particular scoring method for the domains or the subtasks.

4.2. Implications for measuring early childhood development globally

As global discussions continue about the utility of national monitoring of child development and learning in light of Sustainable Development Goal 4, finding assessment tools that are comparable for and sensitive to a range of diverse groups within a country is critical. Understanding an assessments' sensitivity to developmental change, as well as to a large range of program and policy initiatives, is needed. The findings in this paper show that the IDELA appears to be sensitive to exposure to formal ECCE and to program impacts of the ELM intervention. Indeed, a recent national study of ECCE impact in collaboration with Bhutan's Ministry of Education and UNICEF reinforce this conclusion at a scale above that attempted in this sample (Pisani et al., 2017).

The measurement properties identified in this study suggest that the IDELA is a useful addition to the spectrum of direct assessment tools available to measure of children's school readiness and development (e.g., Panter & Bracken, 2009; Rao et al., 2014) in the preschool and early school years, and to understanding the impacts of particular interventions on young children's development. This may particularly be the case in low-income countries. Its combination of several domains of development in one comprehensive instrument, its relatively short assessment time (~30 min), and the minimal material resources (i.e., 12 picture cards, small items for counting, children's book, blank paper, and pencils) required to administer the assessment make it relatively feasible in low-resource settings.

Table 4
Parameter estimates and standard errors of latent regression analysis.

	Motor development	Socio-Emotional development	Early Literacy	Early Numeracy
	b (SE)			
Child is female	.144* (.069)	.044 (.051)	−.068 (.049)	−.045 (.069)
Child age	.442*** (.096)	.266*** (.055)	.301*** (.084)	.465*** (.066)
Child enrolled in ECCE	.704*** (.116)	.664*** (.087)	.780*** (.126)	.821*** (.132)
Mother's education (at least primary school)	.146^{+} (.076)	.137* (.068)	.105 (.070)	.197 (.082)
Father's education (at least primary school)	.228** (.083)	.210** (.068)	.192** (.074)	.272** (.088)
Household asset index	−.021 (.028)	−.018 (.025)	.026 (.027)	.018 (.038)

Note: Estimates are derived from a single latent regression model. Unstandardized parameter estimates shown. Goodness of fit statistics for the model are as follows: χ^2 = 3431.23 (3059); RMSEA = .013; TLI = .982.

$^{+}$ $p < .10$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table 5
Model fit statistics for measurement invariance.

	Configural invariance	Metric invariance	Scalar invariance
Boys vs. girls			
Chi-square (df)	5571.51	5628.28	5730.47
contributions:	(5266)	(5337)	(5437)
Boys (N = 323)	2769.62	2822.25	2872.94
Girls (N = 359)	2801.90	2806.03	2857.54
RMSEA	.013	.013	.013
TLI	.989	.990	.990
χ^2 difference (df)		76.33 (71)	186.82 (171)
χ^2 difference p-value		.331	.193
ECCE vs. home care			
Chi-square (df)	5463.06	5583.67	5655.95
contributions:	(5269)	(5340)	(5411)
ECCE (N = 519)	3144.66	3123.28	3097.26
Home care (N = 163)	2318.40	2460.39	2558.69
RMSEA	.010	.012	.012
TLI	.988	.985	.985
χ^2 difference (df)		140.68 (71)	223.40 (142)
χ^2 difference p-value		$< .001$	$< .001$
Treatment vs. control			
Chi-square (df)	4788.06	4848.12	4913.62
contributions:	(4566)	(4632)	(4698)
Treatment (N = 465)	2551.51	2662.93	2707.73
Control (N = 160)	2236.55	2185.18	2205.89
RMSEA	.012	.012	.012
TLI	.992	.992	.992
χ^2 difference (df)		78.20 (66)	152.31 (132)
χ^2 difference p-value		.145	.109

Table 6
Standardized mean differences (with large-sample standard errors) on the IDELA domains, for three subgroup comparisons.

	Motor development	Social-Emotional development	Early Literacy	Early Numeracy
ECCE vs. home care	1.152 (0.225)	1.655 (0.298)	1.644 (0.289)	1.802 (0.337)
Boys vs. girls	0.260 (0.180)	0.082 (0.179)	−0.062 (0.182)	−0.038 (0.164)
Treatment vs. control	0.499 (0.179)	0.125 (0.222)	0.423 (0.160)	0.443 (0.123)

Note: The mean differences are reported in standard deviation units and were obtained by subtracting the mean of the second group from that of the first. No differences by gender were significant. All other differences were significant beyond the $p < .001$ level, except for Social Emotional in the Treatment versus Control comparison. The values reported for the ECCE versus Home Care comparison are from the partial invariance model described in the text, not the full invariance model, which was rejected (see Table 5).

To date, the IDELA has been used in over 30 countries by Save the Children and increasingly by other researchers and practitioners to measure early childhood development and school readiness. The current study shows that, at least in the Oromia Region in Ethiopia, accounting for the four unique domains and the subtask structure within each domain is important. The analyses provide insight into the promise of using factor analytic methods to understand the utility of measures developed to assess children's school readiness, and lay the groundwork for similar research with the IDELA and other measures across countries. Understanding the relations between the IDELA assessment and other assessments across contexts would be an important undertaking to help researchers and governments decide which assessment best meets their needs.

4.3. Implications and future directions for research and practice

Two key implications for future research emerge from the current study. First, while global and regional assessments of school readiness already exist, future studies should assess the measurement properties of these other tools. Understanding whether the empirical factor structure of other assessments do (or do not) mirror the conceptual structure will allow researchers to know how to most accurately score and present the measure. We were not able to locate published presentations of confirmatory factor structure, for example, for any multi-domain direct assessments or parent/teacher reports of school readiness in LMICs. We were also not able to locate any published examples of measurement invariance across conceptually important subgroups for either direct assessments or parent/teacher reported measures.

Second, with the growth in coordinated efforts to develop measures of early childhood development and early education contexts – such as the multi-agency Measuring Early Learning Quality and Outcomes (MELQO) Consortium and the UNICEF Multiple Indicators Cluster Survey (MICS) – there is a need for cross country measurement invariance analyses to assess the factor structure of measures across different countries and contexts (both within and across countries). This work is an important next step in understanding how results from global assessment tools can be compared across countries, as well as helping governments measure the state of learning and development for young children in their own countries.

Acknowledgements

This paper reflects contributions from several individuals. We would like to acknowledge and thank Sarah Kabay, Natalia Rojas, and Mackenzie Whipps, who provided analytic and conceptual contributions to this work. We would also like to thank Larry Aber for his guidance throughout the project. We thank NYU Abu Dhabi Research Institute for core support to the Global TIES for Children Center, which provides intramural support for research activities that directly contributed to this paper. Finally, we thank Celine Gustavson, Dan Stoner and Sara Poehlman without whose support for the development and field testing of IDELA to better the lives of children in Ethiopia and around the world this work would not be possible.

Appendix A. Power analyses

Note that the calculation reported here do not take into account clustering of students within schools and villages, and should therefore be regarded as approximate. However, with the exception of testing the Motor domain in the exploratory sample, we may safely conclude that the present study is very highly powered to assess all hypotheses of interest.

A.1 Establishing the size of the confirmatory sample

Power analyses were conducted at the domain level and for the overall unconstrained model to establish the sample size of the confirmatory sample. Results of the power analyses are summarized in Table A1. Degrees of freedom for each domain-level model were computed using a bi-factor model in which each subtask with more than one item was allocated to both a general factor and a residual factor. Degrees of freedom for the overall model were obtaining by combining the domain-level models, and assuming that the correlations among the four domains were freely estimated.

Table A1
Power analysis for factor analysis determining sample size of confirmatory sample.

Model	Degrees of freedom	Minimum sample size	Power at $N = 454$	Power at $N = 228$
Gross and Fine Motor	29	454	.800	.437
Social-Emotional	66	268	.979	.708
Literacy/ Numeracy	324	106	~1	.999
Overall	2483	37	~1	~1

Note: Sample sizes $N = 454$ and $N = 228$ denote the confirmatory and exploratory samples, respectively. See text for additional details.

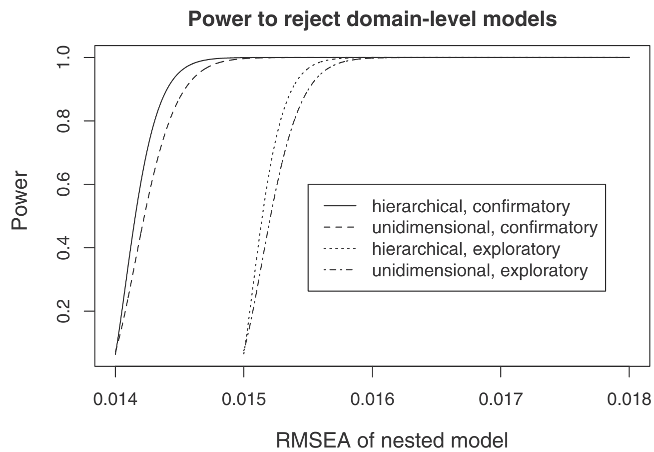

Fig. A1. Power curves for the chi-square difference test of overall IDELA models.

Minimum sample size was computed for $\alpha = .05$ and power = .80, and using RMSEA = .05 for the alternative distribution, values above were considered to represent poor model fit. We selected RMSEA = .02 as the value of the null distribution, values below which were considered to represent acceptable model fit. See MacCallum, Browne, and Sugawara (1996) for details on calculations and further discussion.

Based on the minimum sample size of the Motor domain, which had the fewest degrees of freedom, we randomly selected 454 children to be in the confirmatory sample. The remaining $N = 228$ observations were assigned to the exploratory sample.

A.2 Power of the tests of domain-level models

Fig. A1 reports the power curves for rejecting the Unidimensional Model and the Heirarchical Model, tested against the Unconstrained Model, in the exploratory and confirmatory samples. The RMSEA of the Unconstrained model was obtained from the estimated model in each sample (see Tables 3 and 5 of the main paper). See MacCallum, Browne, & Cai (2006) for details on calculations and further discussion.

A.3 Power of the tests of measurement invariance models

Fig. A2 reports the power curves for rejecting the Metric and Scalar Invariance, tested against the Configural model, for each of the three comparisons reported in the paper. The RMSEA of the Configural model was obtained from the estimated model, in each comparison (see Table 7 of main paper). See MacCallum et al. (2006) for details on calculations and further discussion.

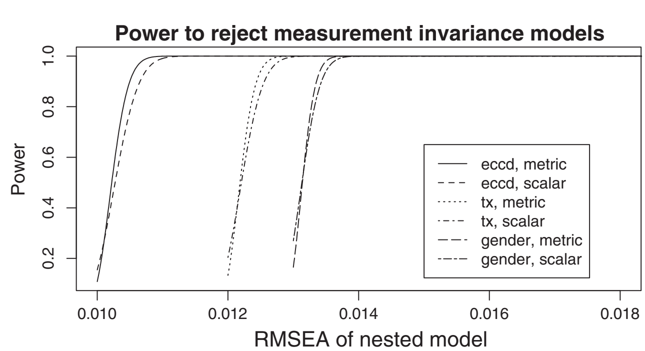

Fig. A2. Power curves for the chi-square difference test of measurement invariance models.

A.4 Appendix references

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130–149.

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, 11, 19–35.

Appendix B. Summary of item omissions and Heywood cases

As described in Section 2.3, we omitted the last 10 items on the Letter Identification and Number Identification subtasks, due to low response rates resultant from stopping rules. In the course of the analyses, a total of 5 additional items were omitted for reasons described below. In addition to the item omission, we also discuss how we dealt with Heywood cases when these occurred.

Omissions due to colinearity. Two items on the Draw a Human Figure subtask of the Motor domain were collinear with each other and with the other items on that subtask. Most children who were not able to draw a recognizable head for their figure were also not able to draw any other features of the figure. A second item asked if children were able to draw a second facial feature, which was collinear with drawing a head and drawing the first facial feature.

Omissions due to model misfit. A total of three items on three different subtasks were omitted due to model misfit in the exploratory analyses. Two items on the Social-Emotional domain were omitted because they had large residual correlations with other Social-Emotional items on different subtasks. One item on the Motor domain was omitted because it had large residual correlations with items on the Early Literacy and Early Numeracy domains. These omissions are described in more detail below.

Heywood cases. In the factor analysis literature, a "Heywood case" occurs when an item has a negative residual variance (see, Bartholomew et al., 2011, sec. 3.12). Two features of the present analysis are known to lead to Heywood cases: relatively small numbers of items per subtask, and relatively small sample size, the latter of which occurred in our subgroup analyses. Although there were no Heywood cases in the full sample at baseline or endline, when Heywood cases resulted in sub-group analyses, we used equality constraints with one or more additional items on the same subtask to address the issue. The rationale behind this approach is that using a simpler model ("partial tau–equivalence") ensures a valid solution for all parameters. A total of six items were problematic in one or more analyses.

References

Amente, A., Takele, M., Pisani, L., & Anis, K. (2015). *Ethiopia ELMI parenting endline*. Washington D.C: Save the Children.

Anders, Y., Rossbach, H. G., Weinert, S., Ebert, S., Kuger, S., Lehrl, S., & von Maurice, J. (2012). Home and preschool learning environments and their relations to the development of early numeracy skills. *Early Childhood Research Quarterly*, *27*(2), 231–244.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis* (3rd ed.). London: Arnold.

Bartlett, L., Dowd, A. J., & Jonason, C. (2015). Problematizing early grade reading: Should the post-2015 agenda treasure what is measured? *International Journal of Educational Development*, *40*, 308–314. http://dx.doi.org/10.1016/j.ijedudev.2014.10.002.

Black, M. M., Walker, S. P., Fernald, L. C., Andersen, C. T., DiGirolamo, A. M., Lu, C., ... Devercelli, A. E. (2017). Early childhood development coming of age: Science through the life course. *The Lancet*, *389*(10064), 77–90.

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, *78*, 647–663.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, *75*, 33–57.

Cameron, C. E., Brock, L. L., Murrah, W. M., Bell, L. H., Worzalla, S. L., Grissmer, D., & Morrison, F. J. (2012). Fine motor skills and executive function both contribute to kindergarten achievement. *Child Development*, *83*(4), 1229–1244.

Collaborative for Academic, Social, and Emotional Learning (2012). *2013 CASEL guide: Effective social and emotional learning programs, preschool and elementary school edition*. Chicago: CASEL.

Chavan, M., & Yoshikawa, H. (2013). *The future of our children: Lifelong, multi-generational learning for sustainable development*. New York, Paris, New Delhi: U.N., Sustainable Development Solutions Network.

Cueto, S., León, J., Miranda, A., Dearden, K., Crookston, B. T., & Behrman, J. R. (2016). Does pre-school improve cognitive abilities among children with early-life stunting? A longitudinal study for Peru. *International Journal of Educational Research*, *75*, 102.

Dowd, A. J., Borisova, I., Amente, A., & Yenew, A. (2016). Realizing Capabilities in Ethiopia: Maximizing Early Childhood Investment for Impact and Equity. *Journal of Human Development and Capabilities*, *17*(4), 477–493. http://dx.doi.org/10.1080/19452829.2016.1225702.

ECD Action Network (2017). *Task Force on Data, Measurement and Research: Preliminary report*Washington, DC: ECD Action Network.

Ethiopian Ministry of Education (2012). *Education statistics annual abstract 2011–12*. Addis Ababa, Ethiopia: Education Management Information System.

Ethiopian Government (2016). *Federal Democratic Republic of Ethiopia Government*. Retrieved from: http://www.ethiopia.gov.et/stateoromia.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Thousand Oaks, CA: Sage Publications.

Hasan, A., Hyson, M., & Chang, M. C. (2013). *Early Childhood Education and Development in Poor Villages of Indonesia: Strong Foundations, Later Success*. World Bank Publications.

Hood, M., Conlon, E., & Andrews, G. (2008). Preschool home literacy practices and children's literacy development: A longitudinal analysis. *Journal of Educational Psychology*, *100*(2), 252–271.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55.

Janus, M., & Offord, D. R. (2007). Development and psychometric properties of the Early Development Instrument (EDI): A measure of children's school readiness. *Canadian Journal of Behavioural Science*, *39*, 1–22.

Jennrich, R. I., & Bentler, P. M. (2012). Exploratory bi-factor analysis: The oblique case. *Psychometrika*, *77*, 442–454.

Kassahun, T., & Kedir, B. (2006). Girls' performance in mathematics in upper primary schools of Addis Ababa. *Indian Journal of Gender Studies*, *13*(3), 401–424.

Little, T. D., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives*, *7*(4), 199–204.

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, *11*, 19–35.

Martinez, S., Nadeau, S., & Periera, V. (2012). *The promise of preschool in Africa: A randomized impact evaluation of early childhood development in rural Mozambique*. The World Bank & Save the Children.

McClelland, M. M., Morrison, F. J., & Holmes, D. L. (2000). Children at risk for early academic problems: The role of learning-related social skills. *Early Childhood Research Quarterly*, *15*, 307–329. http://dx.doi.org/10.1016/S0885-2006(00)00069-7.

McCoy, D. C., Black, M., Daelmans, B., & Dua, T. (2016a). *Measuring population-level development in 0–3. Early childhood matters*. Netherlands: Bernard van Leer Foundation.

McCoy, D. C., Peet, E. D., Ezzati, M., Danaei, G., Black, M. M., Sudfeld, C. R., ... Fink, G. (2016b). Early childhood developmental status in low-and middle-income countries: National, regional, and global prevalence estimates using predictive modeling. *PLoS Medicine*, *13*(6), e1002034.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. New York: Routledge.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, *80*(392), 993–997.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177–196.

Muthén, B. O., & Muthén, L. (2014). *Mplus 7.3 [computer software]*. Los Angeles, CA: Muthén.

Office of Head Start (2010). *The Head Start child development and early learning framework*. Washington, DC: USDHHS, Office of Head Start. Retrieved May 22, 2017 from http://eclkc.ohs.acf.hhs.gov/hslc/tta-system/teaching/eecd/Assessment/Child%20Outcomes/HS_Revised_Child_Outcomes_Framework(rev-Sept2011).pdf.

Panter, J. E., & Bracken, B. A. (2009). Validity of the Bracken School Readiness Assessment for predicting first grade readiness. *Psychology in the Schools*, *46*, 397–409.

Pisani, L., Borisova, I., & Dowd, A. J. (2015). *International Development and Early Learning*

Assessment technical working paper. Washington, DC: Save the Children Available at: http://resourcecentre.savethechildren.se/library/international-development-and-early-learning-assessment-technical-paper.

Pisani,, L., Dyenka, K., Sharma, P., Chhetri, N., Dang, S., Gayleg, K., & Wangdi, C. (2017). Bhutan's national ECCD impact evaluation: Local, national, and global perspectives. *Early Child Development and Care*, *4430*(March), 1–18.

Rao, N., Sun, J., Ng, M., Becher, Y., Lee, D., Ip, P., & Bacon-Shone, J. (2014). *Validation, finalization and adoption of the East Asia-Pacific Early Child Development Scales (EAP-ECDS)*. New York: UNICEF.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the Testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, *47*(3), 361–372.

Sénéchal, M., & LeFevre, J. A. (2002). Parental involvement in the development of children's reading skill: A five-year longitudinal study. *Child Development*, *73*(2), 445–460.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling*. New York, NY, USA: Chapman Hall/CRC.

Snow, C. E. (2006). What counts as literacy in early childhood? In K. McCartney, & D. Phillips (Eds.), *Blackwell handbook of early childhood development*. Blackwell Publishing Ltd..

Snow, C. E., & Van Hemel, S. B. (2008). *Early childhood assessment: Why, what and how*. Washington, D.C: The National Academy Press.

Sörbom, D. (1989). Model modification. *Psychometrika*, *54*(3), 371–384.

Tafere, Y., & Woldehanna, T. (2012). *Beyond food security: Transforming the productive safety net programme in Ethiopia for the well-being of children*. Working Paper 83Oxford: Young Lives.

UNESCO (2013). *Toward universal learning: Recommendations from the learning metrics task force*UNESCO-UIS/Brookings Institution 2013 Available at: http://www.brookings.edu/~/media/research/files/reports/2013/09/learning-metrics-task-force-universal-learning/ltmf-recommendationsreportfinalweb.pdf.

United Nations (2015). *Transforming our world: The 2030 agenda for sustainable development*. New York: Author.

United Nations Development Programme (2015). *2015 Human Development Report* Retrieved from: http://hdr.undp.org/en/2015-report.

UNICEF (2013). *Statistics – Ethiopia*. Retrieved from: http://www.unicef.org/infobycountry/ethiopia_statistics.html.

UNICEF (2014). *The formative years: UNICEF's work on measuring early childhood development*. New York: UNICEF.

Woldehanna, T., Tafere, Y., Pankhurst, A., & Gudisa, R. (2011). *Understanding changes in the lives of poor children: Ethiopia*Country Report. Oxford: Young Lives.

Wolf, S., Halpin, P., & Yoshikawa, Y. (2017). *Evaluating the Factor Structure and Measurement Invariance of the International Development and Early Learning Assessment (IDELA) across Five Countries*Washington, D.C: World Bank Research Report.

Wuermli, A. J., Helm, J., Hastings, P., Yoshikawa, H., Wolf, S., Halpin, P., Dowd, A. J., Pisani, L., & Borisova, I. (2016). *The Home Environment and Early Child Development and Learning in Bhutan: Measurement and Associations (Dissertation)*. Retrieved from ProQuest Dissertations and Theses database. (No. 10247298).

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M., Espinosa, L., Gormley, W. T., et al. (2013). *Investing in our future: The evidence base on preschool*. Washington, D.C: ECD Action Network.

Yoshikawa, H., Leyva, D., Snow, C. E., Treviño, E., Barata, M., Weiland, C., ... Arbour, M. C. (2015). Experimental impacts of a teacher professional development program in Chile on preschool classroom quality and child outcomes. *Developmental Psychology*, *51*, 309.