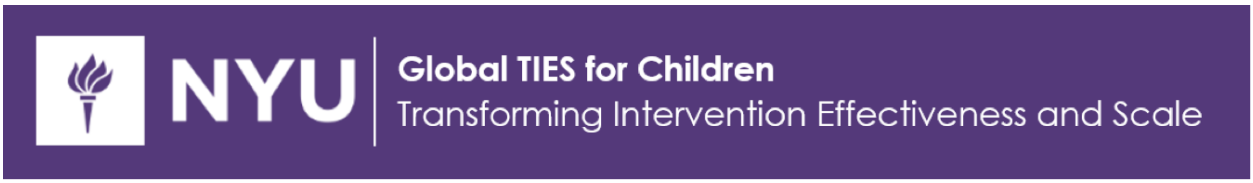


Assessing the Construct Validity of Save the Children's International Development and Early Learning Assessment (IDELA)

Sharon Wolf, Peter Halpin, Hirokazu Yoshikawa

Global TIES for Children, New York University

Lauren Pisani, Amy Jo Dowd

Save the Children

Ivelina Borisova

Formerly with Save the Children, now with UNICEF

Abstract

The International Development and Early Learning Assessment (IDELA) was developed by Save the Children to assess early childhood development holistically. It has been used in 30 mostly low- and middle-income countries. High-quality measurement plays a critical role in improving any outcome and has become an especially important for early childhood development given its recent inclusion in the Sustainable Development Gains (SDGs). Thus IDELA is drawing growing attention from the research and donor communities. This study examines the psychometric properties of the IDELA using exploratory and confirmatory bi-factor analyses to assess the structure and validity of the IDELA items that measure four domains of development (Early Numeracy, Early Literacy, Social-Emotional Development, and Motor Skills) across a regional sample in Oromia, Ethiopia. Results support the presence of a general factor for each of the four domains, with additional residual factors that correspond to individual task-based item groups. The relationships among the four domains are consistent with the hypothesis of a single over-arching construct, while each provides unique information about domains of children's development. These constructs are replicated in the confirmatory sample which provides reason to be optimistic that the IDELA measures (1) children's holistic development, and (2) four unique domains of development. Next steps are to consider the concurrent and predictive validity of the IDELA as well as to investigate measurement invariance across different subgroups within Ethiopia and across different countries to speak to the applicability and comparability of IDELA across countries.

Introduction

Early experiences form the foundation for future development (Shonkoff & Phillips, 2000). A number of studies have demonstrated that early developmental and learning-related skills are particularly important for children's transition and adaptation to school (e.g., Blair & Razza, 2007; Cueto et al., 2016; McClelland, Morrison, & Holmes, 2000). As a result of the growing evidence base, governments worldwide are realizing that the skills young children bring at the start of school are a major national issue, and there is increasing interest by governments in improving children's early skills and knowledge to increase their success in the early primary grades and beyond.

With the release of the Sustainable Development Goals in 2015 (United Nations, 2015), Target 4.2 under Education Goal 4 aims to ensure equitable access to "high quality early childhood development, care and education so that all children will be ready for primary school." In addition, the targets call for monitoring and reducing learning inequalities both within and across countries. Availability and use of high-quality and feasible measures of children's early learning skills are critical to achieving these targets. Thus, these international goals highlight the importance of developing assessments of early childhood development (ECD) and school readiness that are easily administered, conceptually and psychometrically validated across contexts, and are aligned with national monitoring systems. One of the key challenges to many other direct assessments is the length of time of administration, and thus a central issue in global assessments is to balance rigor and feasibility.

Efforts are being made by researchers to develop regionally valid assessments of early childhood development and learning, such as the Inter-American Development Bank's Regional Project on Child Development Indicators (PRIDI) in Latin America (Verdisco, Cueto, Thompson

& Neuschmidt, 2014) and the East Asia Pacific-Early Child Development Scales (EAP-ECDS) (Rao, Sun, Ng, Becher, Lee, Ip et al., 2014). There are also efforts to develop assessments that can be used internationally, such as the Early Development Instrument (EDI) (teacher-reported; Janus & Offord, 2007) and the UNICEF MICS Early Childhood Development Index (parent-reported; Bornstein, Britto, Nonoyama-Tarumi, Ota, Petrovic & Putnick, 2012). Most of these assessments cover multiple dimensions of children's development and learning outcomes. Some are direct assessments of children, while others are reported by caregivers or teachers. Given biases that can emerge from parent- or teacher-reports on children's outcomes, emerging best practices recommend direct assessments with children (Domitrovich, Gest, Jones, Gill & DeRousie, 2010; Mendive, Weiland, Yoshikawa & Snow, 2015), though for children under five years of age, multiple sources of evidence (including direct assessment, parent and teacher reports) are ideal (Snow & Van Hemel, 2008).

In some cases, psychometric analyses have been conducted to examine if the conceptual structure of a particular assessment holds empirically. For example, the Bracken School Readiness Assessment was statistically validated as a composite measure of children's school readiness in a U.S. sample of kindergarten children by assessing the measure's predictive and discriminant validity (Panter & Bracken, 2009). The Early Development Instrument (EDI) was analyzed on a sample of Canadian children using factor analytic methods, and found to measure six 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 less common in low- and middle-income countries. However, one effort to assess a contextually relevant scale is the East Asia Pacific–Early Child Development Scales (Rao et al., 2014). By assessing content validity,

internal consistency reliability, and item discrimination, the results of this study revealed that the seven developmental domains in the assessment were validated for use in six Asian countries. The domains include approaches to learning, cognitive development, cultural knowledge and participation, language and emergent literacy, motor development, health, hygiene, and safety, and social-emotional development.

The present study builds on this body of work to examine the measurement properties of the International Development Early Learning Assessment (IDELA), an international tool developed by Save the Children and designed to assess early childhood development and learning in low- and middle-income countries (Pisani, Borisova & Dowd, 2015). The assessment was designed with both rigor and feasibility at the forefront. This study uses a bi-factor analytic approach to assess the empirical structure of the IDELA, which builds on previous measurement analyses conducted by Pisani and colleagues (2015). By addressing the construct validity of the assessment using a rigorous methodological approach, this study advances the feasibility of measuring ECD and school readiness globally, as well as comparing the impacts of program and policy interventions across countries.

Assessing School Readiness Requires a Multi-dimensional View of Learning and Development

In a seminal study, Grantham-McGregor and colleagues (2007) estimated that 219 million children under the age of five were not reaching their developmental potential in 2004, using data on stunting and poverty. Although these two indicators are critical predictors of children's short- and long-run well-being, they only partially explain the variation in children's readiness to learn. Young children entering school require a broad array of behaviors and skills in order to succeed. For example, children who are emotionally "well-adjusted" and can regulate their emotions have

a significantly greater chance of early school success, while children who experience serious emotional difficulty face an increased risk of early school difficulty (Blair, 2002; Raver, 2002). A panel of experts commissioned by the National Academy of Sciences differentiated and discussed five key domains of early childhood development based on current theory and research: physical well-being and motor; socioemotional; approaches to learning; language and emergent literacy, and cognitive skills, including early numeracy (Snow & Van Hemel, 2008; pp. 58). These domains were also included in a global framework of learning from the Learning Metrics Task Force, recommending that all children and youth develop competencies across these domains, as well as science/technology and culture/arts (UNESCO, 2013).

Thus, "school readiness" can be defined broadly as an outcome of the early years that covers multiple dimensions of development, including early academic knowledge, behavioral skills, social-emotional development, and aspects of physical health including motor development, and feasible measures of these domains are needed to understand the current state of early childhood development worldwide. In addition, children progress developmentally through different levels of thinking (Clements & Sarama, 2007), and thus while children may meet a cut-off specification for "readiness", some children may be further along in this progression than others, and this may vary across domain.

While several studies have assessed how behavioral development predicts academic outcomes in early childhood education, most of these studies have not considered multiple dimensions of school readiness (including behavioral and early academic skills) simultaneously. A study using six nationally representative samples in the United States did attempt to do this, and identified that within three key elements of school readiness—school-entry academic, attention, and socioemotional skills— the strongest predictor of later achievement were school

entry math skills, followed by reading skills, and children's attention capacities. Early social-emotional skills and behaviors, on the other hand, were not significant predictors of later academic outcomes (Duncan et al., 2007).

The Need in Low- and Middle-income Countries: Establishing Global Metrics

While school readiness and its measurement have received quite a bit of attention in the developmental and educational literature in high-income countries, conversations are only now starting on similar issues in low- and middle-income countries. The ability to measure domains of school readiness and their association with children's academic outcomes over time in LMICs has been limited due to a lack of validated measures and available data. There is a need for monitoring assessments that capture multiple domains of early childhood development and school readiness skills, are easily and feasible to administer, and can be used and compared across regions, countries, and contexts. In addition, there is a need for an assessment that is sensitive enough to be used for program evaluation so that the various efforts of organizations and governments to improve early childhood outcomes can be assessed and compared. The IDELA is an assessment that was developed to address these needs.

History and Background of the IDELA

In the ECD field, there are few international assessment tools available that can be used to holistically measure children's development and emergent skills. In 2011, Save the Children completed a comprehensive review of the existing child development assessments and documented a number of important limitations with existing assessments. Many of the assessments available were limited in their approach, either targeting only one skill area or a specific age group and many were reliant on parent or teacher report rather than directly assessing children's skills. Further, costs associated with using the assessments across countries

or projects were a concern as many required special permission and purchase. Most importantly, the majority of existing assessments had been used primarily in high-income countries, such as the United States, United Kingdom and Australia, making them difficult to adapt and use across countries with diverse populations and resource-poor settings.

Save the Children's review concluded that despite the existence of early childhood care and development (ECCD) assessments in the global space, none of the assessments available at the time offered a balance between (1) international applicability, especially within low and middle income country contexts, (2) feasibility and ease of administration and adaptation, and (3) supporting psychometric research. With these criteria in mind, and lessons learned from years of early childhood programming, Save the Children began the process of developing and validating the International Development and Early Learning Assessment (IDELA). The initial set of items was 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 (EDI), among other assessments.

The goal for IDELA was to develop a holistic, rigorous, open source assessment that is feasible and easily adapted to different national and cultural contexts. IDELA was developed with an aim to support continuous program improvement across Save the Children's and partners' numerous country sites, to increase accountability among ECD initiatives globally, and to offer comparable and ongoing data and evidence about children's learning and development across countries that can help governments and global actors to bring successful ECD programs to scale. Pisani and colleagues (2015) describe aspects of IDELA that make the tool feasible, reliable, and adaptable in varied, low-resource settings, as well as the testing done to ensure rigorous implementation across sites. Unlike many ECD assessments developed in wealthier

contexts, IDELA does not need numerous or expensive materials for administration, but rather minimal materials that can be locally sourced. This both decreases the financial burden of the assessment, and increases the local relevance of the tool.

Further, IDELA does not require specially trained professionals and can therefore engage local teachers, government officials, university students, community organizers and others in administration. While priority is typically given to individuals who have previous experience working with young children, no formal training is required. Enumerators in Malawi, for example, were a mix of men and women from the local community and all testing locations were rural. Enumerators in Egypt were women working as community organizers for local NGOs or for the local government, and the testing occurred in rural and semi-urban areas. To enable reliable data collection, training on IDELA typically lasts for four to five days and includes in-office exercises and hands-on field training. Enumerators first practice using IDELA with each other in a controlled setting, and then in pilot testing locations with young children in communities similar to those that will be included in the study sample. IDELA teams document inter-rater reliability by systematically having 10 percent of children in the overall sample scored by two enumerators simultaneously. Inter-rater reliability results from data collected at sites the in Malawi and Egypt described above found that the intra-class correlation was “excellent” (>.75; Fleiss, 1986) across all domains in both sites.

To date, Save the Children has used the IDELA in over 20 countries. In Ethiopia specifically, the IDELA has been used for a number of years across multiple regions and samples. The data for this study come from Ethiopia.

Sample Context and Background on Ethiopia

Ethiopia is a country in the horn of Africa with a population of close to 100 million. 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 ranked in the bottom tier, number 174 out of 188 countries[1]. Currently, the Federal Democratic Republic of Ethiopia is implementing "Early Childhood Care and Education (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 21.6% in 2011-12 academic year[2]. Though this government program is very ambitious, pre-school education is marred by many challenges such as lack of trained and independent facilitators/teachers, unavailability of curriculum and guidelines, lack of adequate center facilities, developmentally appropriate learning materials, play grounds and lack of incentives/salary for teachers assigned for this program, among others (Dowd, Borisova, Amente & Yenew, under review).

In 2012, Save the Children began piloting a toolkit 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 Development (ECCD) in the regions of Tigray and Oromia through its child sponsorship funding. The sample for this study comes from the baseline data collected as part of a larger intervention evaluation of the ELM program in Ambo and Dendi of West Showa district, Oromia where SC supported 36 ECCD centers. Of these, 19 were community-based, built in partnership with communities and 17 are government-run centers, located on the grounds of primary schools. One trained adult at these centers engaged children in structured and unstructured play, using learning and teaching materials provided by the program. The children attended the centers 5 days a week, on average 3 hours per day.

[1] http://hdr.undp.org/en/composite/HDI
[2] Ethiopian Ministry of Education, 2012. Education Statistics Annual Abstract 2011-12.

The Current Study

The IDELA consists a set of play-based items administered through groups of tasks (or, subtasks), covering five domains of development. *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., Personal Awareness; Comparison by Size and Length). A previous technical report addressed the internal consistency reliability, inter-rater reliability, and construct validity of the IDELA (Pisani et al., 2015). This study adds to this measurement research by examining a series of exploratory and confirmatory factor analyses assessing the internal structure of the IDELA.

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, Social-Emotional Learning)[3]. The research questions addressed in this report are as follows:

1. Are the items of the IDELA currently used to measure each of the domains – motor, social-emotional, early literacy, and early numeracy – 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 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 measurement model that accounts for the subtask structure of the IDELA.

[3] Note that full instrument includes an additional 15 items that are administrator reported, and also an additional domain for Self-Regulation. See the Measures section below for more details on the items used in this analysis.

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, are the correlations among the domains compatible with the hypothesis of a higher-order factor?

Methods

Participants

Data for this study come from the baseline sample of an evaluation study of the Emergent Literacy and Math Intervention (ELMI) program. Children (N = 682) were sampled from 36 Early Childhood Care and Development (ECCD) centers and neighboring villages without centers in Ambo and Dendi Districts of West Showa in the Oromia Region. Children were 5.9 years of age, on average (SD = .40, range 4-7) and 52% female. At baseline, about three-quarters of the sample (N = 519; 76.1%) was enrolled in center-based early childhood education center.

Measures

All children were assessed using the IDELA assessment. The two subtasks assessing children's self-regulation skills were excluded given that this developmental domain in the assessment has only two subtasks which are considered complementary/supplementary items. Thus, four domains of development are assessed: Gross and Fine Motor, Social-Emotional, Early Literacy, and Early Numeracy.

Analytic Plan

At the domain level, we assessed whether the items used to measure each of the 4 domains were consistent with the hypothesis of a single underlying construct or factor. Additionally, we sought to address the subtask structure of the items using a bi-factor approach (see Rijmen 2010, for discussion). At the level of the overall assessment, we wanted to assess

the specificity of the items (i.e., whether any items loaded onto domains other than their target domain), and also the relationship among domains.

Data: Exploratory and confirmatory samples. Before conducting any analyses, we randomly divided the $N = 682$ observations into two halves: 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 and for the overall assessment. The purpose of the confirmatory sample was to ensure that the proposed models demonstrated out-of-sample generalizability.

A sample size of $N = 365$ for the confirmatory sample was determined by conducting power analyses using a bi-factor model at the domain level. The power analyses were conducted at the domain level to ensure that the conclusions about the individual domains, and not only the overall assessment, were based on sufficient sample size. The power analyses are summarized in Table 1 (see Preacher & Coffman, 2006, for details on calculations). After randomly selecting 365 children to be in the confirmatory sample, the remaining sample size of $N = 317$ observations were assigned to the exploratory sample.

Table 1. Power Analysis for Determining Sample Size of Confirmatory Sample.

Model	Degrees of Freedom	Minimum Sample Size
Gross and Fine Motor	25	362.5
Social-Emotional	63	180.5
Early Literacy	493	49.2
Early Numeracy	663	41.6
Overall	4838	14.7

Note: Degrees of freedom for each domain-level model were computed using a bi-factor model. Degrees of freedom for the overall model were obtaining by combining the domain-level models. Item thresholds were not included in

the degrees of freedom calculations. Minimum sample size was computed for α = .05, power = .80, *RMSEA* = .05 for the null distribution, and *RMSEA* = .08 for the alternative distribution. The minimum sample size from the power analysis is not intended to reflect the number of observations needed for estimation of model parameters.

Step 1: Exploratory factor analysis within domains. For each domain, we conducted three exploratory analyses. The first was to fit a unidimensional factor model to the items of each domain, but without modeling the subtask structure. By examining overall goodness of fit of the Unidimensional Model, this allowed us to assess the need for a bi-factor model. By examining targeted misspecification indices ("modification indices"), we could also assess whether any deviations from the single factor model were consistent with the subtask structure of the IDELA. Second, we conducted an exploratory factor analysis with a bi-factor rotation (Jennrich & Bentler, 2011, 2012). We ran a series of bi-factor models, varying the number of residual factors, to identify the best-fitting model. We also examined the factor pattern of the bi-factor rotations for each number of residual factors, to see these were consistent with the subtask structure of the IDELA.

The first two analyses were intended to provide complementary sources of information for formulating a "proposed model" for each domain. In the third exploratory analysis, we tested the proposed models. These are the models we present below.

Step 2: Exploratory factor analysis across domains. Having conducted exploratory analyses for each domain, the next step was to analyze all items simultaneously by combining the proposed models from step 1. We fit a total of three models to the full IDELA assessment, again using the exploratory sample. These three models differed in terms of how we modeled the correlations among the factors representing the domain-level constructs.

The first model simply combined the domain-level models without placing any restrictions on the correlation matrix of the factors. By examining overall model fit and targeted

model misspecification indices, this allowed us to assess whether the any items loaded on more than one domain. This first model also provided a nesting model for chi-square difference testing of the two following models. We therefore refer to it as the "Unconstrained Model." The second 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 an over-arching construct. We refer to this as the "Hierarchical Model." The third model replaced the four domain-level factors with a single factor. This model tests whether the four domains were really providing unique information, or whether IDELA *only* measures a single over-arching construct. We refer to this as the "Undiminesional Model."

As mentioned, the second and third models are nested within the first. This allows for χ^2 difference testing of each nested model against the Unconstrained Model. If the χ^2 difference test is statistically significant, then the Unconstrained Model fits the data significantly better and therefore it should be favored over the nested model. If the test is not significant, then the nested model fits the data just as well as the Unconstrained Model, and the nested model can be preferred on a basis of parsimony (i.e., because it has less parameters). This approach provides a more rigorous test of the Hierarchical Model and the Unidimensional Model than using goodness of fit indices alone.

Step 3: Confirmatory model. The final analytic step involved assessing the out-of-sample generalizability of the models established in Steps 1 & 2 using the confirmatory sample.

Results

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 were conducted using the *diff-test* module.

Step 1: Exploratory Factor Analysis Within Domains. The proposed model for each domain is summarized in Figures 1-4, which reports the factor loadings on the general factor as well as the subtask structure (defined above) used for the subtask. Table 2 summarizes the goodness of fit of the proposed models for each of the domains. For the sake of brevity we omit details of the preliminary exploratory analyses that led to the proposed models.

Table 2. Summary of Goodness of Fit for Proposed Models at the Domain Level: Exploratory Sample.

Domain	χ^2 (df)	RMSEA (90% CI)	TLI
Gross and Fine Motor	54.98 (27)	.057 (.035, .079)	.993
Social-Emotional	157.57 (74)	.060 (.047, .073)	.934
Early Literacy	1037.02 (649)	.043 (.038, .048)	.992
Early Numeracy	390.06 (692)	.033 (.027, .038)	.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.

Motor development. The domain of motor development consists of 10 items, and 4 item-clusters. Results from the bi-factor analysis, the final model selected, indicated strong evidence of a general factor for motor development (see Table 2). Beyond one general factor, there was no indication of any residual factors. Notably, one item – HUMAN1 – had to be removed from the final bi-factor model because it resulted in computational problems (a "Heywood case," due to its very high correlations the other items on the "human" subtask.). This may indicate a general concern with how the human subtask is scored, or it may be due to

sampling error. We can address this issue in the replication study. The proposed model is shown in Figure 1.

Social-emotional development. The domain of social-emotional development consists of 14 items, and 5 item-clusters. One item-cluster consists of only 1 item. Results from the bi-factor model, the final model selected, provided strong evidence of a general factor for social-emotional development, with three residual factors: *empathy, conflict/problem solving*, and *emotion identification* (see Table 2). The subtasks with residual factors did not have sufficient items to create a bi-factor model (3 items or more are required per subtask). Instead, residual covariances were added rather than a residual factor: empathy2 and empathy3; conflict1 and conflict2; and emotion1 and emotion4. The final model is shown in Figure 2.

Early literacy. The domain of early literacy consists of 38 items and 6 item-clusters. Results from the bi-factor model, the final model selected, indicated strong evidence of a general factor for early literacy, with three residual factors (see Table 2). There was evidence of residual covariation for five of the item-clusters (the exception was letter identification). To address this, two residual factors (word pairs and oral comprehension), as well as three additional sets of residual covariances (print awareness, writing level, and expressive vocabulary), were included. The final model is shown in Figure 3.

Early numeracy. The domain of early numeracy consists of 39 items and 8 item-clusters. Two item-clusters consist of only 1 item. Results from the bi-factor model, the final model selected, indicate provide strong evidence of a general factor for early numeracy, with three residual factors (see Table 2). There was evidence of residual covariation of three item-clusters: size1-size4, one-to-one correspondence, and addition/subtraction. The final model is shown in Figure 4.

Step 2: Exploratory Factor Analysis Across Domains

The goodness of fit for the three overall models is reported in Table 3. Figures 5 and 6 show the parameter estimates for the Unconstrained Model and the Hierarchical Model, in the exploratory sample.

Table 3. Summary of Goodness of Fit for the Overall IDELA Model: Exploratory Sample

Model	χ^2 (df)	RMSEA (90% CI)	TLI	χ^2 -diff (df)	*p*-value
Unconstrained	5310.32 (4815)	.018 (.015, .021)	.982	NA	NA
Hierarchical	5312.57 (4817)	.018 (.015, .021)	.982	5.67 (2)	.059
Unidimensional	5717.91(4821)	.024 (.020, .027)	.967	140.66 (6)	< .001

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; the p-value is reported in the last column.

Unconstrained Model. After the domain specific models were established, the next step in the analysis involved fitting all of the items in one model. As shown in the first row of Table 3, the Unconstrained Model for all items showed very good overall fit to the data. Examination of modification indices did not indicate that any items were cross-loading on factors other than their target domain. As shown in Figure 5, the four domain-level factors had correlations ranging from .60 to .85.

Hierarchical Model. The second row of Table 3 indicates that the Hierarchical Model also showed good fit to the data overall. Importantly, the chi-square different test did not lead to a rejection of this model when tested against the Unconstrained Model. Thus we conclude that the relationships among the four domains are consistent with the hypothesis of a single higher-order factor. As shown in Figure 6, the factors loadings on the higher order factor ranged from .67 to .95.

Unidimensional Model. The final line of Table 3 shows the goodness of fit of the unidimensional factor fitted to all 101 items (while preserving the subtask structure established in Step 1). Although the overall fit of the model is acceptable, the chi-square difference test reveals that this model did not fit the data as well as the Unconstrained Model. We therefore reject the hypothesis that a single construct was sufficient to explain the relationships among all items on the IDELA. Because the model was rejected, we do not present its parameter estimates.

Step 3: Confirmatory Analyses

As final step, we consider the out-of-sample generalizability of the results reported in Steps 1 & 2 using the confirmatory sample.

Factor Analysis within Domains: As shown in Table 4, all of the domain level models were replicated with acceptable goodness of fit in the confirmatory sample.

Table 4. Summary of Goodness of Fit for Proposed Models at the Domain Level: Confirmatory Sample.

Domain	χ^2 (df)	RMSEA (90% CI)	TLI
Gross and Fine Motor	77.25 (27)	.071 (.053, .090)	.976
Social-Emotional	167.34 (74)	.059 (.047, .071)	.949
Early Literacy	1102.71 (649)	.044 (.039, .048)	.993
Early Numeracy	846.78 (692)	.025 (.018, .030)	.995

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.

As with the exploratory sample, we again omitted HUMAN1 due to a Heywood case. The replication of the Heywood case a second sample suggests that the item is problematic; again it was correlated very highly with other items on the HUMAN subtasks. Additionally, one of the oral comprehension items (ORALCOMP3) also resulted in a Heywood case in the

confirmatory sample. We attempted to address this in several ways: (a) by fixing the residual variance of the item; (b) by constraining the residual factor loadings of all oral comprehension items to be equal; and (c) by omitting the item from the residual factor. In the first case the model did not converge, and in the latter two cases the negative residual variance persisted. Therefore it was necessary to omit ORALCOMP3 from the confirmatory analyses. Further investigation is required to assess the root of this problem.

Factor Analysis Across Domains. Table 5 reports the goodness of fit and the Unconstrained Model and the Hierarchical Model in the confirmatory sample. Overall, the results closely parallel the exploratory data analyses. There was no evidence of crossloading items in the Unconstrained Model, and the chi-square different test did not lead to a rejection of the Hierarchical Model.

Table 5. Summary of Goodness of Fit for the Overall IDELA Model: Confirmatory Sample

Model	χ^2 (df)	RMSEA (90% CI)	TLI	χ^2 -diff (df)	*p*-value
Unconstrained	5225.80 (4718)	.017 (.014, .020)	.987	NA	NA
Hierarchical	5227.90 (4720)	.017 (.014, .020)	.987	5.30 (2)	.071

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; the p-value is reported in the last column.

Unconstrained Model. Again the Unconstrained Model showed very good overall fit to the data Figure 7 shows the correlations among the domains in the confirmatory sample, which ranged from .67 to .87.

Hierarchical Model. As with the exploratory sample, the Hierarchical Model showed good fit overall as shown in Figure 8, the factors loading on the higher order factor ranged from .78 to .94.

Discussion

We first describe what was learned empirically about the IDELA in terms of the empirical support for its conceptual structure. Second, we turn to additional analyses that should be taken into consideration based on the findings in this paper. Finally, we discuss the implications for use of the IDELA tool in global metrics of early childhood development, both low- and middle-income countries and in high-income countries, and its use for program impact evaluations.

The IDELA Structure

The IDELA aims to measure children's school readiness holistically—both by measuring unique domains of development, and by measuring young children's overall school readiness. The existing IDELA structure in this analysis consisting of 4 domains, 24 subtasks, and 101 items was found to have very good fit with the data collected in Ethiopia. Moreover, the four domains are consistent with the statistical hypothesis of a single over-arching construct–which we call school readiness–while each provides unique information about children's development. The structure reflects constructs 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 (Panter & Bracken, 2009; Janus & Offord, 2007) and in East Asia and the Pacific (Rao et al., 2014). The fact that these constructs are replicated in a sample of children in Ethiopia provide good reason to be optimistic that the IDELA measures (1) children's holistic development, and (2) four unique domains of development, that resonate across varied settings.

While the four constructs measured in the IDELA are related to one another and also to a higher-order construct, they remain distinct. In particular, replacing the four constructs with a single construct did not provide acceptable fit to the data. This again provides supportive evidence for a four-factor approach distinguishing these domains of early childhood learning and development. Finally, the results from the exploratory analysis were replicated and confirmed, with the final model showing strong empirical support for the conceptual structure of the IDELA.

Implications for Use in Early Childhood Development Globally

The several-year process of developing the IDELA has resulted in a rigorous, holistic, feasible and valid international assessment for measuring early learning and development of children 42 to 78 months. With 24 subtasks and an average of 30 minutes per child to complete the assessment, the IDELA is one of the shortest assessments that provides a holistic picture of learning and development in the early years. Importantly, IDELA has not been validated at the diagnostic level for screening individual children for developmental delays in any of its skill domains, and is not intended as a screening tool. Currently, IDELA does not set thresholds for "school readiness" and an IDELA score in a certain range cannot be taken as indicative of a serious problem. It's most important contribution globally is its ability to measure holistic early learning and development at the group or population level.

To date, the assessment has been used successfully in over 20 countries to measure children's school readiness, assessing the effectiveness of different program approaches, and in program evaluations. Further work to document IDELA's concurrent and predictive validity will bolster the evidence of its value in contributing to dialogue both on school readiness as well as successful transition into and through the first grades of primary schooling.

Next Steps in Analyzing the IDELA Factor Structure

There are two primary additional analytic areas for consideration to identify the optimal items for administration of the IDELA. First, most of the IDELA items have good reliability and a large portion of the item variance (30% or more) is explained by the respective domain factor. However, in this sample, there are eight items that have relatively low reliability (see Table 1). Moving forward, it would be useful to conduct similar analyses in other databases to consider eliminating or replacing low reliability items.

The second issue concerns the letter identification and number identification subtasks. The items on both subtasks are highly reliable in terms of the amount of variance the items explain in the respective subscale. However, reducing their number, either during the administration of the assessment (e.g., removing some items, or adaptively administering a subset of items to each child) or during the scoring procedure (e.g., not applying all items to obtain students' scores; down-weighting the entire subtask), should be considered. Reasons for this include the following: (a) the subtasks account for a large proportion of the items on their respective subscales and in this sense dominate the interpretation of the constructs; (b) some items have very high correlations, perhaps indicative of stopping or skipping rules, which is in turn indicative that too many items are being administered to some children; (c) preliminary IRT analyses (see Figures 3 and 4) indicate that a sizeable portion of the items on both subscales have the same level of difficulty, and in this sense are redundant. However, decisions would need to be made based on the alphabet system used in each respective country that use the Latin/Roman alphabet, as well as in those that use different alphabets or orthographic representation. This is an area future research could explore.

There are two immediate primary next steps for future analyses with IDELA. The first

will consider the concurrent and predictive validity of the IDELA factor scores using data from children's caregiver reports on the home environment. The second will also consider measurement invariance analyses to assess if the factor structure of the assessment is similar across different subgroups within Ethiopia (e.g., children in center care versus home care), and across different countries. The findings will speak applicability and comparability of the assessment across 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, as well as Ali Amente, Alene Yenew and Motuma Takele who led the data collection teams in Ethiopia. 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.

References

Blair, C. (2002). School readiness: Integrating cognition and emotion in a neurobiological conceptualization of children's functioning at school entry. *American Psychologist, 57*, 111 – 127.

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.

Bornstein, M. H., Britto, P. R., Nonoyama-Tarumi, Y., Ota, Y., Petrovic, O., & Putnick, D. L. (2012). Child development in developing countries: Introduction and methods. *Child Development, 83*, 16-31.

Clements, D. H., & Sarama, J. (2007). Effects of a preschool mathematics curriculum: Summative research on the Building Blocks project. *Journal for Research in Mathematics Education, 38*, 136-163.

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*, 102–114.

Domitrovich, C., Gest, S., Jones, D., Gill, S., & Sanford DeRousie, R. (2010). Implementation quality: Lessons learned in the context of the Head Start REDI trial. *Early Childhood Research Quarterly, 25*, 284-298.

Dowd, A.J., Borisova, I., Amente, A. & Yenew, A. (under review). Realizing Capabilities in Ethiopia: Maximizing Early Childhood Investment for Impact and Equity. *Journal of Human Development and Capabilities.*

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

Fleiss, J. L. (1986). Reliability of Measurement. *The Design and Analysis of Clinical Experiments.* John Wiley & Sons, Inc., Hoboken, NJ, USA.

Grantham-McGregor, S., Cheung, Y. B., Cueto, S., Glewwe, P., Richter, L., Strupp, B., & International Child Development Steering Group. (2007). Developmental potential in the first 5 years for children in developing countries. *The Lancet, 369*, 60-70.

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. (2011). Exploratory bi-factor analysis. *Psychometrika*, *76*, 537-549.

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

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. doi:10.1016/S0885-2006(00)00069-7.

Mendive, S., Weiland, C., Yoshikawa, H., & Snow, C. (2015). Opening the black box: Intervention fidelity in a randomized trial of a preschool teacher professional development program. *Journal of Educational Psychology, 108*, 130-145.

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

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

Preacher, K. J., & Coffman, D. L. (2006, May). Computing power and minimum sample size for RMSEA [Computer software]. Available from http://quantpsy.org/.

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). UNICEF. Retrieved (11 February, 2016) from: [http://www.arnec.net/wp-content/uploads/2015/07/EAP-ECDS-Final-Report1.pdf].

Raver, C.C. (2002). *Emotions matter: Making the case for the role of young children's emotional development for early school readiness*. Social Policy Reports of the Society for Research in Child Development.

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. doi:10.1111/j.1745-3984.2010.00118.x

Shonkoff, J. P., Boyce, W. T., & McEwen, B. S. (2009). Neuroscience, molecular biology, and the childhood roots of health disparities. *JAMA*, 301, 2252-2259.

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

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.

Verdisco, A., Cueto, S., Thompson, J., & Neuschmidt, O. (2014). Urgency and possibility: First initiative of comparative data on child development in Latin America. *Inter-American Development Bank.* Washington, D.C. Available at https://publications.iadb.org/handle/11319/6849?locale-attribute=en&

Figure 1. Diagram of the final exploratory model of the motor development domain

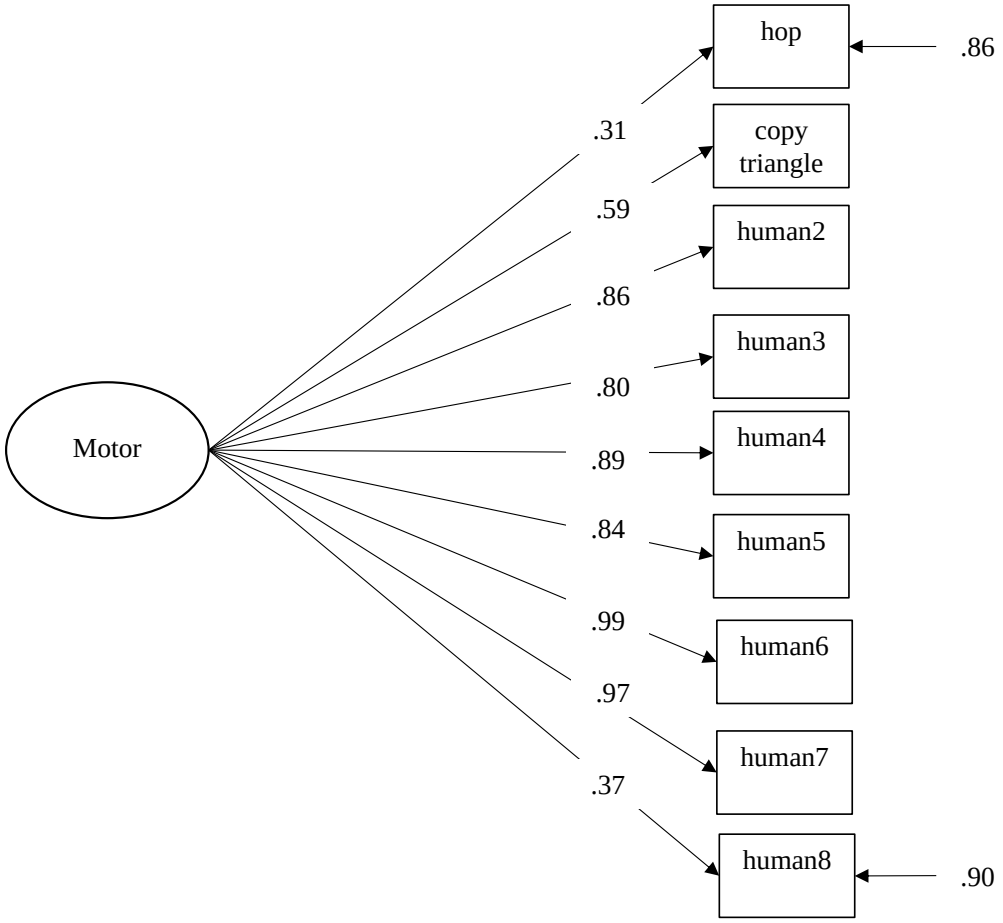


Figure 2. Diagram of the final exploratory model of the social-emotional development domain

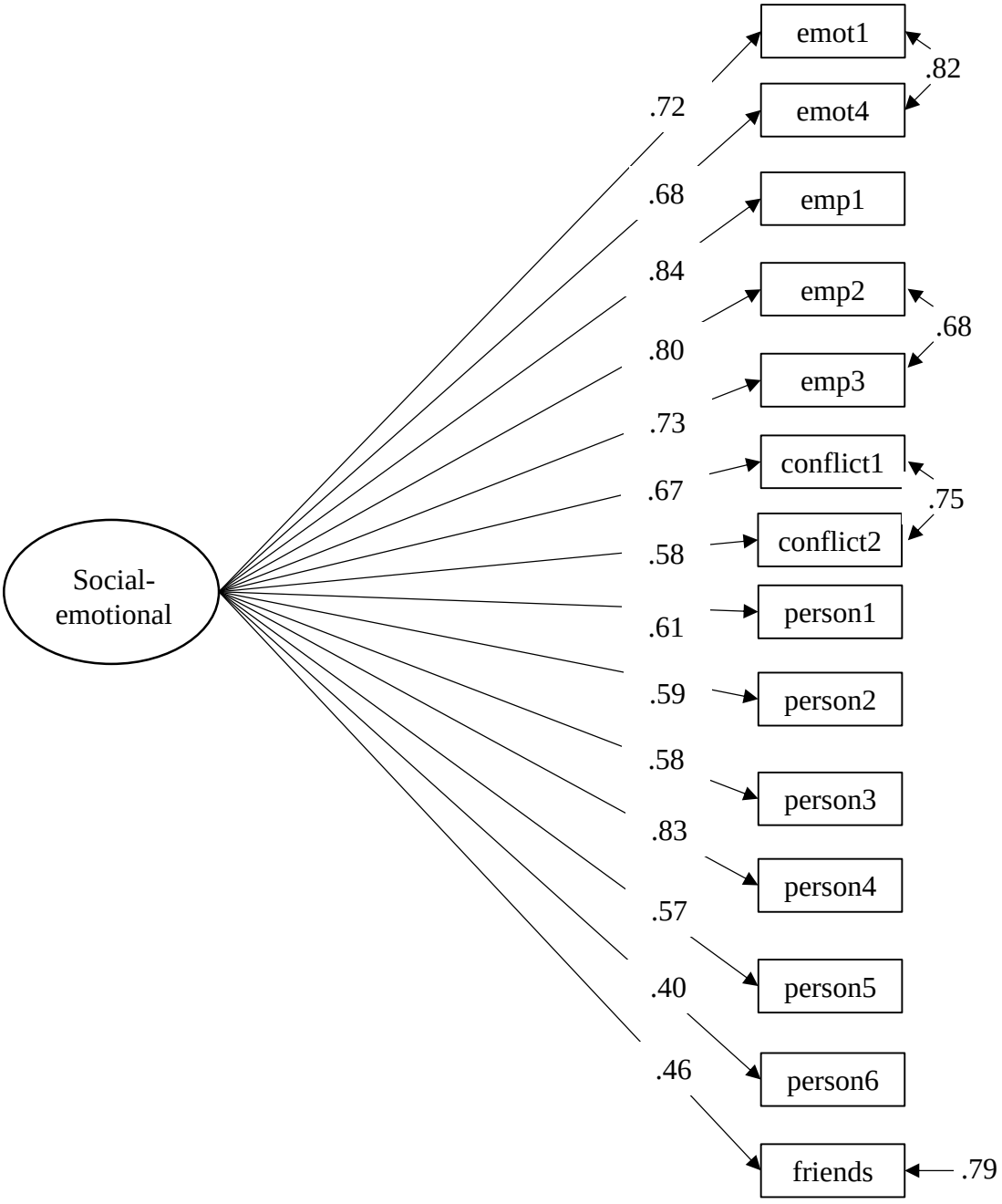


Figure 3. Diagram of the final exploratory model of the early literacy domain

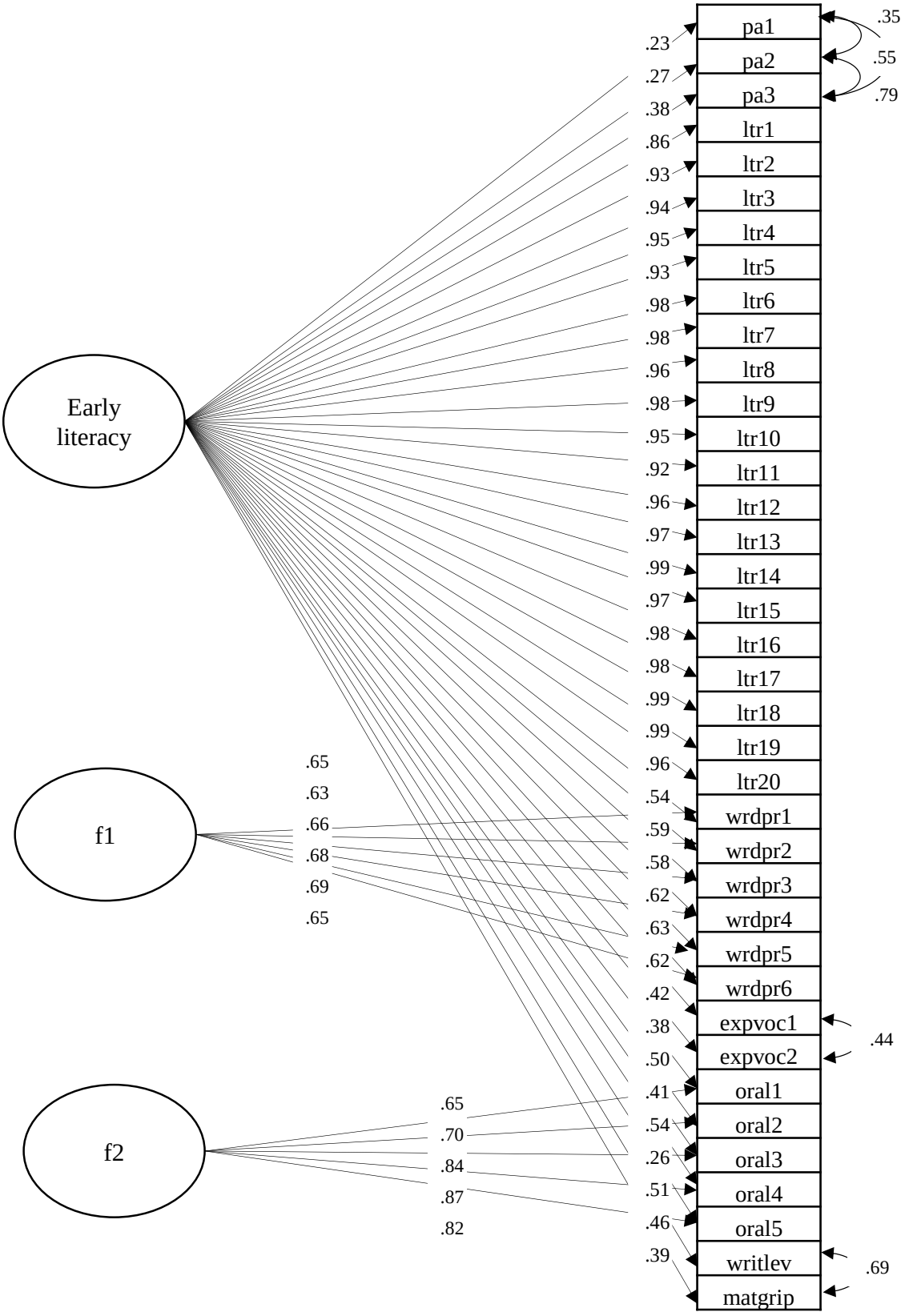


Figure 4. Diagram of the final exploratory model of the early numeracy domain

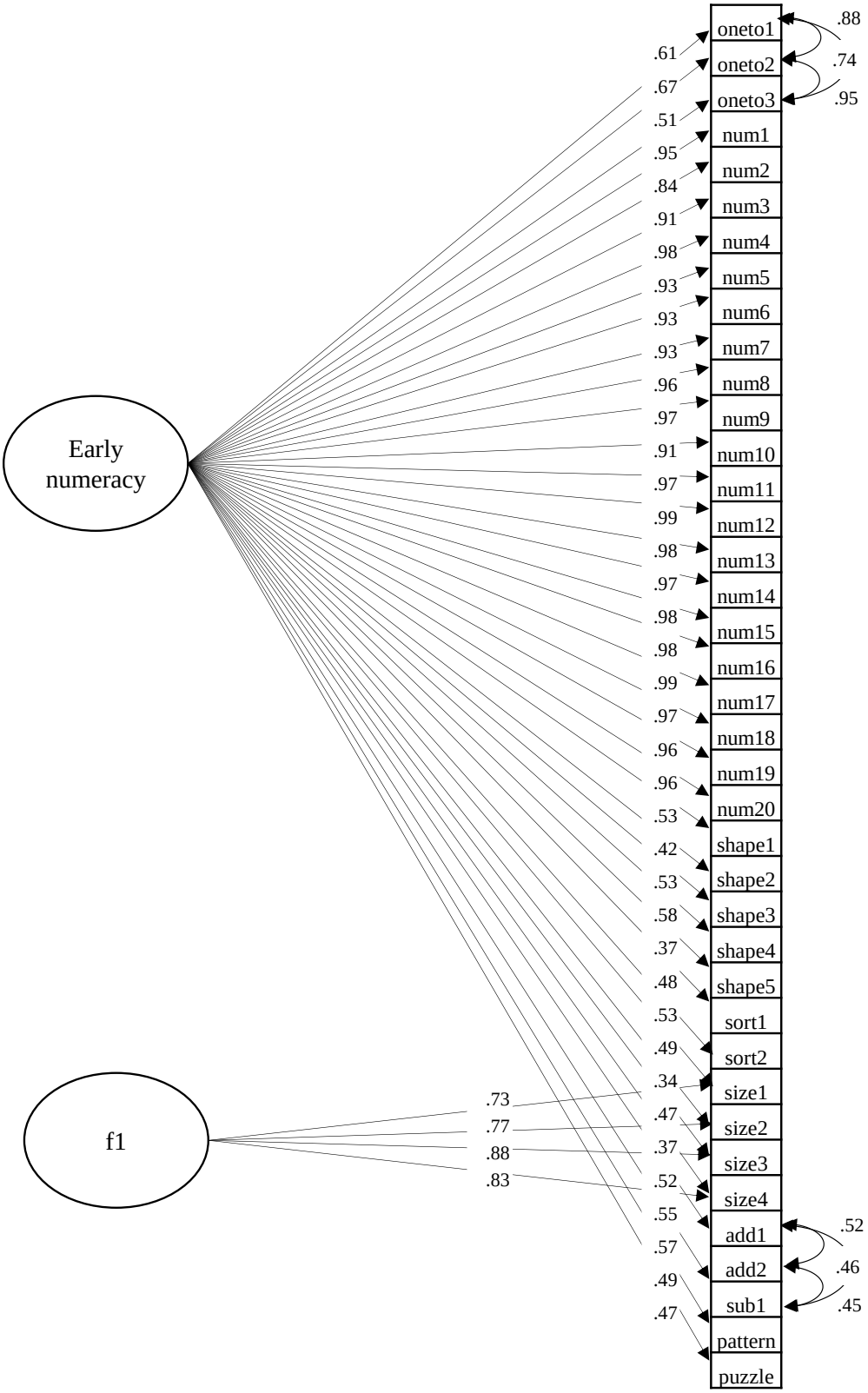
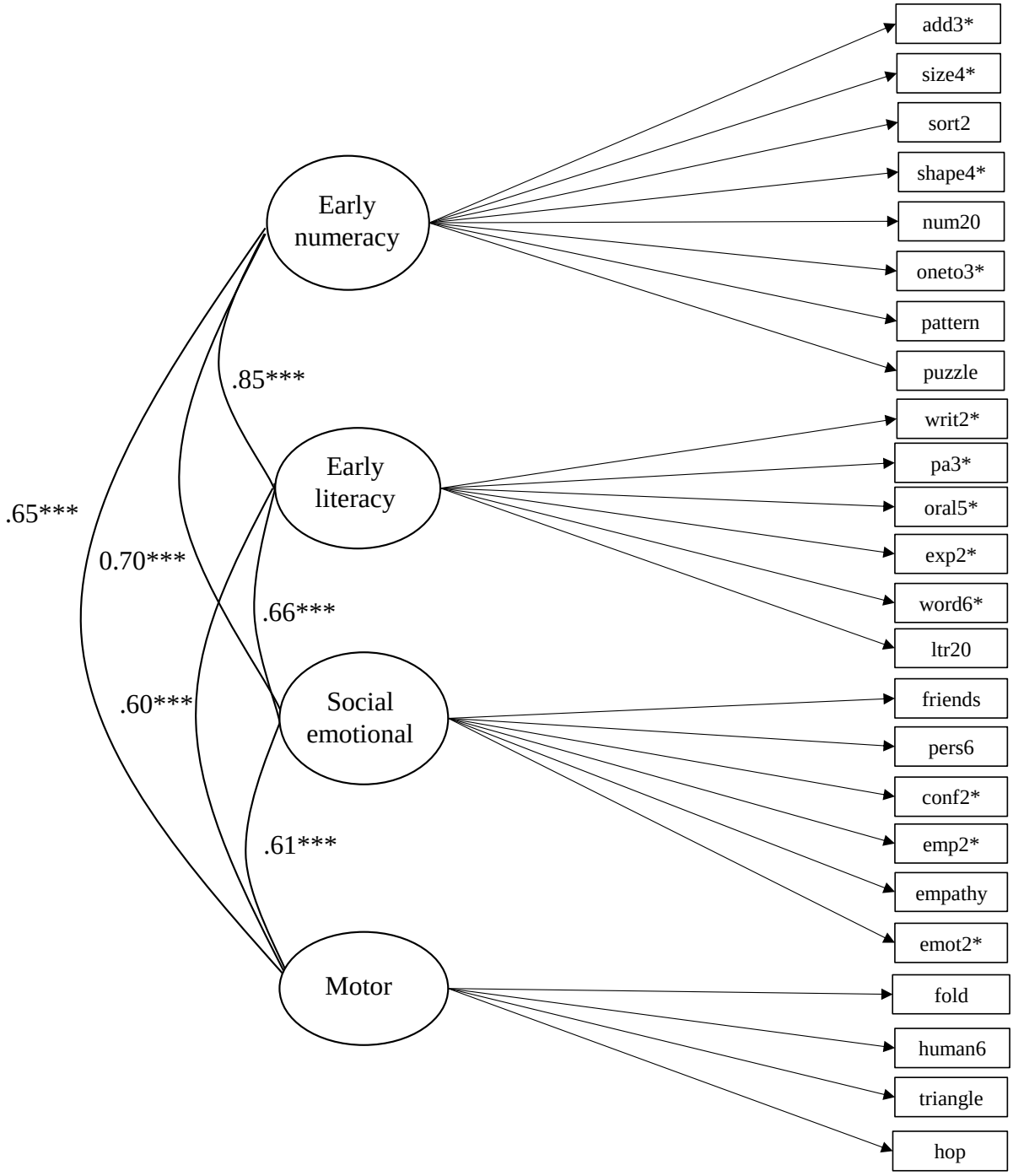
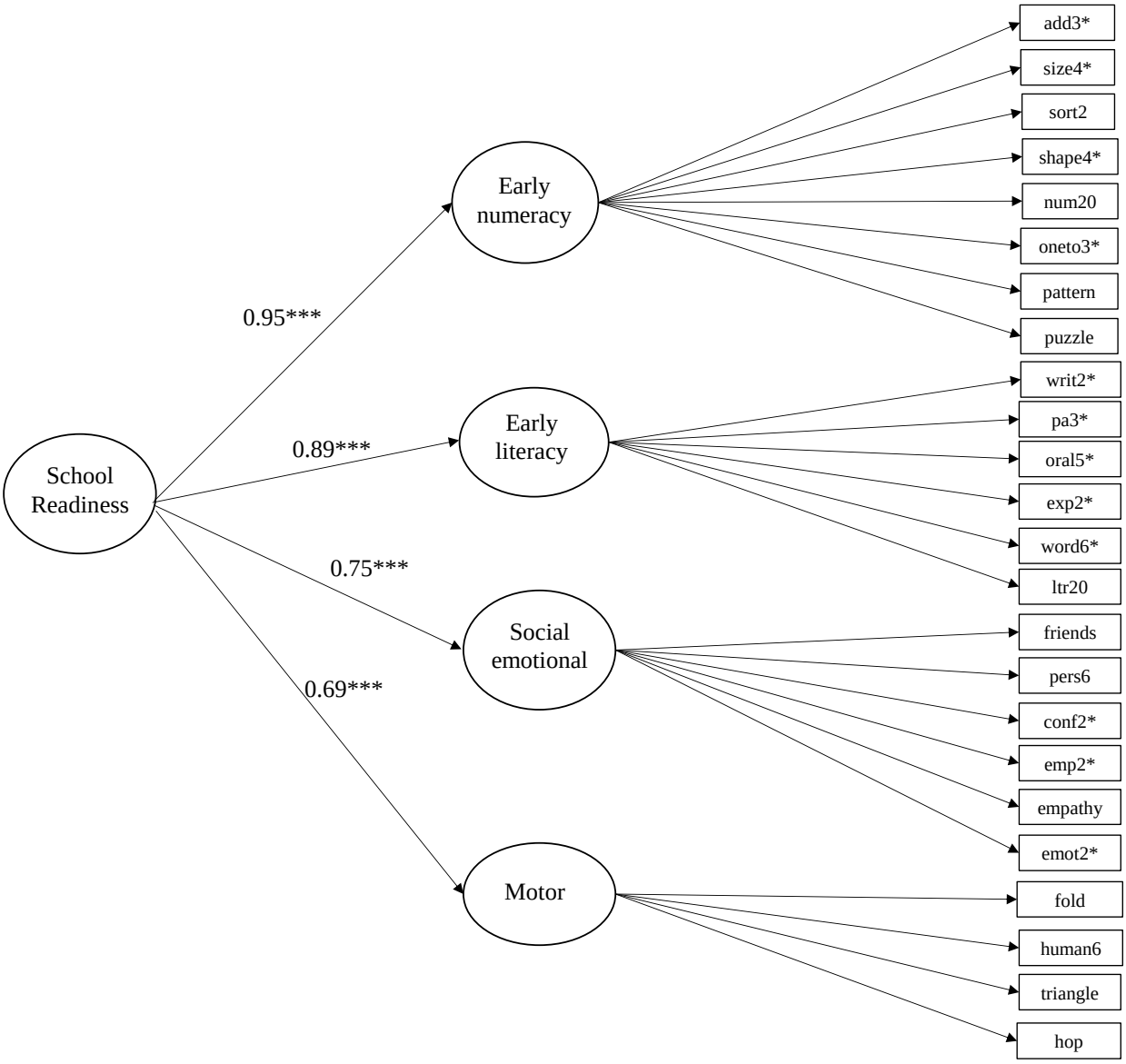


Figure 5. Diagram of the unconstrained factor analysis across four domains



Note. Diagram depicts subtasks only, though item level responses were included in analysis.
* indicates that a residual factor was fit for the items in the subtask.

Figure 6. Diagram of the higher order factor model, with all domain factors loading onto one general factor



Note. Diagram depicts subtasks only, though item level responses were included in analysis.
* indicates that a residual factor was fit for the items in the subtask.

Appendix Table 1. Percent of variance explained in by each item, by domain

		Percent of variance explained by factor			
		0-10%	**10-20%**	**20-30%**	**30% +**
Motor	COPYTRIA				0.672
	HUMAN2				0.879
	HUMAN3				0.444
	HUMAN4				0.784
	HUMAN5				0.562
	HUMAN6				0.952
	HUMAN7				0.901
	FOLD				0.347
Literacy	PA1		0.161		
	PA2	0.089			
	PA3		0.152		
	LTR1_V2				0.572
	LTR2_V2				0.790
	LTR3_V2				0.772
	LTR4_V2				0.763
	LTR5_V2				0.871
	LTR6_V2				0.824
	LTR7_V2				0.927
	LTR8_V2				0.962
	LTR9_V2				0.898
	LTR10_V2				0.803
	LTR11_V2				0.884
	LTR12_V2				0.857
	LTR13_V2				0.956
	LTR14_V2				0.904
	LTR15_V2				0.879
	LTR16_V2				0.921
	LTR17_V2				0.953
	LTR18_V2				0.916
	LTR19_V2				0.916
	LTR20_V2				0.917
	WORDPAIR				0.749
	WORDPAIR				0.776
	WORDPAIR				0.749
	WORDPAIR				0.828
	WORDPAIR				0.749
	WORDPAIR				0.686
	EXPRESSV			0.288	
	EXPRESSV				0.364
	ORALCOMP				0.707
	ORALCOMP				0.626
	ORALCOMP				0.824
	ORALCOMP				0.829
	ORALCOMP				0.887
	WRITELEV				0.343
	MATUREGR				0.411
Math	ONETOONE				0.701
	ONETOONE				0.449
	ONETOONE				0.354
	NUM1				0.859

	NUM2				0.701
	NUM3				0.862
	NUM4				0.906
	NUM5				0.892
	NUM6				0.916
	NUM7				0.876
	NUM8				0.921
	NUM9				0.938
	NUM10				0.868
	NUM11				0.877
	NUM12				0.932
	NUM13				0.950
	NUM14				0.941
	NUM15				0.981
	NUM16				0.904
	NUM17				0.950
	NUM18				0.914
	NUM19				0.913
	NUM20				0.937
	SHAPEID1			0.278	
	SHAPEID2				0.307
	SHAPEID3				0.355
	SHAPEID4	0.097			
	SHAPEID5				0.445
	SORT1				0.331
	SORT2				0.311
	SIZE1				0.879
	SIZE2				0.771
	SIZE3				0.951
	SIZE4				0.941
	ADD1				0.387
	ADD2				0.334
	SUB1				0.346
	PATTERN				0.300
	PUZZLE		0.178		
Social-Emotional	EMOTION4				0.506
	EMOTION1				0.505
	FRIENDS			0.215	
	EMPATHY1				0.591
	EMPATHY2				0.557
	EMPATHY3				0.631
	CONFLICT				0.466
	CONFLICT				0.345
	PERSONAL1				0.320
	PERSONAL2				0.443
	PERSONAL3				0.302
	PERSONAL4				0.671
	PERSONAL5				0.418
	PERSONAL6		0.190		