APA NLM

tapraid5/z2p-devpsy/z2p-devpsy/z2p99918/z2p4178d18z | xppws | S=1 | 10/9/18 | 3:37 | Art: 2018-1591

American Psychological Association

Developmental Psychology

0012-1649/18/$12.00

http://dx.doi.org/10.1037/dev0000626

Measuring Early Learning and Development Across Cultures: Invariance of the IDELA Across Five Countries

AQ: au
AQ: 1

Peter F. Halpin
University of North Carolina at Chapel Hill

Sharon Wolf
University of Pennsylvania

AQ: 2

Hirokazu Yoshikawa, Natalia Rojas, and Sarah Kabay
New York University

Lauren Pisani and Amy Jo Dowd
Save the Children, Washington, DC

Relatively little research has addressed whether conceptual frameworks of early learning generalize across different national contexts. This article reports on a cross-country measurement invariance analysis of the International Development and Early Learning Assessment (IDELA). The IDELA is a direct assessment tool for 3- to 6-year-old children, intended to measure Early Literacy, Early Numeracy, Motor, and Social-Emotional development. Its generalizability is evaluated using samples from 5 countries: Afghanistan (N = 2,629); Bolivia (N = 480); Ethiopia (N = 682); Uganda (N = 504); and Vietnam (N = 675). The 4-domain model of the IDELA was supported in each country, although the domains were highly correlated. Measurement invariance analysis revealed that most IDELA items do not provide a basis for comparing children's development over the 5 countries. This research supports the use of the IDELA for program evaluation and within-country monitoring purposes, but cautions against its use for international comparisons.

Keywords: early learning, international development, cross cultural, measurement invariance

AQ: 3 *Supplemental materials:* http://dx.doi.org/10.1037/dev0000626.supp

During the early childhood years, foundational cognitive, socioemotional, motor, and language skills develop (Black et al., 2017; Mwaura, Sylva, & Malmberg, 2008; Rao et al., 2014; Snow, Burns, & Griffin, 1998). Increasing recognition of the importance of supporting early learning and development (ELD) has led to calls for valid, reliable, and feasible assessments at the local, national, and global levels (e.g., World Health Organization, UNICEF, & World Bank Group, 2016; United Nations, 2017). Of particular relevance are skills and behaviors that enable children to learn in school, which have been referred to in terms of school readiness (Snow & Van Hemel, 2008; UNESCO, 2013).

From a policy perspective, at least three rationales have been put forward for assessments that measure normative development, rather than screening or diagnostic measures designed to identify delays or disabilities. First, evaluators wish to use such assessments to estimate the impact of early childhood programs and policies (e.g., Raikes, Yoshikawa, Britto, & Iruka, 2017). Second, an increasing number of governments and NGOs are interested in national or local monitoring of young children's development beyond survival and basic health and disease. Third, recent global initiatives have brought to the forefront the development of metrics that can be used to monitor and compare ELD across countries. For example, children's readiness for primary school is identified as a core component of the United Nations' Sustainable Development Goals (SDGs; United Nations, 2017). The lifelong learning goals proposed by the Learning Metrics Taskforce (UNESCO, 2013) and UNICEF's Early Learning and Development Standards (Kagan & Britto, 2005; UNICEF, 2017) also emphasize the importance of measuring ELD in international contexts.

Relatively few studies have addressed the psychometric properties of ELD assessments that are intended for use across national settings. In particular, research supporting distinct subdomains of ELD (e.g., cognitive, language, and social development) has been largely conducted in affluent, Western nations (e.g., Duncan et al., 2007; NICHD Early Child Care Research Network, 2005; Purpura, Hume, Sims, & Lonigan, 2011; Snow et al., 1998). While there is growing empirical support for the hypothesis of multiple subdomains of ELD in other regions of the world (Gladstone et al., 2010; Janus & Offord, 2007; Rao et al., 2014; Verdisco, Cueto, & Thompson, 2016; Wolf et al., 2017), it is not yet apparent the extent to which these multidomain conceptual models of ELD are

AQ: 20 Peter F. Halpin, School of Education, University of North Carolina at Chapel Hill; Sharon Wolf, Graduate School of Education, University of Pennsylvania; Hirokazu Yoshikawa, Natalia Rojas, and Sarah Kabay, Department of Applied Psychology, New York University; Lauren Pisani and Amy Jo Dowd, Save the Children, Washington, DC.

Correspondence concerning this article should be addressed to Peter F. Halpin, Halpin, School of Education, University of North Carolina at Chapel Hill, Campus Box 3500, Peabody Hall Office 111, Chapel Hill, NC 27599-3500. E-mail: peter.halpin@unc.edu

generalizable over countries, or whether the scores obtained from current ELD assessments provide unbiased comparisons over countries.

In this study, we address these issues in the context of one multidomain measure of ELD: The International Development and Early Learning Assessment (IDELA). The IDELA is a direct child assessment developed by Save the Children to measure ELD for children aged 3 to 6 in low- and middle-income countries (Pisani, Borisova, & Dowd, 2015). The assessment has been used in over 40 countries and has drawn growing attention from the international research and donor communities. The present study reports a secondary data analysis of samples drawn from five countries in AQ: 4 which Save the Children (SC) has conducted program evaluation research using the IDELA: Afghanistan, Bolivia, Ethiopia, Uganda, and Vietnam. These five samples were chosen because they each contained a sufficient number of observations for psychometric analysis and represented multiple regions of the developing world. T1 Table 1 provides a brief overview of the five countries in terms of several indicators, and in the Method section we describe the samples in more detail.

The International Development and Early Learning Assessment (IDELA)

In 2011, Save the Children completed a comprehensive review of existing child development assessments. Many of the instruments available at that time were limited in their approach, either targeting only one skill area or a specific age group, and many relied on parent or teacher report rather than directly assessing children's skills. Additionally, many instruments required special permissions and purchase. Most importantly, the majority of existing tools had been used primarily in high-income countries, such as the United States, the United Kingdom, and Australia, and they were not designed for use across countries with diverse populations and resource-poor settings.

The development of the IDELA was informed by multiple sources and existing tools, including the Early Development Instrument (Janus & Offord, 2007), the Ages and Stages Questionnaire (Squires & Bricker, 2009), the Malawi Developmental Assessment Tool (Gladstone et al., 2010), and the East Asia-Pacific Early Child Development Scales (Rao et al., 2014). Items were designed and then adapted to offer a balance between (a) international applicability, especially within low- and middle-income country contexts; (b) feasibility and ease of administration and adaptation; and (c) psychometric rigor. Initial versions of the IDELA were tested and modified over 3 years in multiple sites across 12 different low- and lower-middle-income countries. Pisani et al. (2015) provide details of the development of the individual IDELA items.

The conceptual model of the IDELA is summarized in Figure 1. The IDELA aims to measure four distinct domains of child F1 development: Emergent Literacy, Emergent Numeracy, Social-Emotional Skills, and Gross and Fine Motor Skills. Each domain is conceptualized in terms of a number of component skills. For example, Shape Identification is a component skill of the Emergent Numeracy domain. The component skills are in turn operationalized in terms of one or more specific behavioral responses to be provided by the child (e.g., "Can you point to the circle?"). We will refer to the skills within domains as *subtasks*, and the specific behavioral responses that make up each subtask as *items*. The version of the IDELA used in the present study has 21 subtasks and 76 items, which are described in more detail in the Measures section. The current version of the tool is available at https://idela-network.org.

Two recent studies have provided empirical evidence about the proposed conceptual model of the IDELA in Ethiopia (Wolf et al., 2017) and Bhutan (Wuermli, Helm, Hastings, Yoshikawa, & Dowd, 2017). Both of these studies concluded that a bifactor model (see, e.g., Rijmen, 2010; also, the Method section) with four general factors provided acceptable fit to the item-level data of the IDELA. The bifactor model was used to allow for dependence among items on the same subtask. The four general factors corresponded to the posited four-domain structure of the IDELA and were found to have plausible relations with external variables. Moreover, both studies evaluated whether the internal structure of the IDELA was invariant over key subgroups within countries, including (a) gender, (b) urbanicity, and (c) child's enrollment status in early child care and development (ECCD) programs. Wolf et al. (2017) additionally examined measurement invariance over experimentally induced subgroups used for program evaluation. The IDELA domains were found to be invariant in the majority of these comparisons, with a main exception being the Motor domain over urbanicity in Bhutan.

The Current Study

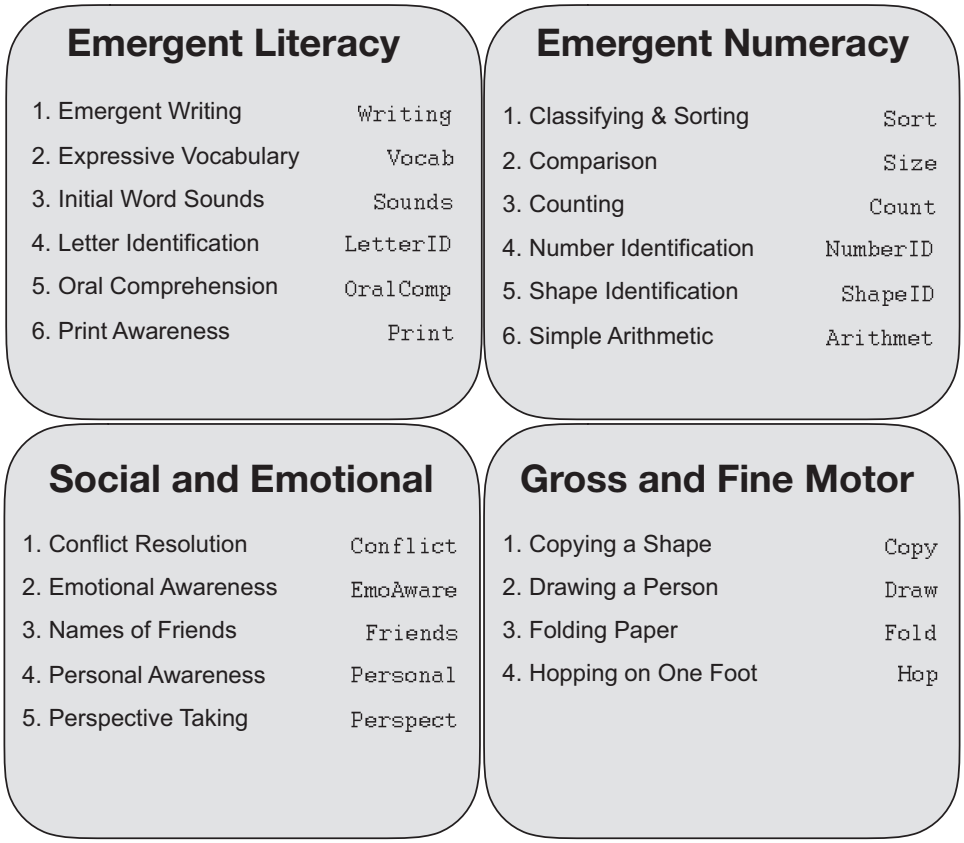
This study builds on previous research in two ways. First, we evaluate the fit of the bifactor model in four new samples of

Table 1
Country Information

Country	Country income level	HDI ranking/grouping	Poverty rate	GER primary school	GER preprimary school
Afghanistan	Low	169/Low	35.8%	111.9%	.8%
Bolivia	Lower-middle	118/Medium	38.6%	97.1%	70.7%
Ethiopia	Low	174/Low	29.6%	102.1%	30.4%
Uganda	Low	163/Low	19.5%	109.9%	11.6%
Vietnam	Lower-middle	115/Medium	13.6%	108.9%	83.1%

Note. HDI denotes Human Development Index (source: hdr.undp.org/en/countries). It is a composite of life expectancy, education, and income per capita indicators used to rank 188 participating countries by overall human development. All other statistics were originally reported by the World Bank (source: data.worldbank.org/country). GER denotes Gross Enrollment Ratio. It is the ratio of the number of students who live in a country to those who qualify for a particular grade level. It indicates the proportion of the population that a country is able to accommodate, not actual students enrolled. The GER can be over 100% as it includes students who may be older or younger than the official age group.

tapraid5/z2p-devpsy/z2p-devpsy/z2p99918/z2p4178d18z | xppws | S=1 | 10/9/18 | 3:37 | Art: 2018-1591



AQ: 21

Figure 1. The IDELA conceptual model: Core domains and skills. Variables names used for reporting statistical analyses are also indicated.

children from Afghanistan, Bolivia, Uganda, and Vietnam, building on results from a fifth country (Ethiopia) previously reported (Wolf et al., 2017). We thus conduct four replication studies addressing the generalizability of the conceptual model of the IDELA. Second, we address whether the scores obtained from the IDELA provide unbiased comparisons over countries. In general, scores from assessments that are translated or otherwise adapted for use in new populations cannot be assumed to be comparable to the original assessment or to one another (Joint Committee on Educational and Psychological Testing, 2014). In particular, cross-country comparisons using the IDELA will be biased unless its items, or a subset thereof, have equivalent psychometric properties over countries.

In the psychometric literature, procedures for comparing an assessment over groups have been discussed in terms of measurement invariance (e.g., Millsap, 2011) and differential item functioning (DIF; e.g., Holland & Wainer, 1993). Bauer (2017) provided a synthesis of these research areas, and Zwitser, Glaser, and Maris (2017) addressed their importance in international contexts. We briefly review this literature to facilitate interpretation of our research questions.

Measurement invariance means that an assessment can be used to make unbiased comparisons over populations (American Psychological Association, 2014, p. 211). There are three commonly assessed types of measurement invariance, which can be interpreted in terms of increasingly strict levels of invariance. The first level, *configural* invariance, assesses whether the number of factors and the general pattern of factor loadings are the same across groups. Our four replication studies address configural invariance of the IDELA. The second level, *metric* invariance, assesses whether the factor loadings are numerically equivalent across groups. Lack of metric invariance is also referred to as *nonuniform* DIF. This level of invariance allows for comparisons of the (co-) variances of the IDELA domains, but not their means, across countries. The third level, *scalar* invariance, assesses whether the thresholds of the items are equivalent across groups. Thresholds describe the difficulty of an item (e.g., the proportion of correct responses), and lack of scalar invariance is also referred to as *uniform* DIF. In the present study we focused on scalar invariance because it is required for unbiased estimation of mean differences on the IDELA domains.

Additional types of measurement invariance have been proposed for evaluation of the bifactor model, depending on how the residual factors are treated (Jeon, Rijmen, & Rabe-Hesketh, 2013). In this study, we require only configural invariance on the residual factors, as we do not seek to make subtask-level comparisons over countries. We also note that scalar invariance of the full set of IDELA items is a relatively strict requirement. As an alternative to scalar invariance, it is also possible to use a small "anchor set" of items to statistically equate other items that are not directly com-

tapraid5/z2p-devpsy/z2p-devpsy/z2p99918/z2p4178d18z | xppws | S=1 | 10/9/18 | 3:37 | Art: 2018-1591

parable across groups. This is referred to as *partial* measurement invariance, and is also a key principle behind test equating and linking (e.g., Kolen & Brennan, 2014). In the present study we consider partial scalar invariance as a plausible alternative to full scalar invariance. It is important to note that, even with full or partial scalar invariance, the bifactor model requires latent-variable methods for scoring; scoring methods based on observed scores (e.g., raw totals) are generally not compatible with the model.

The proposed analyses are summarized in terms of the following five research questions (RQs).

Within-Country RQs (Configural Invariance)

RQ1 (*Dimensionality of domains*). Do the item response data from each IDELA domain support the hypothesis of a single, general construct?

RQ2 (*Item specificity*). Do any items on the IDELA measure a domain other than the one intended?

RQ3 (*Relations among domains*). How are the IDELA domains related to one another? In particular, are the four domains statistically distinct from one another, or are they better represented as an undifferentiated general factor?

Across-Country RQs

RQ4 (*Scalar invariance*). Are the IDELA domains, each taken as a whole, invariant at the scalar level over the five countries?

RQ5 (*Partial scalar invariance*). If the IDELA domains, taken as a whole, are not invariant over countries, (a) to what extent does the noninvariance lead to biased comparisons (i.e., mean differences) among the countries, and (b) are there subsets of items that may be invariant?

Method

Participants

Secondary analysis of the data sets included in this report was approved by the internal review board of Save the Children (project title: "IDELA measurement invariance study"; protocol number FWA00022738). Table 2 summarizes the data sets used for each country. In all countries except Afghanistan, data were collected during the baseline phase of quasi-experimental studies to assess the efficacy of ECCD programs implemented by Save the Children. The children recruited into these studies were a convenience sample based on geographic proximity to locations in which Save the Children had established field offices. The data reported here were collected prior to service delivery from Save the Children. In Afghanistan, data were collected as part of a large-scale assessment of young children's development, covering sites in addition to those in which Save the Children conducted regular operations. In all cases, the samples should not be considered nationally representative. In Afghanistan, Ethiopia, and Uganda, samples included children enrolled in ECCD programs as well those who were not. Past research has found that the IDELA is invariant over populations of children enrolled and not enrolled in ECCD programs (Wolf et al., 2017), although children enrolled in ECCD programs tended to perform better than those who were not. In our secondary analyses we restrict comparisons over countries to only those children who were enrolled in ECCD programs.

Afghanistan. Data were collected in September 2015 in order to provide a cross-sectional summary of young children's skills and development as part of a monitoring/advocacy effort, the first large-scale assessment of young children's development in the country. Across the four provinces, 2,629 children between 3 and 6 years old and their families were included in this study, half of whom were enrolled in various SC-supported ECCD programs (with instruction predominantly in Uzbek and Dari), and half of whom were living in the same villages but not enrolled in an ECCD program. For the ECCD group, 106 ECCD centers supported by Save the Children were randomly sampled. Within each of those centers, an average of 25 children were randomly sampled and included in the study if they were reported to be between 3.5 and 6.5 years old. The comparison group comprised children who were living in villages with ECCD centers, but who were not attending the centers.

Bolivia. The objective of the study was to explore children's learning over one school year and identify equity gaps. The sample included 8 *centros infantiles* (child development centers) and 20 *niveles iniciales* (the equivalent to a pre-K/Kindergarten class located within an elementary school), all in a peri-urban area of Cochabamba, the fourth largest city in Bolivia. All of the centers were publicly run programs receiving support from Save the Children in the development of learning materials, teacher training, and community outreach events for parents. The data were collected by a team of five enumerators trained at the Cochabamba office as part of a baseline assessment.

Table 2
Sample Information

Country	*N*	Region	Assessment language	Mean age (range)	% female	% ECCD	Urbanicity	Mother/Father literate (%)
Afghanistan	2,629	Yen Bai and Quang Nam provinces	Vietnamese	5.4 (3–8)	57%	44.60%	Urban + rural	17/38
Bolivia	480	Wakiso district	Luganda	4.7 (3–6)	49%	100.00%	Peri-urban	96/98
Ethiopia	682	West Shewa region	Oromiffa	5.9 (4–7)	52%	76.10%	Rural only	26/60
Uganda	504	Cochabamba	Spanish	4.6 (4–6)	48%	48.60%	Rural only	Not available
Vietnam	675	Faryab, Saripol, Kandahar and Kabul provinces	Uzbek and Dari	4.3 (3–5)	50%	100.00%	Rural only	82/90

Note. *N* denotes sample size and ECCD denotes enrollment in an early childcare and development program.

tapraid5/z2p-devpsy/z2p-devpsy/z2p99918/z2p4178d18z | xppws | S=1 | 10/9/18 | 3:37 | Art: 2018-1591

Ethiopia. These data came from an IDELA baseline assessment in West Shewa, Ethiopia, from November 3–24, 2014. The goal of the assessment was to understand the effectiveness of a particular Save the Children preprimary program (Emergent Literacy and Math) in improving children's learning and development outcomes. Data were collected from nine villages within four rural Woredas (districts), for a total of 36 villages. In half of the Woredas, children had access to ECCD centers, while in the other half they did not. Assessors randomly sampled an average of 20 children aged 5–6 years from within each village, as well as one caregiver per child. Data collection occurred in November 2014 and ran for 3 weeks. Eighteen data collectors were trained to collect data using Android tablets with Tangerine software.

Uganda. Data was collected in June 2016 as part of a baseline assessment within a Save the Children Sponsorship area in the Wakiso district of Uganda. Twenty children aged 4, 5, and 6 years were randomly sampled from 30 ECCD centers across four rural subdistricts. Instruction at all centers was in Luganda. Data collectors were trained to collect data using Android tablets with Tangerine software.

Vietnam. These data were collected as part of a baseline assessment from Save the Children-supported areas in the Yen Bai and Quang Nam provinces. The study targeted 30 ECCD centers and 15 children within each center. Traditionally, instruction in ECCD centers in these communities was primarily in Vietnamese, but facilitators were being trained in how to incorporate mother-tongue instruction. The selection of the sample was based on the proximity of the school to a paved road, the economic situation of the area, and the ethnic minority population.

Measures

All children were assessed using the IDELA. The assessment was administered by a trained enumerator, usually a field officer recruited from the local population, and required about 30 min per child. The instructions for children were translated into the local language(s) and adapted using a process of review and field-testing. All text-based items used the country site's alphabetic script and the Arabic number system. In three of the five samples, communities were monolingual, but administration in Afghanistan and Vietnam included multiple languages. In Afghanistan, the languages used varied depending on the area of the country, and in Vietnam, children in the sampled communities spoke a mixture of Vietnamese and a local language. In cases where children's mother tongue differed from the national language, assessors were asked to deliver instructions in children's mother tongue. We make note of subtasks that require significant adaptions beyond translation.

The number of items per subtask has been modified throughout IDELA's revision process, and three different versions were used in the five samples reported on in this study. In all analyses, we used only the items that were common to all three versions of the IDELA. Most items are scored as "correct/incorrect," but a few are scored as ordered categorical, with higher numbers denoting better performance on the item. Below we summarize the content of each domain and subtask, indicate which items are ordered categorical, and describe a number of example items. For a full description of the current version of the IDELA and the revision process, see Pisani et al. (2015).

Emergent literacy. This domain consisted of 24 items grouped into six subtasks: Emergent Writing (1 item with 4 categories), Expressive Vocabulary (2 items each with 11 categories each), Initial Word Sounds (3 items), Letter Identification (10 items), Listening Comprehension (5 items), and Print Awareness (3 items).

For the Letter Identification subtasks, the enumerator points to a sequence of letters arranged in a grid and the student is asked to name each letter. The letters are not presented in alphabetical order. Rather, high-frequency letters were determined by local staff, and the letters are randomly ordered on the grid. This means that the 10 items corresponded to 10 different letters, in each of the different countries. To address this situation, we reordered the items by the proportion of correct responses within each country. Consequently, the first item on the Letter Identification subtask was always the easiest in that country, and the last item was always the hardest.

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 the 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.

Emergent numeracy. This domain consisted of 27 items grouped into six subtasks: Classifying and Sorting (2 items), Comparison (4 items), Counting (3 items), Number Identification (10 items), Shape Identification (5 items), and Simple Arithmetic (3 items).

The Number Identification subtask used the numbers 1–10 in a manner analogous to the Letter Identification subtask. The numbers are presented in a random order, but the same order was used across countries. The main adaptations over countries appear on the Comparison and Counting subtasks. On the Counting subtask, a number of small, familiar objects (e.g., beads, pebbles) is placed in front of the child, with the choice of objects depending on what is available at the field site. The enumerator asks the child to give him/her a certain number of the objects, and if the child provides the requested number, the response is recorded as correct. The same small objects are used in part of the Simple Arithmetic subtask. Picture cards, which can be locally adapted, are used for both the Simple Arithmetic subtask and the Comparison subtask.

Motor development. The Motor domain consisted of 12 items, grouped into four subtasks: Copying a Shape (1 item), Drawing a Person (7 items), Folding a Piece of Paper (1 item; 4 categories), and Hopping on One Foot (1 item; 11 categories). These subtasks 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 fold at each step (within 1 cm).

Social-emotional development. The Social-Emotional domain consisted of 13 items grouped into five subtasks: Conflict Resolution (2 items), Emotional Awareness (2 items), Names of Friends (1 item; 11 categories), Personal Awareness (6 items), and Perspective Taking (3 items). The Conflict Resolution items ask

tapraid5/z2p-devpsy/z2p-devpsy/z2p99918/z2p4178d18z | xppws | S=1 | 10/9/18 | 3:37 | Art: 2018-1591

the child to decide what to do if he or she were playing with a toy and another child wanted to play with the same toy. "Correct" answers, as agreed upon with local educational staff, typically included talking to the child, taking turns, sharing, or getting another toy. Similar modifications are made for items on the Emotional Awareness subtask, which asks children to describe situations in which they have been happy or sad.

Analytic Approach

As mentioned, the IDELA is structured such that items are nested within subtasks, which are nested within domains. Items on the same subtask use the same stimulus materials and require similar responses (e.g., pointing, verbalization). Consequently, it is important to control for similarities among items on the same subtask when making inferences about the domain-level constructs. The bifactor model is one way of dealing with this situation, and is a more general version of both the hierarchical factor (second-order) model and the testlet model (see, e.g., Rijmen, 2010). Using the bifactor model, each item is modeled via two factors, a general factor corresponding to the domain, and a specific factor corresponding to its subtask. The subtask factors are typically interpreted as representing residual variation, because they are assumed to be uncorrelated with the general factor and with each other (although more general models can be identified).

The bifactor model has been criticized for its tendency to overfit data (e.g., Bonifay, Lane, & Reise, 2017; Reise, Kim, Mansolf, & Widaman, 2016). Exploratory research with the IDELA has found that the residual factors map directly to the subtask structure of the instrument (Wolf et al., 2017). In particular, using an exploratory bifactor rotation (Jennrich & Bentler, 2011), items on the same subtask were consistently found to load on the same residual factor, rather than on different residual factors. Additionally, fitting a single-factor model to each of the IDELA subtasks, modification indices (Sorbom, 1989) clearly indicated that fit could be improved by adding residual correlations among items on the same subtask, but not items on different subtasks. Although not reported here, we replicated these exploratory analyses with random subsets of the data from each of the five countries reported on in this study. While the degree of residual association among items on the same subtask differed over countries, there was no case in which a different residual structure was apparent in the data. Based on prior research and exploratory work with the present sample, we concluded that a bifactor model was appropriate for application with the IDELA.

A bifactor model for the IDELA can be described as follows. Let X_{ijkm} be a binary variable denoting the observed response of child $i = 1, \ldots, N$ to item $j = 1, \ldots, J_k$ on subtask $k = 1, \ldots, K_m$ of domain $m = 1, \ldots, M$. Let η_{ijkm} denote the probit of X_{ijkm}. Then a bifactor model for the binary IDELA items can be written

$$\eta_{ijkm} = \alpha_j + \beta_{jm} f_{im} + \gamma_{jk} g_{ik} \quad (1)$$

where:

α_j is the threshold for the item.
β_{jm} is the factor loading of the item on the domain factor.
f_{im} is the child's (latent) level on the domain factor.
γ_{jk} is the factor loading of the item on the subtask factor.
g_{ik} is the child's (latent) level on the subtask factor.

It is assumed that the random vector $f_i = (f_{i1}, \ldots, f_{iM})$ is multivariate normal with mean vector μ and (unstructured) covariance matrix Ψ, that the random vector $g_i = (g_{i1}, \ldots, g_{iK})$ is multivariate normal with mean zero and diagonal covariance matrix Φ, and that $\mathrm{COV}(f_i, g_i) = \mathbf{0}$. For the IDELA items that are ordered categorical rather than binary (see Measures section), an ordered-probit model is used in place of the probit model. This implies that the threshold parameter of the item is replaced by a vector of thresholds, but the model specification is otherwise the same.

Interpretation of the model parameters. In the present application, the latent factor f_m describes children's competency on IDELA domain m. The parameters μ and Ψ describe the population of children within a country with respect to their performance on the IDELA domains. The latent factors g_k describe residual variation in children's competency that is associated with a specific subtask k. In the present study, these residual factors are not of substantive interest.

The model parameters (i.e., α_j, β_{jm}, and γ_{jk}) describe the IDELA. To distinguish them from model parameters that describe the population of children, they are collectively referred to as the *measurement parameters*. The threshold parameters, α_j, describe the proportion of students who respond to each category of the item j. For binary items, this corresponds to the difficulty of the item (i.e., the proportion of correct responses). The loadings β_{jm} describe how strongly an item is associated with its focal domain f_m, and similarly for γ_{jk}. In general, it is important that performance on an item is more strongly related to the domain it is intended to measure than its specific subtasks (i.e., $|\beta_{jm}| \geq |\gamma_{jk}|$). This is important because, if an item loads more strongly onto its subtask factor than the general factor, this suggests that the subtask factor is the principal source of variation and should not be interpreted as merely a residual.

Path diagrams representing the conceptual model of each domain of the IDELA are presented in Figure 2, with the variable codes used for each subtask listed in Figure 1. Note that most but not all subtasks contain multiple items. For subtasks comprised by a single item, the residual factor is omitted. For subtasks with only two items, both items' factor loadings on the residual factor are fixed to one to identify the variance of the residual factor.

Comments on estimation. Using the probit link function, the model parameters can be consistently estimated using a weighted least squares (WLS) loss function with the polychoric correlation matrix of the observed data (see, e.g., Muthén & Satorra, 1995a). This approach can also accommodate robust standard errors and goodness-of-fit statistics under the usual framework for model misspecification and complex sampling designs (Muthén & Satorra, 1995b), with corresponding adjustments to tests of model fit for nested models (Muthén, Du Toit, & Spisic, 1997; Satorra & Bentler, 2010). A logit link function is more commonly used when estimation is based on the likelihood of the full contingency table of the items, although estimation of high dimensional models using maximum likelihood has only recently become tractable (see Cai, 2010; Rijmen, 2010). The identification of the model is similar using either estimation method. In the present study, we scaled latent response variables by setting their total variance to one, and scaled the factors (domain and subtask) by setting their total variance to one. All analyses were conducted using Mplus 8,

tapraid5/z2p-devpsy/z2p-devpsy/z2p99918/z2p4178d18z | xppws | S=1 | 10/9/18 | 3:37 | Art: 2018-1591

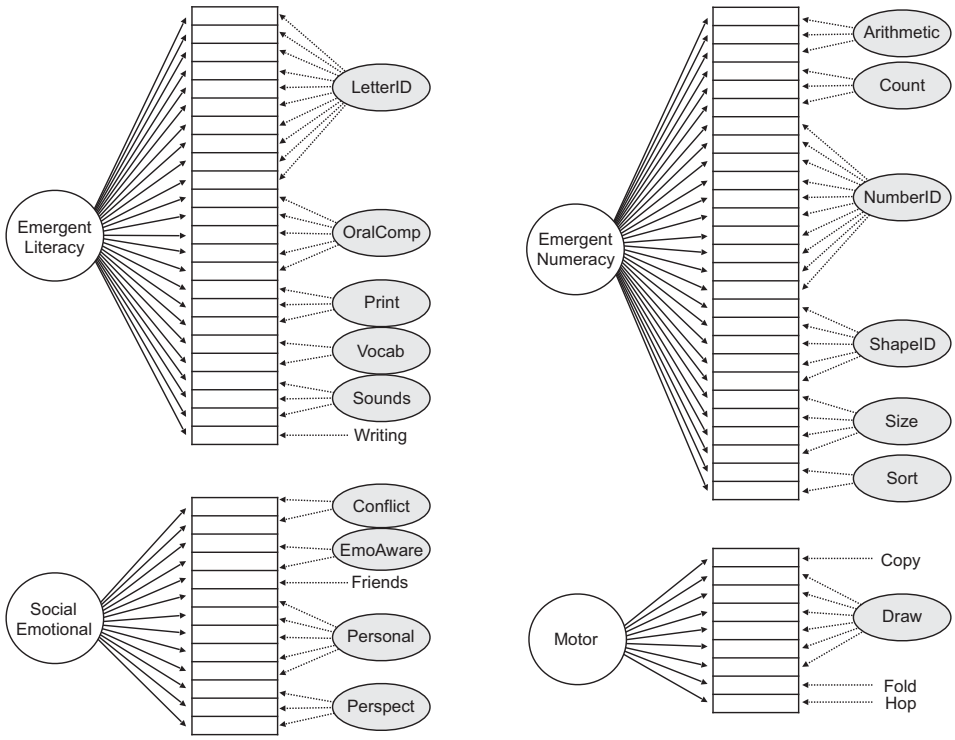


Figure 2. Bifactor models for the IDELA domain. Subtask factors are assumed to be uncorrelated with the domain factor and with one another. In the Unconstrained model, the four domain factors are assumed to be correlated. In the Unidimensional model, the four domain factors are replaced with a single, overarching factor. Note that subtasks comprised by a single item do not have a residual factor; this is indicated by omitting the latent variable (oval) for those subtasks. Variable codes are listed in Figure 1.

with example code included in the online supplemental materials.

Focal Inferences

Within-country RQs. When assessing RQ1, the main concern is whether the model in Equation (1) fits the data, for a given domain, in a given country. This leads to a total of four (domains) × five (countries) analyses. Using the probit approach, the null hypothesis for each analysis may be formally stated in terms of polychoric correlation matrices:

$$H_{01}\colon \Sigma_m = \Sigma_m(\theta_m),$$

where Σ_m denotes the population covariance matrix of the items on domain m, θ_m is a vector containing the model parameters for domain m, and $\Sigma_m(\theta_m)$ is the model-implied covariance matrix for domain m. If we reject the null hypothesis in a given country, we conclude that the bifactor model with one general factor does not generalize to that country.

As noted, we are additionally concerned with the interpretation of the model, and in particular, whether the items load at least as strongly on the domain factor as they do on their subtask factor. Because significance tests associated with factor loadings are relatively uninformative in large sample settings, and since there are a large number of factor loadings to consider (5 countries × 2 factors × 76 items = 760 loadings), we informally evaluate the relative magnitude of domain and subtask loadings using heat maps of the point estimates.

For RQ2, we are again concerned with whether the model fits the data, but this time for all domains simultaneously, in a given country. The focal inference can be stated as

$$H_{02}\colon \Sigma = \Sigma(\theta),$$

where the notation and conclusions parallel those for H_{01} in an obvious way. We are additionally concerned with whether any misfit of the four-domain model can be attributed to lack of item specificity (i.e., nonzero coefficients $\beta_{jm'}$ where m' denotes a domain other than the one the item j is intended to measure). We address this question using modification indices. Again, the number of parameters to be evaluated is very large (5 countries × 3 cross-domains × 76 items = 1,140 cross-loadings), so we make these considerations informally.

For RQ3, we are interested in various structural hypotheses about the correlations among the factors, again within a single country. In particular, prior research has shown that theoretically defined domains of ELD may be highly correlated with one another (Janus & Offord, 2007; Rao et al., 2014; Wolf et al., 2017). Therefore, we focus on comparing the hypothesized four-domain model of the IDELA against one in which the full set of IDELA items is modeled with a bifactor model with a single general factor. We refer to these as the *Unconstrained* model and the *Unidimensional* model, respectively. The null hypothesis of interest is

tapraid5/z2p-devpsy/z2p-devpsy/z2p99918/z2p4178d18z | xppws | S=1 | 10/9/18 | 3:37 | Art: 2018-1591

$$H_{03}: \Sigma(\theta_1) = \Sigma(\theta_2),$$

where θ_1 and θ_2 denote the parameters of the two different models. Because the Unidimensional model is nested within the Unconstrained model, H_{03} can be tested using a chi-square test for nested models. If the null hypothesis is rejected, we conclude that the Unidimensional model fits the data worse than the Unconstrained model, and therefore reject the Unidimensional model as insufficient. If the null hypothesis is retained, we conclude that the Unidimensional fits the data just as well as the Unconstrained model, and therefore reject the Unconstrained model as unnecessary.

Across-country RQs. For RQ4 and RQ5, the overall question is whether the measurement parameters of the bifactor model are invariant over countries. We used the MIMIC model (see Bauer, 2017; Muthén, 1989) to address this question. The MIMIC model allows for group differences on the item thresholds (difficulty parameters) only, which is referred to as *uniform* DIF. To allow for separate conclusions about each domain, we conducted the analyses one domain at a time. Additionally, to facilitate comparison over countries, we balanced the size of each sample by selecting all children from the smallest sample ($N = 480$ in Bolivia), and selecting a random subsample of $N = 480$ from the remaining countries. We also balanced the subsamples in each country on gender. A number of sensitivity analyses were conducted to ensure that our conclusions did not depend on the model specification or the choice of samples.

AQ: 7

Using the MIMIC model requires respecification of Equation (1) as

$$\eta'_{ijkmc} = \eta_{ijkm} + \delta_{jc} \quad (2)$$

where δ_{jc} denotes a country-specific item "effect" for the $c = 1, \ldots,$ five countries. We use deviation (sum-to-zero) coding for the country effects, so that δ_{jc} may be interpreted as each country's deviation from the overall mean of the item thresholds (or vector of thresholds), α_j. We also respecify the population model so that the means on the general factor for each domain, μ_{mc}, may vary over countries. For the population means we again utilize deviation coding, so that the μ_{mc} can be interpreted as each country's deviation from the overall mean, μ_m, which is set to zero to identify the model.

We address RQ4 by comparing two specifications of Equation (2), for each domain. In the scalar invariance specification, we assume that all item thresholds are constant across all countries, and estimate country means on the domain factor (i.e., fix $\delta_{jc} = 0$ for all j and c, and estimate μ_{mc} freely). For configural invariance specification, we estimate the item thresholds freely in each country, but fix the country means on the domain factor to equal zero for the purposes of model identification (i.e., estimate δ_{jc} for all j and c, but fix $\mu_{mc} = 0$ for all c). Because the scalar model is nested within the configural model, the two models can be compared using a chi-square test for nested models. The null hypothesis of this test is

$$H_{04}: \Sigma_m(\theta_{m1}) = \Sigma_m(\theta_{m2}),$$

where θ_{m1} and θ_{m2} denote the model parameters for the two models, for each domain m. The interpretation is analogous to that of RQ3.

We conducted a number of sensitivity analyses to ensure the robustness of conclusions about H_{04}. To address the possibility of nonuniform DIF (i.e., group-specific differences among the loadings on the domain factors), we used a multigroup model for measurement invariance (see Bauer, 2017; Millsap, 2011). We also considered whether balancing the samples in each country on age and/or exposure to ECCD led to substantially different conclusions. When balancing on age, it was necessary to omit the Ethiopian sample, because it did not include sufficient numbers of younger children. The sensitivity analyses are summarized in the online supplemental materials. We note that the alignment method for assessing measurement invariance (Asparouhov & Muthén, 2014) is not currently implemented for multidimensional models (Muthén & Muthén, 2018); for this reason, we did not compare the MIMIC approach to the alignment method.

If the scalar model is rejected, a number of follow-up analyses become relevant. These analyses are summarized under RQ5, which we approach from an exploratory, post hoc perspective. There exist a large number of procedures that can be used to identify items that are biased over groups (Holland & Wainer, 1993). In the present study, we use an approach that is logically similar to forward selection in stepwise regression. We start with the scalar invariance model, and at each step remove the one constraint $\delta_{jc} = 0$ that produces the largest modification index. The procedure stops when the chi-square difference of the modified model against the configural model is nonsignificant at the 5% alpha level. The resulting partial invariance model provides estimates of the population means, which we denote as μ'_{mc} to distinguish them from the population means from the scalar model, denoted μ_{mc}. The practical importance of DIF on the IDELA can be judged by considering how our conclusions about differences among countries would change when using μ'_{mc} in place of μ_{mc}. For example, the rank orderings of the countries' means may change, or the difference among countries may become more or less pronounced. In order to consider whether any invariant items are meaningfully grouped with subtasks or domains, we also report the δ_{jc} in the partial invariance model. Here we reestimate the MIMIC model, treating the μ'_{mc} as fixed and report confidence intervals on all δ_{jc}.

The exploratory nature of RQ5 should be emphasized—these analyses are intended as post hoc procedures to follow up RQ4, in the case that scalar invariance over countries is rejected for one or more domains. These follow-up analyses can provide initial evidence about the pragmatic importance of modeling DIF when making comparisons over countries with ELD assessments, and can help to identify particular subsets of items on the IDELA that may be suitable for making comparisons over countries.

Results

Summary of Item Omissions and Model Modifications

Item omissions. Two items on the Draw a Person subtask of the Motor domain were collinear with the other items on that subtask, in all countries. 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. Another item asked the child to draw a second facial feature, which was collinear with drawing the first facial feature. These two items were omitted from all analyses.

tapraid5/z2p-devpsy/z2p-devpsy/z2p99918/z2p4178d18z | xppws | S=1 | 10/9/18 | 3:37 | Art: 2018-1591

One item on the Personal Awareness subtask of the Social-Emotional domain was answered incorrectly by nearly all children in Bolivia, Uganda, and Vietnam. This resulted in unstable estimates of the correlations between this item and many other items, and therefore the item was omitted from all analyses. The item asked the child to name their state/country.

We also mention here one item on the Motor domain (Hopping on One Foot) that was identified during our analysis of RQ2 and removed from all subsequent analyses (RQ3–RQ5) due to lack of specificity. Nearly all children were able to hop the full 10 times, but children who could not do so were also more likely to do poorly on many of the literacy and numeracy items.

Model modifications. A number of items were highly correlated with one or more items on the same subtask, but only in some countries. For subtasks with few items, this resulted in very large factor loadings on the residual factors (i.e., Heywood cases). When this happened, we constrained the factor loadings on the residual factors to be equal to one another, and used the constrained model for all analyses. This allowed us to retain the items while also ensuring that equivalent models were being used across countries. In each country, the resulting respecifications of the residual factor loadings resulted in negligible decrements to model fit. The modified subtasks were (a) Emergent Literacy: Initial Word Sounds, Print Awareness; (b) Emergent Numeracy: Comparison, Simple Arithmetic; and (c) Social Emotional: Perspective Taking.

Finally, the analyses for RQ2 revealed that the residual factors of the Number Identification subtask and the Letter Identification subtask were highly correlated in all countries. This was plausibly due to both subtasks using printed text arranged on a grid. We included the correlation between the residual factors in models reported for RQ2 and RQ3.

Within-Country Analyses (RQ1, RQ2, and RQ3)

Table 3 summarizes the model fit for each domain considered separately, in each country. In the factor analysis literature, adequate model fit is typically considered to be indicated by a root mean square error of approximation (RMSEA) of $<.05$, and a Tucker-Lewis index (TLI) of $>.95$ (e.g., Hu & Bentler, 1999). Based on these criteria, we conclude that the bifactor model with a single general factor had acceptable fit to each domain of IDELA, in each of the five samples. The estimated factor loadings on the domain and subtasks factors are summarized in Figure 3. For most items and most countries, the loadings on the domain factors were at least as large as the loadings on the subtask factors. There were, however, some marked exceptions, including: (a) the Number Identification and Letter Identification subtasks in Vietnam; (b) various subtasks on the Early Numeracy domain in Bolivia and Uganda; and (c) the Conflict Resolution and Perspective Taking subtasks in Uganda.

Table 4 summarizes the model fit statistics relevant for evaluating RQ2 and RQ3. Focusing first on RQ2, we found that the Unconstrained model fit the data much better than required by usual standards for covariance-based modeling. Other than the Hopping on One Foot subtask of the Motor domain, examination of modification indices did not indicate any items that loaded strongly onto other domains.

As reported in Table 5, in the Unconstrained model the factor correlations among the IDELA domains were quite large in all countries, with the highest correlations involving the Emergent Literacy domain. Given the strong correlations among the domains, it was plausible that a simpler model might be viable. We tested this hypothesis by comparing the fit of the Unconstrained

Table 3
Summary of Goodness of Fit for Within-Country, Within-Domain Models

Domain	Country	$\chi^2(df)$	RMSEA [90% CI]	TLI
Emergent literacy	Afghanistan	378.565 (232)	.016 [.013, .018]	.993
	Bolivia	284.873 (232)	.022 [.011, .030]	.984
	Ethiopia	331.857 (232)	.025 [.019, .031]	.993
	Uganda	267.575 (232)	.017 [.000, .026]	.997
	Vietnam	264.26 (232)	.014 [.000, .022]	.992
Emergent numeracy	Afghanistan	492.308 (302)	.015 [.013, .018]	.992
	Bolivia	378.516 (302)	.023 [.015, .030]	.989
	Ethiopia	387.751 (302)	.020 [.014, .026]	.996
	Uganda	340.642 (302)	.016 [.000, .024]	.993
	Vietnam	392.172 (302)	.021 [.015, .027]	.982
Motor	Afghanistan	72.592 (10)	.049 [.039, .060]	953
	Bolivia	13.060 (10)	.025 [.000, .059]	991
	Ethiopia	21.958 (10)	.042 [.018, .066]	987
	Uganda	20.620 (10)	.046 [.016, .074]	996
	Vietnam	11.710 (10)	.016 [.000, .046]	999
Social-Emotional	Afghanistan	143.738 (58)	.024 [.019, .029]	.984
	Bolivia	86.634 (58)	.032 [.016, .046]	.981
	Ethiopia	138.489 (58)	.045 [.036, .055]	.965
	Uganda	87.968 (58)	.032 [.017, .045]	.971
	Vietnam	116.053 (58)	.039 [.028, .049]	.969

Note. $\chi^2(df)$ 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. Sample sizes for each analysis are reported in Table 2.

tapraid5/z2p-devpsy/z2p-devpsy/z2p99918/z2p4178d18z | xppws | S=1 | 10/9/18 | 3:37 | Art: 2018-1591

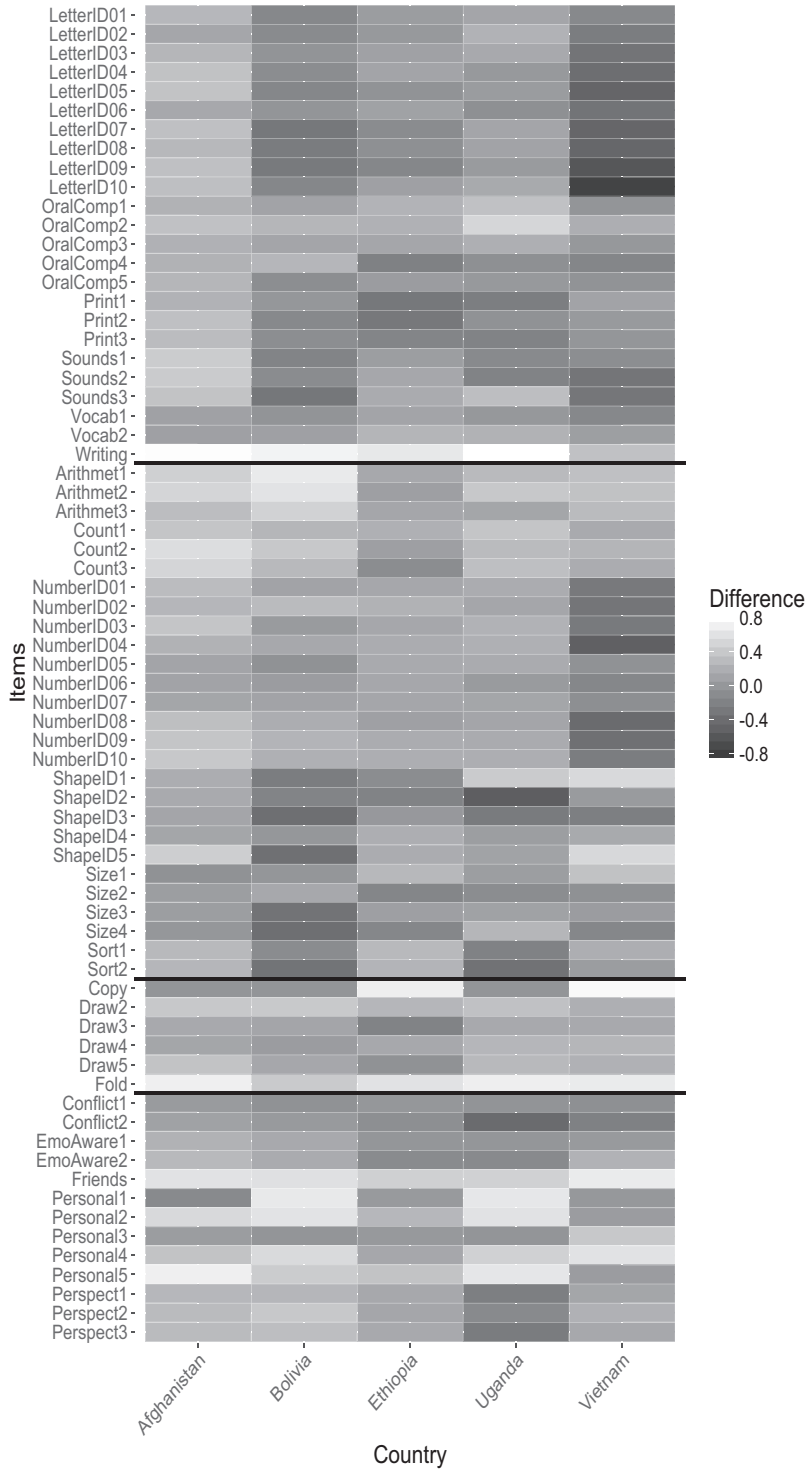


Figure 3. Difference between general and residual factor loadings for the IDELA items. Darker values denote items for which the standardized loading on the subtask was larger than the standardized loading on the focal factor. Estimates reported for the Unconstrained model. Horizontal lines separate the four domains, and variable codes are listed in Figure 1.

tapraid5/z2p-devpsy/z2p-devpsy/z2p99918/z2p4178d18z | xppws | S=1 | 10/9/18 | 3:37 | Art: 2018-1591

Table 4
Summary of Goodness of Fit for Within-Country, Across-Domain Models

Model	$\chi^2(df)$	RMSEA [90% CI]	TLI	χ^2 diff (*df*)	*p*-value
Afghanistan					
Unconstrained	2742.727 (2285)	.009 [.007, .010]	.987		
Unidimensional	2805.673 (2291)	.011 [.009, .013]	.986	119.496 (6)	<.001
Bolivia					
Unconstrained	2504.484 (2285)	.014 [.010, .018]	.972		
Unidimensional	2576.870 (2291)	.016 [.012, .019]	.963	102.427 (6)	<.001
Ethiopia					
Unconstrained	2570.255 (2285)	.014 [.010, .016]	.988		
Unidimensional	2593.022 (2291)	.014 [.011, .017]	.987	38.94 (6)	<.001
Uganda					
Unconstrained	2424.256 (2285)	.011 [.005, .015]	.990		
Unidimensional	2450.040 (2291)	.012 [.006, .016]	.988	53.524 (6)	<.001
Vietnam					
Unconstrained	2760.415 (2285)	.018 [.015, .020]	.960		
Unidimensional	2863.813 (2291)	.019 [.017, .021]	.952	156.539 (6)	<.001

Note. $\chi^2(df)$ 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 its degrees of freedom, with the *p*-value reported in the last column. Sample sizes for each analysis are reported in Table 2.

model to that of the Unidimensional model. As shown in Table 4 the Unidimensional model also fit the data from each country quite well, but in each case it was found to have worse fit than the Unconstrained model, as assessed by a chi-square difference test. We also considered various other simplified domain models, which are not reported in the table. In particular, we collapsed only Emergent Literacy and Emergent Numeracy in each country, but these models were also rejected by chi-square difference tests against the Unconstrained model. Thus, while the domain factors were often highly correlated, they were nonetheless statistically distinct in each of the countries. We address this point further in the Discussion.

Measurement Invariance Analysis (RQ4 and RQ5)

The foregoing analyses provided evidence that the conceptual four-domain model of the IDELA provided acceptable fit to the data from all five countries. Our next step was to evaluate whether the domains were measurement invariant over countries. The results from the measurement invariance analysis are summarized in Table 6. It can be seen that the scalar invariance model was unequivocally rejected for all domains. The online supplemental materials report a number of secondary analyses described in the Analytic Approach that were conducted to ensure that this conclusion was not contingent on the model specification or the specific sub samples used in the analyses. In addition, based on Figure 3 we also suspected that the Vietnamese sample might be driving the dissimilarity among the countries. Therefore we also conducted secondary analyses omitting the Vietnamese sample. Each of these robustness checks led to the same conclusion: unequivocal rejection of the scalar invariance model.

Next we explored whether a partial invariance model was feasible for each domain using a forward selection procedure with the MIMIC model (see Analytic Approach). Based on examination of the resulting partial invariance model, we concluded that most items on each domain exhibited DIF over two or more countries. It was not apparent that the invariant items corresponded to any particular subtask or domain, or that any one country was disproportionately responsible for the DIF. Figure S1 in the online supplemental materials summarizes the resulting partial invariance model in terms of the country-specific deviations on the item thresholds.

Finally, Figure 4 addresses the pragmatic implications for comparisons over countries. This figure reports the factor means on each domain, from the scalar invariance models and the partial invariance models. The difference between the two sets of estimates is attributable to lack of fit of the scalar models, and consequently the estimates from the partial invariance models are regarded as more appropriate for the present data. For all domains except Social-Emotional, the magnitude of bias induced by assuming scalar invariance was severe enough to lead to incorrect comparisons among countries. For instance, using the scalar invariance model for Early Literacy, it would be concluded that Bolivian children are about .3 standard deviation units below average. However, under the partial invariance model, Bolivian children are slightly above average on Early Literacy. As another example, Afghan children were significantly below average on

Table 5
Correlations Among Factors, Unconstrained Models

Correlations	Country				
	Afghanistan	Bolivia	Ethiopia	Uganda	Vietnam
Literacy/Numeracy	.950	.889	.958	.938	.925
Literacy/Motor	.933	.881	.918	.920	.927
Literacy/SE	.877	.807	.917	.880	.830
Numeracy/Motor	.911	.679	.947	.838	.848
Numeracy/SE	.897	.690	.881	.830	.667
Motor/SE	.820	.624	.781	.846	.649

Note. Table reports point estimates of the factor correlations among the domains, in the Unconstrained model, from each of the five countries. SE denotes the Social-Emotional domain.

Table 6
Summary of Goodness of Fit for Configural and Scalar Models, by Domain

Model	$\chi^2(df)$	RMSEA [90% CI]	TLI	χ^2 diff (*df*)	*p*-value
Emergent literacy					
Configural	458.032 (232)	.020 [.017, .023]	.988		
Scalar	952.6779 (324)	.028 [.026, .031]	.977	512.540 (92)	<.001
Emergent numeracy					
Configural	727.810 (301)	.024 [.022, .027]	.983		
Scalar	2417.892 (404)	.046 [.044, .047]	.941	1811.417 (104)	<.001
Motor					
Configural	4.744 (5)	.000 [.000, .028]	.999		
Scalar	527.316 (25)	.091 [.085, .098]	.881	511.288 (20)	<.001
Social-Emotional					
Configural	224.940 (57)	.035 [.030, .040]	.952		
Scalar	797.192 (105)	.052 [.049, .056]	.893	594.338 (48)	<.001

Note. $\chi^2(df)$ 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 its degrees of freedom, with the *p*-value reported in the last column. All analyses used a random sample of size 480 from each country, with the total sample size equal to 2,400.

Early Numeracy under the scalar model, but were significantly above average under the partial invariance model.

Discussion

With recent calls for inclusion of national data on ELD in global monitoring mechanisms such as the United Nation's SDGs, it is important to critically assess the cross-country comparability of ELD assessments using up-to-date psychometric methods. To our knowledge, the present study is the first to distinguish multiple types of measurement invariance (e.g., configural, partial, and scalar) and examine lack of invariance at the item level in an international context.

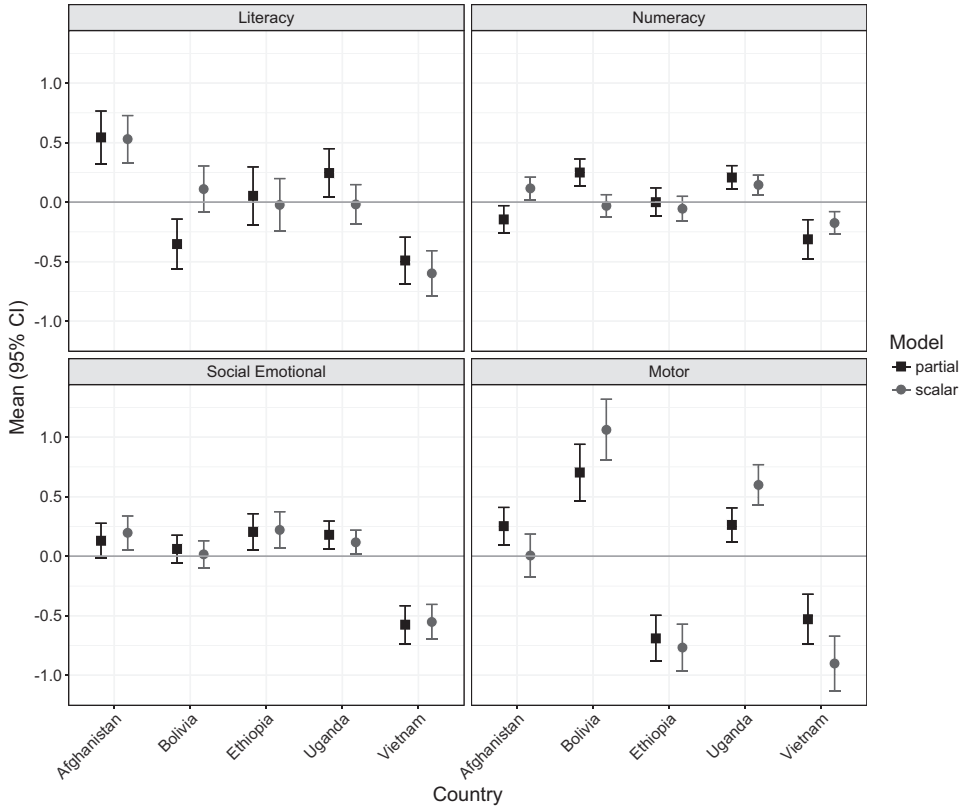


Figure 4. Factor means estimated from the partial invariance models and the scalar invariance models. For each domain and each model, the country means are scaled so that the overall mean for all countries is zero, and the means are standardized by the within-country standard deviation. Error bars denote approximate 95% confidence intervals using a normal reference distribution.

tapraid5/z2p-devpsy/z2p-devpsy/z2p99918/z2p4178d18z | xppws | S=1 | 10/9/18 | 3:37 | Art: 2018-1591

We evaluated whether the conceptual multidomain model of one ELD assessment, the IDELA, was supported by empirical data from five low- and middle-income countries from diverse regions of the world. Based on past research, we represented the IDELA's conceptual model statistically using a bifactor model with four general domains. This model fit the data well in each country. However, the correlations among the domains were quite strong, especially those involving the Emergent Literacy domain. Because the IDELA is administered verbally and involves numerous prompts with printed materials, these strong correlations may indicate that many IDELA items are measuring basic literacy to some extent. Moreover, Wolf et al. (2017) assessed whether the correlations among the four IDELA domains were compatible with a higher-order domain of school readiness, finding that the higher-order model fit the data about as well as the unconstrained four-domain model reported here. These considerations raise the questions about whether multidomain models of ELD are necessary to explain observed data, or if a simpler model may be sufficient.

On the other hand, work conducted in the United States has shown that different domains of ELD are differentially predictive of later academic success (e.g., Duncan et al., 2007). Wolf et al. (2017) examined subgroups used for evaluation of an ECCD program, finding that the program raised performance on all IDELA domains except Social-Emotional. The latter domain was least highly correlated with all other domains across all five countries considered in the present study.

To provide more evidence about the distinctness of the IDELA domains, we tested whether the four-domain model could be replaced by a similar model with a single, overarching domain. The global fit indices of the four-domain and single-domain models were very similar in all countries (see Table 3), which is to be expected since the models are only differentiated by a small subset of their parameters (i.e., the single-domain model shares 201 of its 202 parameters with the four-domain model). To facilitate a comparison that focuses only on the parameters that differentiated the two models, we used a chi-square difference test. In each country, we rejected the hypothesis that the simpler model fit the data as well as the four-domain model. Thus, while the domains were highly correlated, statistical evidence supported the conclusion that they were distinct in each country. Given the complexity of the assessment (4 domains, 22 subtasks, 76 items), the diversity of the samples considered, and the relatively thorough analysis conducted, this is an encouraging finding about the potential to reliably measure multidomain models of ELD in international settings.

The results of our cross-country comparisons were less optimistic. The main take-aways of the measurement invariance analyses were that, for each domain, (a) there is overwhelming evidence against the assumption of scalar measurement invariance over countries, and (b) this conclusion did not change when balancing on preexisting subgroups that might plausibly account for heterogeneity among the samples considered in this study. Based on exploratory follow-up analyses, we concluded that (c) the lack of invariance was not due to a small subset of items or isolated to any single country (Figure S1). In addition, (d) ignoring DIF led to a number of incorrect conclusions about the magnitude, direction, and significance of mean differences among the countries (see Figure 4).

Taken as a whole, the IDELA domains do not support unbiased cross-country comparisons, and it was also not evident that any particular subset of items can serve this purpose. There are several possible explanations of this finding. First, it may be the case that the populations of children considered in this study were simply too heterogeneous to be compared on the domains of ELD represented by the IDELA. The results of our secondary analyses (see online supplemental materials) provided some preliminary indication that, even after balancing the samples on gender, age, and exposure to formal early child care, the overall conclusions about measurement invariance did not change. However, there are numerous potential dimensions on which to balance these samples, and we acknowledge that our attempts at matching were post hoc and the data sets were not nationally representative.

Another explanation is that the translation and adaptation of the IDELA items may have resulted in construct-irrelevant variation that systematically differed over countries. On this hypothesis, we might expect that the Emergent Literacy and Social-Emotional domains would be relatively more susceptible to measurement bias, since these items require relatively more translation and input from local collaborators when they are adapted. This hypothesis is also plausible in light of research on international educational surveys designed for older children, which has shown that mathematics and science tend to be more comparable than reading and language skills (see, e.g., Asil & Brown, 2016; Zwitser et al., 2017). However, we did not find strong evidence of such a pattern in the present study. In fact, in terms of mean differences (see Figure 4), the Social-Emotional domain was the least influenced by DIF.

A third explanation is that cross-country DIF was due to the incommensurability of the domains of interest over different cultures. While children's development consists of overall, basic processes that share some consistency from one place to another, cultural and environmental differences in how these processes manifest themselves in children's behaviors may present a significant impediment to designing highly nuanced measures of ELD for use in international settings. This explanation suggests that, as multidomain conceptualizations of ELD become increasingly detailed and well specified, so too do they become increasingly sensitive to cultural and contextual factors that may not lend themselves to global comparability.

Implications for Measuring Early Learning and Development Globally

At a broad level, our findings suggest that the design of assessments to support international comparisons about ELD and school readiness may face substantial psychometric challenges. In particular, it would seem useful to reconcile the richness and nuance of multidomain conceptualizations of ELD with the complexity of making generalizations in international settings. Our findings show that the domains of ELD measured by the IDELA tend to be highly correlated with one another, suggesting that one way forward is to simplify the conceptual model in question. This approach has proven useful in other large-scale assessment contexts. For example, the National Assessment of Educational Progress (NAEP) treats mathematics in terms of five content domains: numbers and operations, measurement, geometry, probability, and algebra (Johnson, Lazer, & O'Sullivan, 1997). However, NAEP scores are reported

tapraid5/z2p-devpsy/z2p-devpsy/z2p99918/z2p4178d18z | xppws | S=1 | 10/9/18 | 3:37 | Art: 2018-1591

on an overall mathematics scale, rather than on the five different content domains, which is justified by the high correlations among the domains. In the present context, such an approach could simplify assessment design (e.g., reduce the number of items and subtasks), as well as the eventual comparison of scores across countries (e.g., van Buuren, 2014).

While a conceptually simplified ELD assessment may present a viable option for comparisons among countries, it is also important to consider the potential shortcomings of such an approach. For example, the primary purpose of the United Nation's SDGs is to drive country-specific progress on the Goals (United Nations, 2015). While comparison among countries may facilitate this purpose, it will also be important that countries receive sufficiently detailed feedback to support useful policy decisions. In this light, the additional nuance offered by multidomain models of ELD and school readiness may be very valuable, and perhaps more valuable than purely comparative information. The results of the present study suggest that, within countries, current ELD assessments such as the IDELA can be useful for monitoring programs in early childhood development, as well as for evaluating specific quality-improvement initiatives. While these goals are aligned with the original intentions of the developers (Pisani et al., 2015), the appropriateness of the IDELA, or of any ELD assessment, should be evaluated on a case-by-case basis.

Given the large number of purposes that assessments of ELD can serve in international settings (see Goldstein & Flake, 2016), it is unrealistic to expect that any single assessment will provide a proverbial silver bullet. To address both international comparability as well as within-country monitoring and evaluation, one possible solution is to design assessments with both short and long versions, the former comprised of relatively generic but reliable content, and the latter providing detailed, actionable feedback about specific subdomains of ELD and school readiness. Numerous procedures for creating alternative forms (e.g., short versions) of an assessment are available in the psychometric literature (see, e.g., Joint Committee for Educational and Psychological Testing, 2014).

Limitations and Future Directions

As mentioned, a major limitation of the present study is that the samples of children were not nationally representative within countries and were not balanced across countries. It would be informative to replicate and extend the analyses reported here with such a data set. Additionally, this study focused on assessment items that were shared across the versions of the IDELA used in the reported studies, and did not evaluate the full set of items on the current version of the IDELA items; the current version has four items that were not reported on here, in addition to those that were omitted during the analysis (see Results section).

A major methodological limitation was the use of purely exploratory procedures to address DIF. A cross-validation study that first explores item bias and then establishes out-of-sample generalizability would be helpful for making decisions about potential item revisions. It would also be important to ensure that item revisions do not depend on the item selection procedure used. Because the structure of the IDELA is quite complex (domain, subtasks, and items), nonparametric procedures for screening items across countries could greatly simplify further analyses, especially when considering larger numbers of countries.

Conclusion

The results of this study provide initial evidence that the four-domain conceptual model of the IDELA generalizes across each of five diverse low- and middle-income countries (Afghanistan, Bolivia, Ethiopia, Uganda, and Vietnam). However, in its present form, the IDELA is not suitable for making comparisons across countries. We suspect this issue is not particular to the IDELA, but is reflective of cultural and contextual variation in expectations about child development at the level of specific skills and competencies. Additional research is needed to identify a core subset of items that may be used to support cross-country comparisons.

References

American Psychological Association. (2014). *APA dictionary of statistics and research methods*. Washington, DC: Author.

Asil, M., & Brown, G. T. L. (2016). Comparing OECD PISA reading in English to other languages: Identifying potential sources of non-invariance. *International Journal of Testing, 16,* 71–93. http://dx.doi.org/10.1080/15305058.2015.1064431

Asparouhov, T., & Muthén, B. O. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling, 21,* 495–508. http://dx.doi.org/10.1080/10705511.2014.919210

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods, 22,* 507–526. http://dx.doi.org/10.1037/met0000077

Black, M. M., Walker, S. P., Fernald, L. C. H., Andersen, C. T., DiGirolamo, A. M., Lu, C., . . . Grantham-McGregor, S. (2017). Early childhood development coming of age: Science through the life course. *The Lancet, 389,* 77–90. http://dx.doi.org/10.1016/S0140-6736(16)31389-7

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychological Science, 5,* 184–186. http://dx.doi.org/10.1177/2167702616657069

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika, 75,* 33–57. http://dx.doi.org/10.1007/s11336-009-9136-x

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43,* 1428–1446. http://dx.doi.org/10.1037/0012-1649.43.6.1428

Gladstone, M., Lancaster, G. A., Umar, E., Nyirenda, M., Kayira, E., van den Broek, N. R., & Smyth, R. L. (2010). The Malawi Developmental Assessment Tool (MDAT): The creation, validation, and reliability of a tool to assess child development in rural African settings. *PLoS Medicine, 7*(5), e1000273. http://dx.doi.org/10.1371/journal.pmed.1000273

Goldstein, J., & Flake, J. K. (2016). Towards a framework for the validation of early childhood assessment systems. *Educational Assessment, Evaluation and Accountability, 28,* 273–293. http://dx.doi.org/10.1007/s11092-015-9231-8

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. http://dx.doi.org/10.1080/10705519909540118

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/*

tapraid5/z2p-devpsy/z2p-devpsy/z2p99918/z2p4178d18z | xppws | S=1 | 10/9/18 | 3:37 | Art: 2018-1591

Revue Canadienne des Sciences du Domportement, 39, 1–22. http://dx.doi.org/10.1037/cjbs2007001

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika, 76,* 537–549. http://dx.doi.org/10.1007/s11336-011-9218-4.Exploratory

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2013). Modeling differential item functioning using a generalization of the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics, 38,* 32–60. http://dx.doi.org/10.3102/1076998611432173

Johnson, E. G., Lazer, S., & O'Sullivan, C. Y. (1997). *NAEP reconfigured: An integrated redesign of the National Assessment of Educational Progress*. Washington DC: National Center for Education Statistics.

Joint Committee on Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA, APA, & NCME.

Kagan, S. L., & Britto, P. R. (2005). *Going global with indicators of child development*. New York, NY: UNICEF.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking* (3rd ed.). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4939-0317-7

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 557–585. http://dx.doi.org/10.1007/BF02296397

Muthén, B. O., Du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished technical report. Retrieved from http://www.statmodel.com/download/Article_075.pdf

Muthén, B. O., & Satorra, A. (1995a). Technical aspects of Muthén's LISCOMP approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika, 60,* 489–503. http://dx.doi.org/10.1007/BF02294325

Muthén, B. O., & Satorra, A. (1995b). Complex sample data in structural equation modeling. *Sociological Methodology, 25,* 267–316. http://dx.doi.org/10.2307/271070

Muthén, L. K., & Muthén, B. O. (2018). *Mplus user's guide* (8th ed.). Los Angeles, CA: Author.

Mwaura, P. A., Sylva, K., & Malmberg, L. E. (2008). Evaluating the Madrasa preschool programme in East Africa: A quasi-experimental study. *International Journal of Early Years Education, 16,* 237–255. http://dx.doi.org/10.1080/09669760802357121

NICHD Early Child Care Research Network. (2005). *Child care and development: Results from the NICHD Study of Early Child Care and Youth Development*. New York, NY: Guilford Press.

Pisani, L., Borisova, I., & Dowd, A. J. (2015). *International development and early learning assessment technical working paper*. Washington, DC: Save the Children.

Purpura, D. J., Hume, L. E., Sims, D. M., & Lonigan, C. J. (2011). Early literacy and early numeracy: The value of including early literacy skills in the prediction of numeracy development. *Journal of Experimental Child Psychology, 110,* 647–658. http://dx.doi.org/10.1016/j.jecp.2011.07.004

Raikes, A., Yoshikawa, H., Britto, P. R., & Iruka, I. (2017). Children, youth and developmental science in the 2015–2030 global Sustainable Development Goals. *Social Policy Report, 30,* 1–23. http://dx.doi.org/10.1002/j.2379-3988.2017.tb00088.x

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, NY: United Nations Children's Fund.

Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? Application of iteratively reweighted least squares to the Rosenberg Self-Esteem Scale. *Multivariate Behavioral Research, 51,* 818–838.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bifactor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47,* 361–372. http://dx.doi.org/10.1111/j.1745-3984.2010.00118.x

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the saled difference chi-square test statistic. *Psychometrika, 75,* 243–248. http://dx.doi.org/10.1007/s11336-009-9135-y

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

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

Sörbom, D. (1989). Model modification. *Psychometrika, 54,* 371–384. http://dx.doi.org/10.1007/BF02294623

Squires, J., & Bricker, D. (2009). *Ages & Stages Questionnaires (ASQ-3: A parent-completed child monitoring system* (3rd ed.)). Baltimore, MD: Brookes.

UNESCO. (2013). *Toward universal learning: Recommendations from the Learning Metrics Task Force*. Montreal, Canada and Washington, DC: UNESCO Institute for Statistics and Center for Universal Education at the Brookings Institution.

UNICEF. (2017). *Early learning and development standards (ELDS) and school readiness: Evaluation report*. New York, NY: Author.

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

United Nations. (2017). *Revised list of global Sustainable Development Goals indicators*. New York, NY: Author.

van Buuren, S. (2014). Growth charts of human development. *Statistical Methods in Medical Research, 23,* 346–368. http://dx.doi.org/10.1177/0962280212473300

Verdisco, A., Cueto, S., & Thompson, J. (2016). *Early childhood development: Wealth, the nurturing environment and inequality. First results from the PRIDI database*. Washington, DC: Inter-American Development Bank. http://dx.doi.org/10.18235/0000501

Wolf, S., Halpin, P., Yoshikawa, H., Dowd, A., Pisani, L., & Borisova, I. (2017). Measuring school readiness globally: Assessing the construct validity and measurement invariance of the International Development and Early Learning Assessment (IDELA) in Ethiopia. *Early Childhood Research Quarterly, 41,* 21–36. http://dx.doi.org/10.1016/j.ecresq.2017.05.001

World Health Organization, UNICEF, & World Bank Group. (2016). Advancing early childhood development: From science to scale. *The Lancet Series*. Retrieved from http://www.who.int/maternal_child_adolescent/documents/early-child-development-lancet-series/en/

Wuermli, A. J., Helm, J., Hastings, P. D., Yoshikawa, H., & Dowd, A. J. (2017). *Investigating the psychometric properties and measurement invariance of the IDELA in a diverse sample from Bhutan*. Unpublished technical report.

Zwitser, R. J., Glaser, S. S. F., & Maris, G. (2017). Monitoring countries in a changing world: A new look at DIF in international surveys. *Psychometrika, 82,* 210–232. http://dx.doi.org/10.1007/s11336-016-9543-8

Received January 24, 2018
Revision received July 27, 2018
Accepted August 2, 2018 ■