



Validation of motor, cognitive, language, and socio-emotional subscales using the Caregiver Reported Early Development Instruments: An application of multidimensional item factor analysis

Citation

Waldman, M., McCoy, D. C., Seiden, J., Cuartas, J., CREDI Field Team, & Fink, G. (2021). Validation of motor, cognitive, language, and socio-emotional subscales using the caregiver reported early development instruments: an application of multidimensional item factor analysis. *International Journal of Behavioral Development*, 45(4), 368-377.

Published version

<https://doi.org/doi.org/10.1177/01650254211005560>

Link

<https://dash.harvard.edu/handle/1/42717582>

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles (OAP), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available. Please share how this access benefits you. [Submit a story](#)

**Validation of Motor, Cognitive, Language, and Socio-emotional Subscales
using the Caregiver Reported Early Development Instruments: An
Application of Multidimensional Item Factor Analysis**

Marcus Waldman, University of Nebraska Medical Center

Dana Charles McCoy, Harvard Graduate School of Education

Jonathan Seiden, Harvard Graduate School of Education

Jorge Cuartas, Harvard Graduate School of Education

CREDI Field Team*

& Günther Fink, Swiss Tropical and Public Health Institute

*CREDI Field Team comprised of (in alphabetical order): Elisa Altafim, Alexandra Brentani, Andreana Castellanos, Alexandra Chen, Anne Marie Chomat, Wafaie Fawzi, Cristina Gutierrez de Piñeres, Jena Hamadani, Natalia Henao, Pamela Jarvis, Codie Kane, Jeffrey Measelle, Patricia Medrano, Lauren Pisani, Muneera Rasheed, Peter C. Rockers, Jonathan Seiden, Christopher R. Sudfeld, Fahmida Tofail, Christine Wong, Dorianne Wright, & Aisha K. Yousafzai

VALIDATION OF SUBSCALES FROM THE CREDI

A growing body of research shows that early childhood is a sensitive period of brain and skill development and has the largest individual and social returns to investments relative to other periods of human development (Grantham-McGregor et al., 2007; Heckman, 2006; Lu, Black, & Richter, 2016; Moffitt et al., 2011; Nores & Barnett, 2010; Peet et al., 2015). Reflecting this promise, the past several decades has seen a surge in global interest in promoting early childhood development (ECD), particularly during the first one thousand days of life (Black et al., 2017). A broad range of ECD intervention approaches (e.g., home visiting programs, early childhood care and education services, nutritional supports) have been developed to meet the needs of children living in diverse settings around the world, and are increasingly being prioritized by governments and non-governmental organizations for large-scale implementation (Richter et al., 2017). At a policy level, the United Nations' recently ratified Sustainable Development Goals (SDGs) which specifically focus on ECD under Target 4.2. In fact, Target 4.2 under the SDGs represents the first major global policy initiative to specifically focus on ECD.

Central to the success of ECD intervention and policy efforts is access to reliable, valid, and practically feasible methods for measuring young children's outcomes. In particular, experts have highlighted the need for global instruments that can be used to capture multiple domains of development (e.g., motor skills, language skills, etc.) in large, culturally diverse samples (Richter et al., 2019). Such approaches are critical for a multitude of purposes, ranging from improving basic understanding of developmental processes globally to evaluating the impact of programs and policies on child outcomes to monitoring progress toward global policy targets.

The large-scale implementation of existing measures of motor, cognitive, language, and socio-emotional development in children younger than three years of age is likely not feasible in international contexts. Existing ECD instruments include the *Denver Developmental Screening*

VALIDATION OF SUBSCALES FROM THE CREDI

Test (DDST; Frankenburg & Dodds, 1967), the *Bayley Scales of Infant and Toddler Development* (BSID-III; Bayley, 2006), and the *Ages & Stages Questionnaire* (ASQ-3; Squires & Bricker, 2009). These instruments provide information about ECD with enough precision to screen individual children for developmental disabilities or delays. However, these instruments were primarily constructed for U.S. populations and, with some exceptions (e.g., Kerstjens et al., 2009), there is limited evidence on their validity in international contexts (Peña, 2007). Furthermore, the costs and resources associated with purchase and implementation make these instruments difficult to implement in large samples, particularly in resource-limited low- and middle-income countries (LMICs).

In recent years, a number of instruments have been developed to address the need for cross-culturally comparable ECD measures. For example, Save the Children's *International Development and Early Learning Assessment* (IDELA) has shown evidence for validity and easy implementation in international contexts, but it is intended to measure learning and development for children 3.5- to 6.5-years-old (Halpin et al., 2019; Wolf et al., 2017). Similarly, the Inter-American Development Bank's Regional Project on Child Development (PRIDI), a direct assessment tool, seeks to measure two- to four-year-olds' motor, cognitive, language, and socio-emotional development using a brief set of indicators that are considered to be valid in culturally diverse contexts. A final example is the *INTERGROWTH-21st Project Neurodevelopment Package* (INTER-NDA), which was calibrated and tested in eight multiethnic sites across five continents, but only targets 22- to 26-month-old children (Fernandes, 2014). Given the particularly high plasticity of development during the first three years of life (Walker et al., 2011), there is an urgent need for scalable, internationally-validated instruments to monitor child development in this specific developmental period.

VALIDATION OF SUBSCALES FROM THE CREDI

In response to the limitations of existing ECD measures, we developed the Caregiver Reported Early Development Instruments (CREDI; [citation redacted; citation redacted]). The CREDI is a simple, caregiver-reported measure developed for large-scale assessment of ECD for children between the ages of zero and three years. The CREDI exists in both a short form and a long form. The short form aims to provide policymakers and NGOs with a single score of overall ECD, and these scores have demonstrated validity evidence in 17 low-, middle-, and high-income countries (McCoy et al., 2018). In contrast, the purpose of the long form is to provide finer-grain information regarding children's development across multiple domains, including (a) motor skills (including fine and gross motor skills), (b) cognitive skills (including executive functioning, reasoning, problem solving, and pre-academic knowledge), (c) language skills (including expressive and receptive language skills), and (d) socio-emotional skills (including emotional and behavioral self-regulation, emotion knowledge, and social competence). A thorough discussion of the instrument's construction (i.e., item construction, data collection procedures, etc.) as well as the psychometrics of the short form is provided by [citation redacted].

The aim of the present study is to report the validity evidence for the CREDI long form's motor, cognitive, language, and socio-emotional subscale scores obtained from $N = 14,113$ caregiver reports in a multicultural, multinational sample. In assessing the evidence, we followed the recommendations set forth by the *Standards for Educational and Psychological Testing* (henceforth, *Standards*; American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) to evaluate whether subscale scores obtained from the CREDI long form support inferences about the developmental status of children under three years of age. Complementing existing evidence regarding the CREDI's test content and cognitive testing provided in [citation

VALIDATION OF SUBSCALES FROM THE CREDI

redacted], this study provides evidence of the long form's: (1) *Construct validity* and the internal structure of the CREDI long form, including the extent to which observed item response patterns are predicted by theory; (2) *Criterion validity* and the degree to which relations between the CREDI subscale scores and other variables match what would be expected by theory; (3) *Reliability* in that the subscale scores drawn from the CREDI long form are sufficiently precise for the intended purpose of the instrument (i.e., population measurement of young children's developmental status in multiple developmental domains); (4) *Fairness* in that scores result in systemically biased conclusions about the developmental status of particular groups of children. To provide this evidence, we begin by comparing a variety of potential model specifications, including several multidimensional models that allow CREDI items to load onto multiple subscales simultaneously. We argue that such a multidimensional approach is more conceptually valid for capturing ECD during infancy and toddlerhood, when children's observable behaviors often reflect multiple underlying skills or capacities (e.g., pointing as reflecting both expressive communication and motor skills).

Measures and Methods

Participants

We collected data from a sample of 14,113 primary caregivers of children aged 0-35 months old from 21 sites across 17 high- and- LMICs.¹ The mean age of children was 20.3 months ($SD = 9.41$). Approximately half of the children were male (50.2%). Geographically, 53.2% of respondents were from Africa (Ghana, Tanzania, and Zambia), 20.68% from Asia (Bangladesh, Cambodia, India, Jordan, Laos, Nepal, Pakistan, and the Philippines), 21.0% from Latin America (Brazil, Chile, Colombia, and Guatemala), and 6.4% from the United States. (See Table 1 for details about the sample). The CREDI was translated (and back-translated) from

¹ This figure differs from the analytic sample size in [citation redacted] because we did not require that at least 75 percent of the items were observed (not missing) to remain included in the sample.

VALIDATION OF SUBSCALES FROM THE CREDI

English to local languages in all sites. Within each site, surveys were administered to children participating in local research projects. Although samples were predominantly convenience-based, several sites (e.g., Brazil, Nepal, Cambodia) included samples that were representative of subnational units (e.g., districts or zones).

The study was reviewed by each site's [*Institution redacted*] Institutional Review Board, and all data collection was conducted in accordance with local ethical standards. All caregivers gave informed consent.

CREDI

In administering the CREDI, we asked caregivers to report whether their child can or does exhibit a range of milestones, skills, and behaviors compiled to measure motor, cognitive, language, and socio-emotional development for children under 36 months of age. We developed and refined the wording of the items refined using a multi-phase process that has been documented previously (see *citation redacted*). Overall, we field-tested 149 items. Caregivers responded to up to 103 dichotomous items that were identified as appropriate given the child's age. Caregivers could answer all CREDI items with a "yes," "no," or "I don't know" response. We treated all "I don't know" responses as missing values in the analysis.

Of the 149 items tested, we excluded 39 from further analysis as these items: (1) showed >10% "don't know" responses, (2) were understood by fewer than 80% of caregivers on cognitive interviews, (3) showed poor agreement levels (unadjusted for chance agreement) of Cohen's $\kappa < .40$ for the caregivers selected to respond to the same questions 7-10 days later, or (4) were identified as primarily measuring mental health and did not demonstrate a monotonic relationship with age. Further details on the item-screening process are provided in [*citation redacted*]. Consequently, we analyzed responses to a total of 110 items in this study.

Construct Validity

Consistent with the recommendations from the *Standards*, we gathered construct validity evidence by demonstrating that there is a theoretical basis for explaining item response patterns (i.e., the internal structure). In developing ECD instruments, exploratory factor analysis (EFA) is often used as a starting point for evaluating the internal structure of the items, including ascertaining the dimensionality of the instrument and assessing internal structure using factor loadings (e.g., Fernandes, 2014; Ghandour et al., 2019). In contrast, in educational assessment and the item response theory literature, test developers often employ confirmatory approaches (e.g., confirmatory factor analysis [CFA]) in which the loading structure of items to constructs is pre-specified according to a panel of experts (Liu & Kang, 2019). Both approaches—EFA and CFA—have advantages and disadvantages. On the one hand, with CFA, there is no guarantee that the theoretical loading structure specified by a panel of experts best explains item responses. On the other hand, traditional EFA models make strong distributional (i.e., normality of the underlying factors) and parametric (i.e. linearity) assumptions that likely do not hold perfectly in real-world data; consequently, solutions from EFA that differ from theoretical expectations may be reflective of the sensitivity of the parameter estimates to assumption violations when modeling the data, as opposed to an accurate indication of the true underlying structure of the instrument. Indeed, although a traditional EFA was conducted (contact first author for details), the EFA solution was determined to be inconsistent with theory because all but six items (two motor items and four socio-emotional items) loaded onto two factors and these two factors had no clear theoretical delineation to theorized ECD constructs.

Our approach for testing the internal structure of the CREDI long form utilized a hybrid between CFA and EFA. Consistent with CFA, we fit multidimensional item factor analysis models (IFA) models to the data using a theoretically grounded factor loading structure

VALIDATION OF SUBSCALES FROM THE CREDI

developed by a team of 16 external expert advisors. However, we also use disagreements among the panel of experts to specify alternative models and evaluate the corresponding fit. Thus, our hybrid approach attempted to strike a balance between identifying a factor loading specification that maximizes model-data consistency (i.e., the goal of EFA) while ensuring that we remain tied to theory (i.e., the goal of CFA).

To begin this process, our panel of expert advisors analyzed each item and voted for all ECD domains that the item was hypothesized to measure (i.e., motor, cognitive, language, and/or socio-emotional; see Appendix Table 2 for a fully tally of all expert votes). These experts included developmental psychologists and pediatricians representing a range of countries. Items with potential implications for multiple areas of development could be flagged as representing more than one domain of development so as to allow for cross-loadings. In other words, unlike what has been done in traditional ECD instruments that provide subscores (c.f. Bayley, 2006; Fernandes, 2014; Squires & Bricker, 2009), we specify cross-loadings and do not require that items are assigned to one and only one ECD domain. We hypothesized that specifying the presence of cross-loadings would better reflect the internal structure of the data because any given item may indicate children's development across several domains. This is especially likely in the first three years of development when children's observable behaviors often reflect multiple different skills and capacities. For example, most traditional measures of ECD claim infants' use of gestures (e.g., pointing, grabbing) as a "pure" representation of their language abilities, whereas it is likely that these behaviors also reflect skills in motor development (Bowman et al., 2017).

We tested three alternative loading specifications by varying the minimum number of expert votes required to freely estimate a loading across the factors (i.e., domains). These four-factor IFA models are visualized in Figure 1. In the first model (Model A.1), loadings were

VALIDATION OF SUBSCALES FROM THE CREDI

freely estimated if at least eight (of the 16) experts agreed that the item loaded on a domain. The second model (Model A.2) and third model (Model A.3) reduced the required number of votes to free a loading to six and four, respectively. We did not test less restrictive specifications, as freeing loadings with fewer than four votes led to convergence issues.

In all models, we relaxed the (unconditional) multidimensional normality assumption traditional in multidimensional IFA models because such an assumption is likely untenable. For example, we did not think it would be plausible to assume that motor subscale scores for all children aged 0-35 months would follow a symmetric distribution as would be implied by the traditional normality assumption. In other words, a symmetric assumption would imply that motor scores followed linear age gradients, whereas we expect nonlinear gradients with the fastest rate of change occurring early in development and then tapering with age. The result of a tapered age gradient would be a left-skewed marginal distribution of motor scores.

To accommodate nonlinear age gradients, all four factors were modeled with a linear, quadratic, and cubic function of age as covariates. In this way, subscale scores were assumed to be multivariate normally distributed for children of the same age, even if the marginal distribution is not normally distributed. Intercepts were fixed to zero and residual variances were fixed to one for model identification. In addition, we included site fixed effects (with the sample from Jordan as the reference group) to account for planned missingness, as not all items were administered in all sites. (Items exhibited an average missingness rate of 50.3% and ranged from 14.6% to 76.9% across sites.) We employed maximum likelihood estimation, which assumes that data are missing at random (MAR) conditional on the observed responses, age, and between site differences in factor scores. Analyses were conducted in Mplus Version 8.3 (Muthén & Muthén, 2017).

VALIDATION OF SUBSCALES FROM THE CREDI

To minimize overfitting and maximize model-data consistency, we next pruned Models A.1-A.3, fitted using the loading specifications. In theory, after reverse-coding items (as appropriate), all items should be positively correlated with the specified developmental domain, implying that all loading estimates should result in positive values. In fitting the models, however, we encountered overfitting behaviors where negative loading estimates on one domain often accompanied unreasonably strong positive loading estimates on another domain. For a small subset of items, this undesirable compensating behavior was so severe when fitting Model A.3 that it led to convergence problems. We considered the instability induced by this compensation as an indication of overfitting to our sample because theory would suggest that all factor loadings in the population would be positive. Overfitting implies that the model is overly complex and does not optimize predictive fit to out-of-sample data compared to a more parsimonious model (c.f., Hastie et al., 2009). Thus, a current focus in measurement and structural equation modeling is developing methods to minimize overfit by reducing model complexity in order to improve the generalizability of inferences (e.g., Jacobucci et al., 2016). Our approach to reduce model complexity was to specify linear inequality constraints that required loading estimates to take on non-negative values only. Loadings with estimates at the boundary of the constraint (i.e., equal to zero) were removed from consideration. We subsequently fit a model without specifying any constraints and removed any non-significant loadings to arrive at our final solution.

Next, we conducted likelihood ratio tests and compared information criteria for the three pruned models (Models A.1-A.3) to select a final model. After selecting the best fitting model, we assessed whether cross-loadings could be ignored by fitting a new four-factor IFA model (Model B diagrammed in Panel B of Figure 2) in which we assigned items to the factor

VALIDATION OF SUBSCALES FROM THE CREDI

corresponding to the most positive standardized loading from the final model best fitting model in Model A.1-A.3.

We evaluated the dimensionality of the data by assessing model fit of the best fitting of the IFA model with four factors (i.e., Model A or Model B, which include separate factors for motor, cognitive, language, and socio-emotional skills) compared to IFA models with fewer dimensions (Model C and Model D). In fitting Model A and Model B, we consistently found strong, positive residual correlations between the factors representing cognitive and socio-emotional skills (approximately $r = .80$). Consequently, we tested a model that combined these factors (Model C) compared to a four-factor solution. Next, we tested whether a four-factor solution fits better than a unidimensional model (Model D) specified with a single factor representing one general ECD construct. If the data support a model specified with four factors, then we would expect that the best fitting four-factor solution (Model A or Model B) would fit better than the three-factor solution (Model C) and the unidimensional model (Model D).

Criterion-related Validity

Following the recommendations of the *Standards*, we assessed the criterion-related validity evidence by evaluating whether the relations of subscores with other measures and known correlates of children's development are consistent with theory. We studied the correlations between CREDI scores with anthropometric data and household stimulation measures because these variables have been shown to predict children's development (Sudfeld et al., 2015; Walker et al., 2011). Additionally, associations between the CREDI subscores and scores from concurrent ECD measures obtained from a subsample of participants were also studied to investigate convergent and discriminant relations. Local collaborators within the data collection sites selected concurrent measures based on children's age and cultural appropriateness and included: (1) the ASQ Social-Emotional (ASQ:SE; Squires, Bricker, &

VALIDATION OF SUBSCALES FROM THE CREDI

Twombly, 2002) collected from 234 Chilean children, (2) the BSID-III, collected from 1,036 Tanzanian children, (3) the INTER-NDA, collected from 921 Zambian children, (4) the MacArthur-Bates CDI collected from 180 Chilean children, and (5) the PRIDI, collected from 598 Brazilian children from 2 to 3 years old. Appendix Table 1 presents a brief description of each instrument. In analyzing convergent and discriminant validity, we calculated partial correlations using polynomial regression to control for the strong confounding effect of age; we also controlled for between-site differences in scores by specifying fixed effects in the regression model.

We collected anthropometric and household stimulation data for 8,925 children in seven countries. HAZ scores (height-for-age or length-for-age z-scores for children less than 24 months) were calculated using the WHO child growth standards (Onis, 2006). Child stimulation was measured following UNICEF guidelines (2014), totaling the number of adult-child activities as reported by the main caregiver, including reading, telling stories, singing songs, taking outside the child, playing, and naming, counting, or drawing objects.

Reliability

We tested two forms of reliability in this study. First, we examined the stability of scores (i.e., test-retest reliability) using data collected from 575 caregivers in Guatemala, Jordan, and Lebanon, who completed the CREDI twice over a 7- to 10-day administration period. We calculated interclass correlation coefficients (ICCs) to measure the stability of scaled scores. We fit a one-way random effects ANOVA to estimate the intraclass coefficient 1, or ICC(1). We chose the one-way random effects model over two-way alternatives because the one-way model measures the absolute agreement between scores across the two points in time by estimating the correlation between time points (McGraw & Wong, 1996).

VALIDATION OF SUBSCALES FROM THE CREDI

Second, we analyzed internal consistency reliability by studying pairwise tetrachoric correlations and by calculating Cronbach's alpha values. We relied on Cronbach's alpha statistics rather than coefficient omega statistics because the latter assumes unidimensionality (see Bandalos, 2018, p. 395) which is not amenable to the multidimensional measurement approach we adopted in this study. Specifically, for each domain, we evaluated separate Cronbach's alpha values for children aged 0-11 months, 12-23 months, and 24-35 months. We note that reporting a single alpha value across all ages is not appropriate because item responses are so highly correlated with age. Consequently, a single value would suggest greater precision of the instrument than warranted when an important goal of the instrument is to discriminate among children of the same age.

Fairness

We investigated measurement non-invariance by studying whether there is evidence of test-level bias in scores across (a) high, (b) middle-high, and (c) low country income groups, as indicated by differential test functioning. We used only data from the fourth and last round of pilot testing, when the administration of the CREDI most resembled its current form. Thus, the total sample size for assessing invariance was $N = 6,545$ caregivers.

In the present study, we conducted pairwise tests comparing differential test functioning across each income group, separately by domain (i.e., motor, cognition, etc.). We used the simulation procedure advanced by Chalmers' et al. (2016) to form a sampling distribution for the unsigned differential test functioning (uDTF) statistic in order to conduct significance testing. The uDTF is interpreted as the average absolute difference in predicted total scores given children's position on the scale for a particular domain (e.g., motor, cognition, etc.), where we used the maximum-a-posteriori factor scores to approximate a child's position on the scale. As an absolute difference, the uDTF is a conservative statistic and represents an upper bound in

VALIDATION OF SUBSCALES FROM THE CREDI

measuring differential test functioning. If the estimated uDTF statistic is statistically significant, such evidence suggests that abilities differentially predict item response patterns and may indicate possible test-level bias. Relying on Stark et al.'s (2004) proposed Cohen's d , we analyzed the substantive size of the uDTF to ascertain whether evidence of bias is practically important,

$$d = \frac{\widehat{\text{uDTF}}}{s_X}, \quad (1)$$

where $\widehat{\text{uDTF}}$ is the estimate for the unsigned differential test functioning statistic and s_X is the standard deviation of observed total scores. The supplemental material contains technical details on our testing procedure for evaluating differential test functioning.

Results

Construct Validity

Of the 110 initial CREDI items, 108 items exhibited positive loadings on at least one domain across the three initial loading specifications discussed in the Methods section and outlined in Panel A of Figure 1 (Models A.1-A.3). The two items that did not exhibit a positive loading under any of the considered specifications included: (1) "Does the child often cry for no reason (e.g., when he/she is not hungry or tired)?" (reverse coded), and (2) "Does the child cry or whine when he/she is made to wait for something he/she wants (e.g., toy or food)?" (reverse coded). These items were subsequently removed when fitting pruned versions of Model A.1-A.3. Likelihood ratio tests suggested that the more stringent eight-vote threshold (Model A.1) and six-vote threshold (Model A.2) for specifying cross-loadings resulted in a decrement in model fit relative to the less strict four-vote model (Model A.1 vs. Model A.3: $\chi^2(18) = 3,999.64, p < .001$; Model A.2 vs. Model A.3: $\chi^2(7) = 1,425.28, p < .001$).

VALIDATION OF SUBSCALES FROM THE CREDI

Relative to Model A.3, likelihood ratio tests also identified a significant decrement in model fit if cross-loadings were not specified (Model B vs. Model A.3: $\chi^2(26) = 2,106.30, p < .001$), if a three-factor solution was employed by combining the factors representing cognitive and socio-emotional skills (Model C vs. Model A.3: $\chi^2(81) = 16,388.35, p < .001$), or if a unidimensional model (Model D) was utilized (Model D vs. Model A.3: $\chi^2(100) = 15,106.78, p < .001$). Combined with the fact that Model A.3 also minimized both the AIC and the BIC across all fitted models (see Table 2), the data therefore suggest a four-factor model with cross-loadings maximizes model-data consistency. Thus, we selected Model A.3 as the final model for the CREDI long form. Appendix Table 2 reports standardized factor loading estimates for this final model (Model A.3); unstandardized factor loadings and threshold estimates are provided in the supplemental material (see Supplemental Table 1).

The correlations among the residuals of the motor, cognitive, language, and socio-emotional factors from the final model (Model A.3) suggest that the factors themselves displayed adequate discrimination to justify a four-factor solution. Except for the residual between the factors representing cognitive and socio-emotional skill ($r = .81, p < .001$), these values ranged from $r = .49 (p < .001)$ between language and socio-emotional skills to $r = .67 (p < .001)$ between motor and cognitive skills. Children's scores on one factor most often explained less than half the variance in scores on a separate factor, holding age constant and controlling for mean differences in scores between sites.

Criterion-related Validity

For each of the four ECD domains, we found evidence of criterion-related validity. The partial correlation between HAZ and CREDI subscores ranged from $r = .16 (p < .001)$ to $r = .20 (p < .001)$. These partial correlations were similar to or larger than those observed between HAZ and scores from concurrent ECD instruments in this sample. Similarly, CREDI subscale scores

VALIDATION OF SUBSCALES FROM THE CREDI

were positively associated with child stimulation, with partial correlations ranging from $r = .21$ ($p < .001$) to $r = .25$ ($p < .001$). As observed with HAZ, CREDI subscale scores were more positively correlated with stimulation than scores from the previously established ECD measures (although we recognize that this may be in part a function of same-reporter bias). Finally, CREDI subscale scores were positively associated with the PRIDI scores (a composite measure of overall development) in Brazil, and partial correlations ranged from $r = .37$ ($p < .001$) to $r = .47$ ($p < .001$).

We also found that convergent and discriminant relations between CREDI motor and languages subscales with subscores from alternative ECD measures generally matched that expected by theory. Partial correlations with CREDI motor scores were strongest for gross motor scores from the BSID-III ($r = .26$, $p < .001$) and from the INTER-NDA ($r = .50$, $p < .001$), but were also positively associated with fine motor skills (BSID-III: $r = .22$, $p < .001$; INTER-NDA: $r = .18$, $p < .001$). Partial correlations with language, cognitive, and socio-emotional scores from these alternative measures ranged from $r = .12$ ($p < .001$) to $r = .24$ ($p < .001$) for the BSID-III scores and from $r = .16$ ($p < .001$) to $r = .34$ ($p < .001$) for the INTER-NDA scores.

CREDI language scores displayed similar convergent and discriminant validity evidence. Language scores exhibited strong, positive partial correlations with the MacArthur-Bates CDI ($r = .60$, $p < .001$), with expressive language scores from the BSID-III ($r = .26$, $p < .001$), and with expressive language scores from the INTER-NDA ($r = .42$, $p < .001$). In contrast, scores from other ECD domains were less positively correlated with CREDI language scores and were found to range from $r = .12$ ($p < .001$) with BSID-III's socio-emotional scores to $r = .40$ ($p < .001$) with INTER-NDA's gross motor skills. CREDI language scores also exhibited positive partial correlations with receptive language measures (BSID-III: $r = .14$, $p < .001$; INTER-NDA: $r = .20$, $p < .001$). In summary, the CREDI language subscale displayed evidence of both convergent

VALIDATION OF SUBSCALES FROM THE CREDI

validity and discriminant validity, especially as it relates to expressive language subscales from alternative instruments.

Moreover, positive partial correlations with concurrent cognitive and socio-emotional subscales provided evidence for convergent validity; however, there was less evidence for discriminant validity. As expected, CREDI cognitive and socio-emotional scores exhibited positive partial correlations with equivalent subscales from the BSID-III (cognitive: $r = .17, p < .001$; socio-emotional: $r = .13, p < .001$), the INTER-NDA (cognitive: $r = .25, p < .001$), and the ASQ:SE (socio-emotional: $r = .31, p < .001$). However, CREDI cognitive scores exhibited even more positive partial correlations with concurrent expressive language scores (BSID-III: $r = .25, p < .001$; INTER-NDA: $r = .36, p < .001$). Likewise, for children of the same age, CREDI socio-emotional scores were more positively correlated with language scores from the BSID-III (receptive: $r = .15, p < .001$; expressive: $r = .24, p < .001$), while ASQ:SE scores were most positively correlated with CREDI cognitive scores ($r = .33, p < .001$). In summary, although we found evidence that CREDI cognitive and socio-emotional scores were positively correlated with measures from alternative instruments, we did not find that these scores were most correlated with concurrent cognitive and socio-emotional measures. We provide a possible explanation for this in our Discussion. Appendix Table 3 contains a table of partial correlations all measures.

Reliability

Moderate-to-strong correlations between scores provided evidence of test-retest reliability. The ICC(1) model ranged between .70 to .81 across the domains (Motor: ICC(1) = .81, 95% CI [.76, .85]; Cognitive: ICC(1) = .79, 95% CI [.74, .83]; Language: ICC(1) = .70, 95% CI [.63, .76]; Socio-emotional: ICC(1) = .78, 95% CI [.73, .83]). These ICC(1) values indicate moderate levels of stability over time for language scores and good levels of stability for the other domains (Koo & Li, 2016).

VALIDATION OF SUBSCALES FROM THE CREDI

Strong pairwise tetrachoric correlations and acceptable Cronbach's alpha values provide evidence of internal-consistency reliability within each of the four domains. Tetrachoric correlations were all positive and averaged around .80 within each domain (Motor: $M = .78$, $SD = .12$; Cognitive: $M = .80$, $SD = .12$; Language: $M = .78$, $SD = .12$; Socio-emotional: $M = .81$, $SD = .14$).

We also found that the Cronbach's alpha values ranged between .64 and .94 across the four domains and three age-groups (see Table 4). Internal consistency was slightly lower for the socio-emotional subscale relative to the other ECD domains, which is perhaps not surprising given the diversity of socio-emotional skills included (e.g., emotion knowledge, self-regulation, social competence, etc.).

Fairness

For the motor, language and socioemotional domains, we found evidence of statistically significant, but substantively small levels of differential test functioning when comparing scores across country income groups (Table 5). uDTF effect sizes in the motor, language, and socio-emotional domains ranged from $d = 0.04$ (Language, high- vs. middle-high income groups: Est. = 0.33, $p = .206$) to $d = 0.09$ (Socio-emotional, high- vs. middle-high income groups: Est. = 0.39, $p < .001$). These effect sizes are universally accepted as small in substantive size (c.f., Cohen, 1988). The small levels of observed differential test functioning indicate that the statistically significant findings of differential test functioning are artifacts of the large sample size ($N = 6,545$), but likely do not suggest that test-level bias threatens the validity of inferences regarding children's development when comparing across country income groups.

Notably, cognitive scores demonstrated the strongest uDTF effect sizes, with the uDTF strongest between middle-high vs. low-income countries and taking on a value of $d = 0.18$ (Est. = 1.06, $p < .001$). Although such a value arguably classifies the differential test functioning as

VALIDATION OF SUBSCALES FROM THE CREDI

moderate rather than small, we note that the uDTF statistic is a conservative statistic and likely overestimates the amount of differential functioning that would change conclusions when comparing scores across income groups.

Discussion

In this paper we have used a large ($N = 14,113$), multi-country and multicultural sample to assess the validity evidence for the motor, language, cognitive, and socio-emotional subscales for the long form of the CREDI. We found sufficient evidence to justify a four-factor solution, as well as acceptable internal-consistency reliability and test-retest reliability. We also found evidence of concurrent validity, although the adjusted CREDI cognitive and socio-emotional scores were more strongly correlated with concurrent scores representing non-equivalent domains than concurrent scores representing the same domain. Regarding the cognitive domain, CREDI scores were more strongly correlated with concurrent expressive language scores than they were with concurrent cognition scores. Although it may seem that the factor representing cognition is more accurately a measure for a language construct, we believe this explanation is unlikely. If adjusted CREDI cognitive scores represented a language construct, then we would expect an unusually strong residual correlation between the cognitive and language domains from Model A.3. Although there was a moderate-to-strong residual correlation between the cognitive and language factors ($r = .62, p < .001$), this association was weaker than the corresponding residual correlation between the cognitive and motor factors ($r = .67, p < .001$).

An alternative explanation is perhaps that concurrent measures of cognitive and expressive language development in young children (e.g., the BSID-III, INTER-NDA) have not allowed for items to load on multiple domains. As a result, these measures may be confounding cognitive and language development in ways that inflate their expressive language subscales relative to the CREDI cognitive subscale. Conceptually, indicators of expressive language (in the

VALIDATION OF SUBSCALES FROM THE CREDI

CREDI and in the concurrent measures) often tap into children's latent cognitive abilities through asking children to describe complex constructs or explain (i.e., make sense of) situations. In fact, of the 44 items that loaded on the factors representing cognition or language, greater than one third (15 items) loaded on both factors simultaneously. Moving forward, additional work is needed to better understand the relations between these complex constructs and to identify more precise ways to operationalize them in distinct ways.

The weak discriminant validity evidence for CREDI's socio-emotional subscale is unsurprising. Socio-emotional development is an extremely broad construct encompassing a highly diverse set of skills ranging from getting along with others (social competence) to inhibiting impulsive behavior (self-regulation) to identifying and responding to emotions (emotion knowledge; Jones et al., 2016). Accordingly, it is no surprise that the socio-emotional measures from the BSID-III, ASQ:SE, and CREDI all focus on different facets of socio-emotional development, complicating comparisons of these scales. The BSID-III and ASQ:SE tend to emphasize adaptive behaviors (e.g., sleep, behavior during mealtimes), whereas the CREDI does not emphasize these behaviors. Further research extricating and incorporating the distinct constructs that comprise socio-emotional development will be needed.

The evidence suggesting unsubstantial levels of differential test functioning in the motor, language, and socioemotional domains is encouraging as it is suggestive that item-level measurement invariance is likely not acting systemically in one direction so as to bias conclusions when comparing mean differences in scores across country income groups. However, researchers using the CREDI scores proceed cautiously in comparing scores across populations for several reasons. The present study only evaluated evidence of measurement invariance across country income groups. Therefore, we cannot establish whether measurement non-invariance would invalidate conclusions if comparing populations defined by some other set

VALIDATION OF SUBSCALES FROM THE CREDI

criteria; future research should examine whether there is evidence of differential test functioning across alternatively defined populations. Meanwhile, we encourage users of the CREDI to acknowledge that conclusions may be dependent on the assumption of especially important given the finding that the size of the uDTF for cognitive scores arguably does designate it as substantively small. It is difficult to project the implications of this finding in practice because we remain unaware of guidance in the literature for when the substantive size of the uDTF designates it as concerning and jeopardizes the validity of conclusions. Future methodological research should focus on providing such guidance.

Although our findings suggest favorable evidence for construct validity of the subscales, we have developed using CREDI's long form, potential users should consider several limitations. Our culturally and linguistically diverse sample was obtained by convenience and not necessarily representative of any stringently defined global population. Thus, next steps include defining a target population, then obtaining representative samples from this target.

More evidence is also needed to firmly establish criterion-related validity. Longitudinal data would provide predictive validity evidence using distal outcomes, including school readiness, academic performance, and later mental health and emotional wellbeing. Given that we found differential test functioning for the motor domain, researchers should be cautious when comparing mean differences in motor scores across countries. Future work should focus on identifying the sources of this differential functioning and investigate item-level measurement non-invariance. Lastly, the CREDI subscales measure aspects of ECD that are shared across cultures, but they are not designed to meaningfully capture important phenomena measured by culturally specific instruments. Moving forward, we strongly recommend that researchers pair the CREDI with direct assessments that can target culturally-specific processes while mitigating bias (e.g., social desirability) associated with caregiver report.

VALIDATION OF SUBSCALES FROM THE CREDI

We also found that items frequently measured multiple domains simultaneously and that specifying cross-loadings resulted in improved model-data consistency. To assist in scoring in the presence of cross-loadings, we provide users with a web-based scoring application available at [*hyperlink redacted*]. Cross-loadings are consistent with developmental theory in that children's observable behavior often requires the recruitment of skills from multiple domains, especially early in life. Yet, to our knowledge, existing ECD instruments ignore item-level multidimensionality, as items are typically assigned to a single developmental domain during the calculation of subscores. Our findings indicate that such practices may result in a misspecified measurement model. Future research should examine whether conclusions about children's development are sensitive to such misspecification, as would be hypothesized by previous simulation studies (c.f., Curran, 1994).

In conclusion, we have shown that scores from the CREDI long form demonstrate evidence of construct and criterion-related validity and are sufficiently precise for population measurement purposes. The CREDI long form is designed to be globally relevant and applicable across cultures. As a self-report measure, the CREDI long form is efficient to implement and can be used in public policies to monitor child development and to assess interventions, with the goal of improving outcomes of children around the world. Towards this end, recent research suggests that the simple act of interviewing caregivers in measuring their children's development may itself help caregivers become more aware of and attentive to their children's milestone attainment and behaviors (Altafim et al., 2020).

References

- Altafim, E. R. P., McCoy, D. C., Brentani, A., de Ulhôa Escobar, A. M., Grisi, S. J., & Fink, G. (2020). Measuring early childhood development in Brazil: validation of the Caregiver Reported Early Development Instruments (CREDI). *Jornal de Pediatria (Versão em Português)*, 96(1), 66-75.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3), 397-438.
- Asparouhov, T., & Muthén, B. (2012). Comparison of computational methods for high dimensional item factor analysis. Mplus Technical Report.
- Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. New York, NY: The Guilford Press.
- Bayley, N. (2006). Bayley scales of infant and toddler development: Bayley-III (Vol. 7). San Antonio, Tex, USA: Harcourt Assessment, Psych. Corporation.
- Black, M. M., Walker, S. P., Fernald, L. C., Andersen, C. T., DiGirolamo, A. M., Lu, C., ... & Devercelli, A. E. (2017). Early childhood development coming of age: science through the life course. *The Lancet*, 389(10064), 77-90.
- Bowman, L. C., Pierce, L. J., Nelson, C. A., & Werker, J. F. (2018). Neural foundations of cognition and language. In R. Gibb & B. B. T.-T. N. of B. and B. D. Kolb (Eds.), *The*

VALIDATION OF SUBSCALES FROM THE CREDI

neurobiology of brain and behavioral development (pp. 257–290). London: Elsevier.
<https://doi.org/10.1016/B978-0-12-804036-2.00010-8>

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, 35(3), 307-335.

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: Improved differential test functioning statistics that account for sampling variability. *Educational and Psychological Measurement*, 76(1), 114–140. <https://doi.org/10.1177/0013164415584576>

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), 37-46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J.: L. Erlbaum Associates.

Curran, P. J. (1994). The robustness of confirmatory factor analysis to model misspecification and violations of normality (Doctoral dissertation, Arizona State University).

Fenson, L., Bates, E., Dale, P. S., Marchman, V. A., Reznick, J. S., & Thal, D. J. (2007). MacArthur-Bates communicative development inventories.

Fernandes, M., Stein, A., Newton, C. R., Cheikh-Ismail, L., Kihara, M., Wulff, K., ... & Ibanez, D. (2014). The INTERGROWTH-21st Project Neurodevelopment Package: A novel method for the multi-dimensional assessment of neurodevelopment in pre-school age children. *PloS one*, 9(11), e113360.

Frankenburg, W. K., & Dodds, J. B. (1967). The Denver developmental screening test. *The Journal of pediatrics*, 71(2), 181-191.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (Vol. 3). Boca Raton, FL: CRC press.

VALIDATION OF SUBSCALES FROM THE CREDI

- Ghandour, R. M., Moore, K. A., Murphy, K., Bethell, C., Jones, J. R., Harwood, R., ... Lu, M. (2019). School readiness among U.S. children: Development of a pilot measure. *Child Indicators Research, 12*(4), 1389–1411. <https://doi.org/10.1007/s12187-018-9586-8>
- 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*(9555), 60-70.
- Halpin, P., Wolf, S., Yoshikawa, H., Rojas, N., Kabay, S., Pisani, L., & Dowd, A. (2019). Measuring Early Learning and Development Across Cultures: Invariance of the IDELA Across Five Countries. *Developmental Psychology, 55*(1), 23-37.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.
- Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science, 312*(5782), 1900-1902.
- Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(4), 1–12. <https://doi.org/10.1080/10705511.2016.1154793>
- Jones, S. M., Zaslow, M., Darling-Churchill, K. E., & Halle, T. G. (2016). Assessing early childhood social and emotional development: Key conceptual and measurement issues. *Journal of Applied Developmental Psychology, 45*, 42-48. [doi:https://doi.org/10.1016/j.appdev.2016.02.008](https://doi.org/10.1016/j.appdev.2016.02.008)
- Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research, 23*(1), 69-86.

VALIDATION OF SUBSCALES FROM THE CREDI

- Kerstjens, J. M., Bos, A. F., ten Vergert, E. M., de Meer, G., Butcher, P. R., & Reijneveld, S. A. (2009). Support for the global feasibility of the Ages and Stages Questionnaire as developmental screener. *Early Human Development*, 85(7), 443-447.
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Liu, J., & Kang, H.-A. (2019). Q-matrix learning via latent variable selection and identifiability. In M. Von Davier & Y.-S. Lee (Eds.), *Handbook of Diagnostic Classification Models* (pp. 247–263). Cham: Springer. <https://doi.org/10.1007/978-3-030-05584-4>
- Lu, C., Black, M. M., & Richter, L. M. (2016). Risk of poor development in young children in low-income and middle-income countries: an estimation and analysis at the global, regional, and country level. *The Lancet Global Health*, 4(12), e916-e922.
- Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., ... & Sears, M. R. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences*, 108(7), 2693-2698.
- Muthén, L.K. and Muthén, B.O. (1998-2017). *Mplus User's Guide*. Eighth Edition. Los Angeles, CA: Muthén & Muthén
- Nores, M., & Barnett, W. S. (2010). Benefits of early childhood interventions across the world:(Under) Investing in the very young. *Economics of Education Review*, 29(2), 271-282.
- Onis M. WHO child growth standards based on length/height, weight and age. *Acta Paediatrica* 2006; 95(S450):76–85.

VALIDATION OF SUBSCALES FROM THE CREDI

- Peet, E. D., McCoy, D. C., Danaei, G., Ezzati, M., Fawzi, W., Jarvelin, M. R., ... & Fink, G. (2015). Early childhood development and schooling attainment: Longitudinal evidence from British, Finnish and Philippine birth cohorts. *PloS ONE*, 10(9), e0137219.
- Peña, E. D. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child Development*, 78(4), 1255-1264.
- Reckase, M. (2009). *Multidimensional Item Response Theory (Statistics for Social and Behavioral Sciences)*. New York, NY: Springer New York.
- Richter, L. M., Daelmans, B., Lombardi, J., Heymann, J., Boo, F. L., Behrman, J. R., ... & Bhutta, Z. A. (2017). Investing in the foundation of sustainable development: pathways to scale up for early childhood development. *The Lancet*, 389(10064), 103-118.
- Richter, L., Black, M., Britto, P., Daelmans, B., Desmond, C., Devercelli, A., ... & Lu, C. (2019). Early childhood development: an imperative for action and measurement at scale. *BMJ global health*, 4(Suppl 4), e001302.
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology*, 89(3), 497–508. <https://doi.org/10.1037/0021-9010.89.3.497>
- Squires, J., & Bricker, D. (2009). *Ages & Stages Questionnaires®, Third Edition (ASQ®-3): A Parent-Completed Child Monitoring System*. Baltimore: Paul H. Brookes Publishing Co., Inc.
- Squires, J., Bricker, D., & Twombly, E. (2002). *The ASQ:SE user's guide: For the Ages & Stages Questionnaires: Social-emotional*. Paul H Brookes Publishing.

VALIDATION OF SUBSCALES FROM THE CREDI

Sudfeld, C. R., McCoy, D. C., Danaei, G., Fink, G., Ezzati, M., Andrews, K. G., & Fawzi, W. W.

(2015). Linear growth and child development in low-and middle-income countries: a meta-analysis. *Pediatrics*, *135*(5), e1266-e1275.

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

United Nations, 2015. Sustainable Development Goals. New York, NY: United Nations.

Verdisco, A., Cueto, S., Thompson, J., Engle, P., Neuschmidt, O., Meyer, S., ... & Miranda, A.

(2014). Urgency and Possibility: Results of PRIDI A First Initiative to Create Regionally Comparative Data on Child Development in Four Latin American Countries. Technical Annex. Inter-American Development Bank, Washington DC.

Walker, S. P., Wachs, T. D., Grantham-McGregor, S., Black, M. M., Nelson, C. A., Huffman, S.

L., ... & Gardner, J. M. M. (2011). Inequality in early childhood: risk and protective factors for early child development. *The Lancet*, *378*(9799), 1325-1338.

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.

Author Contributions

MW conceptualized the validity study, designed the methodological approach, conducted the psychometric analysis, wrote the initial draft, and approved the final manuscript. DM assisted in the design of the study, conducted preliminary statistical analysis, developed the data collection instruments, reviewed and revised drafts, and approved the final manuscript. The CREDI Field Team assisted in the development of data collection instruments, conducted all data collection, reviewed and revised drafts, and approved the final manuscript. GF assisted in the design of the study, conducted preliminary statistical analysis, developed portions of the data collection instruments, led the study sampling, reviewed and revised drafts, and approved the final manuscript.

Acknowledgements

We would like to acknowledge the intellectual contributions of the CREDI Advisory Panel members and our data collection partners. We thank the thousands of children and caregivers who participated in this research. Finally, we are also grateful to Katherine Masyn for her insights and suggestions, and for generously sharing her computational resources so that we could make progress in a timely manner.

Funding

The authors would like to acknowledge funding and support provided by the Saving Brains Program from Grand Challenges Canada (Grant Number 0073-03).

VALIDATION OF SUBSCALES FROM THE CREDI

Table 1. CREDI sample description

Country	Full sample size (total number of children assessed)	Analytic sample size (children under 36 months)	Country estimated stunting prevalence	Country estimated average daily income per capita in USD
Bangladesh	280	280	39%	\$ 8.6
Brazil	2,359	2,212	7%	\$ 39.8
Cambodia	493	410	40%	\$ 9.0
Chile	244	244	2%	\$ 60.8
Colombia	378	314	13%	\$ 35.6
Ghana	3,000	1,709	19%	\$ 10.8
Guatemala	205	197	47%	\$ 19.9
India	200	200	38%	\$ 15.7
Jordan	317	278	8%	\$ 28.1
Laos	46	43	44%	\$ 14.6
Lebanon	426	384	17%	\$ 35.9
Nepal	363	363	37%	\$ 6.3
Pakistan	250	241	45%	\$ 12.9
Philippines	720	719	30%	\$ 19.0
Tanzania	3,715	3,610	34%	\$ 6.9
USA	1,021	899	2%	\$ 144.4
Zambia	2,012	2,010	40%	\$ 9.9
<i>Total/average</i>	<i>16,029</i>	<i>14,113</i>	<i>27%</i>	<i>\$ 28.1</i>
<i>Min</i>	<i>46</i>	<i>43</i>	<i>2%</i>	<i>\$ 6.3</i>
<i>Max</i>	<i>3,715</i>	<i>3,610</i>	<i>47%</i>	<i>\$ 144.4</i>

Note: Income per person and day computed by dividing purchasing-power-parity adjusted per capita income in each country by 365 days. Stunting data refers to children under age 5 and was retrieved from <http://data.unicef.org/topic/nutrition/malnutrition/>.

VALIDATION OF SUBSCALES FROM THE CREDI

Table 2. Model fit across fitted IFA models.

Model	Factors	Parameters	LL	AIC	BIC
A.1	4	321	-232549.06	465740.11	468160.25
A.2	4	332	-231261.88	463187.76	465690.83
A.3	4	339	-230549.24	461776.48	464332.32
B	4	313	-231602.39	463830.78	466190.60
C	3	309	-232068.34	464754.67	467084.33
D	1	239	-238102.63	476683.26	478485.16

Notes: $N = 14,113$. All models fit to the same $J = 108$ items.

VALIDATION OF SUBSCALES FROM THE CREDI

Table 3. Observed loading patterns from the best-fitting model (Model A.3).

	Motor	Cognitive	Language	Soc.-emo.	# Items
1	✓	-	-	-	35
2	✓	✓	-	-	5
3	-	✓	-	-	10
4	-	✓	✓	-	13
5	-	✓	-	✓	4
6	-	-	✓	-	22
7	-	-	✓	✓	4
8	-	-	-	✓	15
Total	40	32	39	23	

Note: ✓ indicates positive and significant loading estimate.

VALIDATION OF SUBSCALES FROM THE CREDI

Table 4. Cronbach's α values observed across age groups and ECD domain.

	Motor	Cognitive	Language	Soc.-emot.
0-11 months	0.94	0.88	0.86	0.70
12-23 months	0.85	0.85	0.92	0.70
24-35 months	0.74	0.80	0.90	0.64

VALIDATION OF SUBSCALES FROM THE CREDI

Table 5. uDTF by country income group comparison across ECD domains.

	<u>High vs. low</u> <u>income</u>			<u>High- vs. middle-</u> <u>high income</u>			<u>Middle-high vs.</u> <u>low income</u>		
	Est.	<i>d</i>	<i>p</i>	Est.	<i>d</i>	<i>p</i>	Est.	<i>d</i>	<i>p</i>
Motor	0.42	0.06	<.001	0.48	0.07	<.001	0.36	0.05	<.001
Cognition	0.62	0.11	<.001	0.48	0.08	<.001	1.06	0.18	<.001
Language	0.44	0.05	<.001	0.33	0.04	.206	0.51	0.06	.736
Socio-emotional	0.28	0.07	<.001	0.36	0.09	<.001	0.26	0.07	<.001

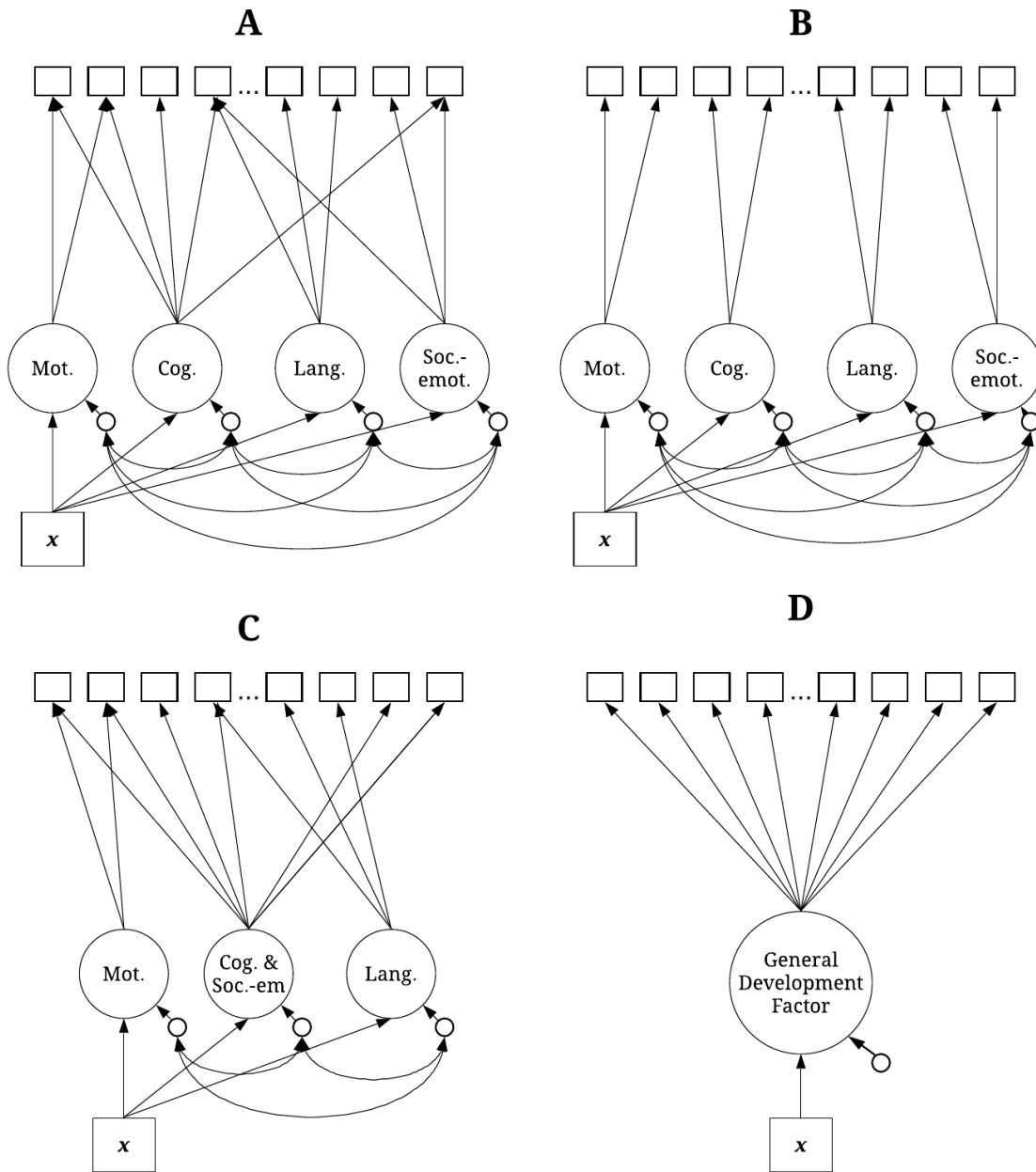


Figure 1. (A) Four-factor IFA model that specifies cross-loadings. (B) Four-factor IFA model without cross-loadings. (C) Three-factor IFA model with the factor representing cognitive and socio-emotional skills combined and with cross-loadings specified (D) Single-factor IFA model. The predictor of each latent variable, x , represents three covariates: a linear, a quadratic, and a cubic term for age, as well as site fixed effects.

VALIDATION OF SUBSCALES FROM THE CREDI

Appendix Table 1

Description of concurrent ECD measures.

Instrument	Description
1 Ages and Stages Questionnaire: Social-Emotional (ASQ:SE)	The ASQ:SE is a caregiver reported screener intended to identify children, aged 1-72 months, at risk of exhibiting behavioral difficulties.
2 Bayley Scales of Infant and Toddler Development III (BSID)	The BSID-III is a series of developmental assessments for children aged 1-42 months. The BSID-III provides subscales for fine-motor skills, gross-motor skills, cognitive development, expressive-language development, and receptive-language developmental using caregiver reports. The BSID-III also includes both caregiver reported and observer-reported measures for socio-emotional development and adaptive behaviors.
3 INTERGROWTH Neurodevelopmental Assessment (INTER-NDA)	The INTER-NDA is a series of maternally-reported and observer-reported assessments for children aged 0-3 years old. The INTER-NDA includes subscales for fine-motor development, gross-motor development, cognitive development, receptive-language development, and expressive-language development. Also included is an overall-language score which is a simple composite using the expressive- and receptive-language items.
4 MacArthur Bates Communicative Development Inventories (MacArthur Bates CDIs)	The MacArthur Bates CDIs measure language development and is appropriate for children aged 8-30 months. The assessment is available in both English and Spanish, and is administered to the child by a trained examiner.
5 Regional Project on Child Development Indicators (PRIDI)	The PRIDI is a single, composite measure that covers motor, cognitive, language, and socio-emotional development. The PRIDI is appropriate for children aged 24-59 months old.

VALIDATION OF SUBSCALES FROM THE CREDI

Appendix Table 2

Subject matter expert votes and standardized loading estimates.

		Motor		Cognitive		Language		Socio-emotional	
		Votes	Std. Est.	Votes	Std. Est.	Votes	Std. Est.	Votes	Std. Est.
LF1	When lying on his/her back, does the child move his/her arms and legs?	16	.90***	-	-	-	-	-	-
LF2	Does the child bring his/her hand to his/her mouth?	16	.90***	-	-	-	-	-	-
LF3	Does the child laugh?	-	-	-	-	-	-	13	.78***
LF4	Does the child smile when others smile at him/her?	-	-	-	-	-	-	15	.82***
LF5	Does the child sometimes suck his/her thumb or fingers?	4	.84***	-	-	-	-	11	-
LF6	Does the child grasp onto a small object (e.g., your finger, a spoon) when put in his/her hand?	16	.84***	-	-	-	-	-	-
LF7	Can the child bring his/her hands together?	16	.75***	-	-	-	-	-	-
LF8	Does the child recognize you or other family members (e.g., smile when they enter a room or move toward them)?	-	-	6	-	-	-	13	.91***
LF9	Does the child hold his/her hands in fists all the time?	15	.44***	-	-	-	-	-	-
LF10	Does the child show interest in new objects that are put in front of him/her by reaching out for them?	6	-	14	.93***	-	-	-	-
LF11	Can the child roll from his/her back to stomach, or stomach to back, on his/her own?	16	.90***	-	-	-	-	-	-
LF12	Does the child show interest in new objects by trying to put them in his/her mouth?	7	-	11	.91***	-	-	-	-
LF13	Does the child often show affection toward others (e.g., hugging parents, brothers, or sisters)?	-	-	-	-	-	-	14	.64***
LF14	Can the child pick up a small object (e.g., a small toy or small stone) using just one hand?	16	.92***	-	-	-	-	-	-

VALIDATION OF SUBSCALES FROM THE CREDI

LF15	Does the child look for an object of interest when it is removed from sight or hidden from him/her (e.g., put under a cover, behind another object)?	-	-	15	.91***	-	-	-	-
LF16	When lying on his/her back, does the child grab his/her feet?	16	.88***	-	-	-	-	-	-
LF17	Can the child make simple sounds like "ba," "da," or "do?"	-	-	-	-	15	.97***	-	-
LF18	When lying on his/her stomach, can the child hold his/her head and chest off the ground using only his/her hands and arms for support?	16	.88***	-	-	-	-	-	-
LF19	Does the child play by tapping an object on the ground or a table?	5	.52***	11	.44***	-	-	-	-
LF20	Can the child hold him/herself in a sitting position without help or support for longer than a few seconds?	16	.92***	-	-	-	-	-	-
LF21	Does the child intentionally move or change his/her position to get objects that are out of reach?	5	.18***	14	.79***	-	-	-	-
LF22	Does the child look at an object when someone says "look!" and points to it?	-	-	5	.44***	15	.46***	-	-
LF23	Does the child recognize his/her -me or nick-me? That is, does he/she respond differently to his/her -me than to other sounds or words?	-	-	5	.56***	15	-	4	.35***
LF24	When you talk to the child, does he/she respond by making a sound (e.g., "ba," "da," or "do") or by saying a word?	-	-	-	-	15	.95***	4	-
LF25	Can the child crawl, roll, or scoot forward on his/her own?	16	.86***	-	-	-	-	-	-
LF26	Can the child pick up and eat small pieces of food with his/her fingers?	16	.89***	-	-	-	-	-	-
LF27	Can the child transfer a small object (e.g., a small toy or small stone) from one hand to the other?	15	.88***	-	-	-	-	-	-
LF28	Does the child clap his/her hands together?	16	.91***	4	-	-	-	-	-
LF29	Can the child maintain a standing position while holding on to a person or object (e.g., wall or furniture)?	16	.91***	-	-	-	-	-	-
LF30	Can the child use gestures to indicate what he/she wants (e.g., put arms up to indicate that he/she wants to be held, or point to water)?	-	-	4	.59***	12	.33***	6	-
LF31	Can the child pick up a small object (e.g., a small toy or small stone) with just his/her thumb and a finger?	15	.85***	-	-	-	-	-	-

VALIDATION OF SUBSCALES FROM THE CREDI

LF32	Can the child pick up and drop a small object (e.g., a small toy or small stone) into a bucket or bowl while sitting?	16	.90***	-	-	-	-	-	-
LF33	Can the child throw a small ball or small stone in a forward direction using his/her hand?	16	.90***	-	-	-	-	-	-
LF34	Can the child walk several steps while holding on to a person or object (e.g., wall or furniture)?	16	.93***	-	-	-	-	-	-
LF35	Can the child say one or more words (e.g., -mes like "Mama" or "ba" for "ball")?	-	-	-	-	15	.92***	-	-
LF36	Can the child maintain a standing position on his/her own, without holding on or receiving support?	16	.96***	-	-	-	-	-	-
LF37	Can the child follow simple directions (e.g., "Stand up" or "Come here")?	-	-	6	.48***	14	.45***	-	-
LF38	Does the child watch what other children do and try to copy them?	-	-	10	-	-	-	9	.87***
LF39	Can the child sit or play on his/her own for at least 20 minutes?	-	-	4	.72***	-	-	13	-
LF40	Can the child walk several steps on his/her own, without holding on or receiving support?	16	.98***	-	-	-	-	-	-
LF41	Can the child bend down to the ground and stand up again without falling and without holding onto a person or object?	16	.95***	-	-	-	-	-	-
LF42	Does the child ask you for help using signs or words when he/she cannot do something on his/her own (e.g., to reach an object up high)?	-	-	-	-	10	.64***	9	.20***
LF43	Does the child try to repeat sounds or words said by other people?	-	-	-	-	15	.83***	4	-
LF44	Can the child climb onto an object such as a chair or bench?	16	.93***	-	-	-	-	-	-
LF45	Can the child figure out how to turn a spoon or object if you give it to him/her the wrong way around?	-	-	14	.77***	-	-	-	-
LF46	Does the child stop at least briefly when told "no" or "stop that"?	-	-	5	.32***	5	-	11	.32***
LF47	Can the child kick a ball or other round object forward using his/her foot?	16	.95***	-	-	-	-	-	-
LF48	Can the child point to a person or object when asked (e.g., "Where is mama?" or "Where is the ball?")?	-	-	4	.32***	15	.62***	-	-

VALIDATION OF SUBSCALES FROM THE CREDI

LF49	Can the child drink from a cup (without a lid) on his/her own without spilling?	15	.86***	-	-	-	-	-	-
LF50	Does the child imitate animal or other sounds (e.g., "vroom" for a car, "moo" for a cow)?	-	-	5	.17***	15	.72***	-	-
LF51	Can the child run more than a few steps without falling or bumping into objects?	16	.95***	-	-	-	-	-	-
LF52	Can the child draw a line or shape on paper with a pen or crayon, or in the dirt with a stick?	16	.34***	6	.56***	-	-	-	-
LF53	Can the child answer simple questions (e.g., "Do you want water?") by saying "yes" or "no", rather than nodding?	-	-	-	-	15	.93***	-	-
LF54	Can the child stack three or more small objects (e.g., blocks, cups, bottle caps) on top of each other?	16	.47***	8	.43***	-	-	-	-
LF55	Does the child imitate others' behaviors (e.g., washing hands or dishes)?	-	-	11	.19**	-	-	8	.66***
LF56	Does the child sometimes share things (e.g., food, toys) with others without being told?	-	-	-	-	-	-	15	.70***
LF57	Can the child follow orders or instructions that have more than one part (e.g., "Go get water and go to bed")?	-	-	9	.40***	12	.51***	-	-
LF58	Can the child say five or more separate words (e.g., -mes like "Mama" or objects like "ball")?	-	-	-	-	15	.93***	-	-
LF59	Is the child kind to younger children (e.g., speaks to them nicely and touches them gently)?	-	-	-	-	-	-	13	.79***
LF60	Can the child walk on an uneven surface (e.g., a bumpy or steep road) without falling?	16	.86***	-	-	-	-	-	-
LF61	Does the child listen to someone telling a story with interest?	-	-	6	.60***	7	-	10	-
LF62	Can the child ask for something (e.g., food, water) by -me when he/she wants it?	-	-	-	-	15	.94***	-	-
LF63	Does the child involve others in play (i.e., play interactive games with other children)?	-	-	-	-	-	-	14	.88***
LF64	Can the child correctly -me at least one family member other than mom and dad (e.g., -me of brother, sister, aunt, uncle)?	-	-	-	-	15	.91***	-	-

VALIDATION OF SUBSCALES FROM THE CREDI

LF65	Does the child play by pretending objects are something else (e.g., imagining a bottle is a doll, a stone is a car, or a spoon is an airplane)?	-	-	13	.86***	-	-	-	-
LF66	Does the child show sympathy or look concerned when others are hurt or sad?	-	-	-	-	-	-	14	.82***
LF67	Can the child walk backwards?	16	.92***	-	-	-	-	-	-
LF68	Does the child show curiosity to learn new things (e.g., by asking questions or exploring a new area)?	-	-	12	.87***	-	-	6	-
LF69	Can the child feed him/herself using a spoon or other utensil without spilling?	15	.80***	-	-	-	-	-	-
LF70	Can the child concentrate on one task (e.g., playing with friends, eating meal) for 20 minutes?	-	-	10	.54***	-	-	9	.18***
LF71	Does the child know the -mes of at least two body parts (e.g., arm, eye, or nose)?	-	-	6	-	14	.91***	-	-
LF72	If you show the child an object he/she knows well (e.g., a cup or animal), can he/she consistently -me it?	-	-	5	-	15	.94***	-	-
LF73	Can the child speak using short sentences of two words that go together (e.g., "Mama go" or "Dada eat")?	-	-	-	-	15	.96***	-	-
LF74	Can the child use a tool (e.g., a stick or spoon) to reach objects that are far away?	-	-	13	.83***	-	-	-	-
LF75	Can the child indicate when he/she needs to go to the toilet?	-	-	4	-	-	-	8	.61***
LF76	Can the child say ten or more separate words (e.g., -mes like "Mama" or objects like "ball")?	-	-	-	-	15	.93***	-	-
LF77	Can the child remove an item of clothing (e.g., take off his/her shirt)?	14	.86***	-	-	-	-	-	-
LF78	Can the child tell you when he/she is tired or hungry?	-	-	-	-	8	.56***	11	.28***
LF79	Does the child usually finish an activity he/she enjoys (e.g., a game or book)?	-	-	9	.78***	-	-	10	-
LF80	Can the child easily switch back and forth between activities (e.g., go back to a game after being interrupted)?	-	-	9	-	-	-	11	.68***

VALIDATION OF SUBSCALES FROM THE CREDI

LF81	Can the child sing a short song or repeat parts of a rhyme from memory by him/herself?	-	-	10	.21***	12	.68***	-	-
LF82	Can the child jump with both feet leaving the ground?	16	.91***	-	-	-	-	-	-
LF83	Can the child speak using sentences of three or more words that go together (e.g., "I want water" or "The house is big")?	-	-	-	-	15	.94***	-	-
LF84	Can the child whisper?	-	-	6	.72***	9	.22***	-	-
LF85	Does the child greet neighbors or other people he/she knows without being told (e.g., by saying hello or gesturing hello)?	-	-	-	-	-	-	14	.68***
LF86	Can the child unscrew the lid from a bottle or jar?	16	.63***	7	.25***	-	-	-	-
LF87	Can the child correctly ask questions using any of the words "what," "which," "where," or "who"?	-	-	6	.30***	13	.67***	-	-
LF88	Can the child correctly use any of the words "I," "you," "she," or "he" (e.g., "I go to store," or "He eats rice")?	-	-	-	-	15	.93***	-	-
LF89	Does the child pronounce most of his/her words correctly?	-	-	-	-	15	.86***	-	-
LF90	Can the child count up to five objects (e.g., fingers, people)?	-	-	13	.39***	5	.54***	-	-
LF91	Does the child ask about familiar people other than parents when they are not there (e.g., "Where is the neighbor?")?	-	-	5	-	6	.45***	10	.49***
LF92	If you show the child two objects or people of different size, can he/she tell you which one is the big one and which is the small one?	-	-	13	-	7	.84***	-	-
LF93	Can the child stand on one foot for several seconds without holding on to a person or object (e.g., wall or furniture)?	16	.86***	-	-	-	-	-	-
LF94	Can the child identify at least one color (e.g., red, blue, yellow)?	-	-	12	.37***	8	.55***	-	-
LF95	Does the child regularly use describing words such as "fast," "short," "hot," "fat," or "beautiful" correctly?	-	-	5	-	15	.92***	-	-
LF96	If you point to an object, can the child correctly use the words "on," "in," or "under" to describe where it is (e.g., "The cup is on the table" instead of "The cup is in the table.")	-	-	8	-	12	.93***	-	-
LF97	Can the child explain in words what common objects like a cup or chair are used for?	-	-	9	-	13	.88***	-	-

VALIDATION OF SUBSCALES FROM THE CREDI

LF98	Can the child dress him/herself (e.g., put on his/her pants and shirt without help)?	14	.91***	-	-	-	-	-	-
LF99	Does the child ask "why" questions (e.g., "Why are you tall?")?	-	-	9	-	13	.90***	-	-
LF100	If you ask the child to give you three objects (e.g., stones, beans), does the child give you the correct amount?	-	-	14	.28***	6	.64***	-	-
LF101	Does the child usually put objects or toys back where they belong after using them?	-	-	4	-	-	-	11	.93***
LF102	Does the child frequently act impulsively or without thinking (e.g., running into the street without looking)?	-	-	4	-	-	-	9	.78***
LF103	Does the child sometimes save things like candy or new toys for the future?	-	-	6	-	-	-	8	.91***
LF104	Can the child say what others like or dislike (e.g., "Mama doesn't like fruit," "Papa likes football")?	-	-	5	-	7	.49***	12	.48***
LF105	Can the child fasten and unfasten buttons without help?	16	.90***	4	-	-	-	-	-
LF106	Can the child talk about things that will happen in the future using correct language (e.g., "Tomorrow he will attend school" or "Next week we will go to the market")?	-	-	7	-	14	.94***	-	-
LF107	Can the child talk about things that have happened in the past using correct language (e.g., "Yesterday I played with my friend" or "Last week she went to the market")?	-	-	7	.18***	14	.75***	-	-
LF108	Does the child know the -mes of any letters (e.g., A, B, C)?	-	-	11	-	8	.82***	-	-

VALIDATION OF SUBSCALES FROM THE CREDI

Appendix Table 3

Matrix of partial correlations between ECD domain scores.

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
CREDI	1. Motor	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	2. Cognitive	.67***	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	3. Language	.58***	.62***	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	4. Socio-emotional	.55***	.81***	.49***	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BSID III	5. Gross motor	.26***	.24***	.21***	.23***	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	6. Fine motor	.22***	.22***	.20***	.21***	.57***	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	7. Cognitive	.17***	.17***	.16***	.17***	.47***	.54***	-	-	-	-	-	-	-	-	-	-	-	-	-
	8. Receptive Language	.16***	.16***	.14***	.15***	.46***	.55***	.63***	-	-	-	-	-	-	-	-	-	-	-	-
INTERNDA	9. Expressive Language	.24***	.25***	.26***	.24***	.45***	.47***	.55***	.70***	-	-	-	-	-	-	-	-	-	-	-
	10. Socio-emotional ^a	.12***	.13***	.12***	.13***	.23***	.31***	.26***	.25***	.24***	-	-	-	-	-	-	-	-	-	-
	11. Gross motor	.50***	.42***	.40***	.37***	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	12. Fine motor	.18***	.20***	.15***	.20***	-	-	-	-	-	-	.25***	-	-	-	-	-	-	-	-
INTERNDA	13. Cognitive	.21***	.25***	.24***	.23***	-	-	-	-	-	.27***	.58***	-	-	-	-	-	-	-	-
	14. Receptive Language	.16***	.18***	.20***	.16***	-	-	-	-	-	.16***	.35***	.64***	.33***	-	-	-	-	-	-
	15. Expressive Language	.34***	.36***	.42***	.33***	-	-	-	-	-	.27***	.34***	.43***	-	-	-	-	-	-	-
	16. PRIDI	.37***	.44***	.47***	.40***	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
INTERNDA	17. MacArthur-Bates	.46***	.55***	.60***	.51***	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	18. ASQ:SE	.23***	.33***	.26***	.31***	-	-	-	-	-	-	-	-	-	-	-	-	.20**	-	-
	19. Stimulation	.22***	.25***	.21***	.22***	.02	.05	.05	.02	.09**	.07*	.13**	.04	.09	.12*	.01	.18***	.15*	.16*	-
	20. HAZ	.20***	.19***	.19***	.16***	.19***	.10**	.07*	.08*	.15***	.05	.20***	.03	.11**	.19***	.12***	.08*	-	-	.03

Notes: ^aCaregiver reported, ***p<0.001, **p<0.01, & *p<0.05. All correlations adjusted for children's age and site location.