Development and validation of an early childhood development scale for use in low-resourced settings

Development and Validation of an Early Childhood Development Scale for Use in Low-Resourced Settings

Dana Charles McCoy[a,b], PhD, dana_mccoy@gse.harvard.edu, corresponding author

Christopher R. Sudfeld[a], ScD, csudfeld@hsph.harvard.edu

David C. Bellinger[a,c], PhD, MSc, David.Bellinger@childrens.harvard.edu

Alfa Muhihi[d], MD, selukundo@gmail.com

Geofrey Ashery[d], giashery@gmail.com

Taylor E. Weary[a], taylorweary@college.harvard.edu

Wafaie Fawzi[a], MBBS, DrPH, mina@hsph.harvard.edu

Günther Fink[a], PhD, gfink@hsph.harvard.edu

[a]Harvard School of Public Health, Boston, MA; [b]Harvard Graduate School of Education, Cambridge, MA; [c]Harvard Medical School, Boston, MA; and [d]Ifakara Health Institute, Ifakara, Tanzania

Abstract

Background: Low-cost, cross-culturally comparable measures of the motor, cognitive, and socioemotional skills of children under 3 years remain scarce. In the present paper, we aim to develop a new caregiver-reported early childhood development (ECD) scale designed to be implemented as part of household surveys in low-resourced settings.

Methods: We evaluate the acceptability, test-retest reliability, internal consistency, and discriminant validity of the new ECD items, subscales, and full scale in a sample of 2,481 18- to 36-month-old children from peri-urban and rural Tanzania. We also compare total and subscale scores with performance on the Bayley Scales of Infant Development (BSID-III) in a subsample of 1036 children. Qualitative interviews from 10 mothers and 10 field workers are used to inform quantitative data.

Results: Adequate levels of acceptability and internal consistency were found for the new scale and its motor, cognitive, and socioemotional subscales. Correlations between the new scale and the BSID-III were high (r>.50) for the motor and cognitive subscales, but low (r<.20) for the socioemotional subscale. The new scale discriminated between children's skills based on age, stunting status, caregiver-reported disability, and adult stimulation. Test-retest reliability scores were variable among a subset of items tested.

Conclusions: Results of this study provide empirical support from a low-income country setting for the acceptability, reliability, and validity of a new caregiver-reported ECD scale. Additional research is needed to test these and other caregiver reported items in children in the full 0 to 3 year range across multiple cultural and linguistic settings.

Keywords: Early child development, low-income countries, measurement, validation, 0-3

Background

Mounting evidence suggests the importance of investing in early childhood development (ECD) for enhancing the economic, health, and educational status of individuals, communities and nations [1-4]. Over the past several decades, a number of well-validated tools have been developed for measuring individual children's motor, cognitive, language, and social functioning during the first years of life (e.g., Griffiths Mental Development Scales, Denver Developmental Screening Test, Bayley Scales of Infant and Toddler Development). These direct assessments are typically done by clinically trained personnel and provide detailed information on individuals' developmental status that can be used for informing clinical decisions, understanding developmental processes, or testing the efficacy of early interventions [5].

Despite their utility in capturing rich data, individual-level assessments are limited in their ability to provide estimates of population-level developmental status for several reasons. First, many of these assessments are quite costly in terms of their copyrights, the time they take to administer, as well as the resources necessary to train assessors, making them impractical for use at scale [6]. Second, the majority of existing developmental assessments have been created with one particular – primarily high-resourced, Western – cultural context in mind. Although great advances have been made recently in developing new tools for non-Western, low-resourced settings (e.g., the Malawi Developmental Assessment Tool, the Inter-American Development Bank's PRIDI tool, the Developmental Milestones Checklist, the East Asia and Pacific Early Child Development Scales), the utility of these assessments for making generalizations outside of the context in which they were developed is unknown [7-8]. Finally, many comprehensive developmental assessments have focused primarily on motor, cognitive, and language development, while neglecting to integrate early manifestations of social, emotional, and

69 regulatory competence. Although many socioemotional skills vary in importance and
70 developmental determinants cross-culturally, research has increasingly shown the early
71 emergence of a core, basic set of these capacities to be strongly related to later-life outcomes in
72 diverse parts of the world [6, 9-13].

73 In recent years, several new tools have been developed to address these limitations and
74 provide comprehensive population-level data in older children (e.g., UNICEF's Early Childhood
75 Development Index for 3- and 4-year-olds, the Early Development Index for school-aged
76 children) [14, 15]. Still, no such scale exists for the under three age period, when children's
77 brains and bodies are developing most rapidly and are most susceptible to intervention [16].
78 Given that target 4.2 of the Sustainable Development Goals aims to "ensure that all girls and
79 boys have access to quality ECD," measuring children's developmental status at the population
80 level is of important policy relevance [17]. Internationally validated and valid tools would
81 provide a new opportunity for global ECD advocates to quantify children's needs across
82 countries and regions, to make more informed decisions regarding policies and resource
83 allocation, and to monitor progress in achieving global goals congruent with the post-2015
84 agenda [18].

85 In this study, we describe the development of a set of caregiver-reported items for
86 quickly and easily measuring the motor, cognitive, and socioemotional skills of children under
87 three living in low-resourced settings, collectively known as the Caregiver-Reported Early
88 Development Index (CREDI). Our focus on a caregiver report format allows us to address
89 several practical and conceptual challenges of using direct assessment with large groups of
90 infants and toddlers. Compared to direct assessments, caregiver reports require limited training
91 and implementation time, provide a more generalizable perspective on children's skills and

behaviors across time and setting, are more appropriate for capturing socioemotional skills, and are less likely to be biased against children who are unfamiliar with clinical assessments, who are shy with strangers, or who do not understand verbal instructions [5]. In particular, the CREDI is designed to be 1) simple and clear enough to be answered by a caregiver with minimal formal education, 2) short enough to be feasibly integrated within large-sample household data collection efforts, 3) sufficiently "culturally neutral" to allow for cross-context comparison, and 4) adequately aligned with "gold standard" direct assessment measures of proven clinical and developmental utility. In creating the CREDI, our ultimate aim is to generate a new tool that will serve to provide conceptually rich, developmentally informed, population-level data on global progress in alleviating ECD-related inequities and meeting target 4.2 of the SDGs. In the present paper, we detail the initial validation of the CREDI using qualitative and quantitative data among 18- to 36-month-old children in peri-urban and rural Tanzania, including evidence of the individual items' and overall scale's acceptability, reliability, and validity. We conclude by describing the implications of this generative work for future validation and expansion efforts.

Methods

Study sample

The sample for the present study was comprised of children 18 to 36 months who had previously participated in a neonatal vitamin A supplementation trial in the Morogoro region of Tanzania (registered at anzctr.org.au as ACTRN12610000636055) [19], as well as the person in the household who reported to spend the most time caring for that child (i.e., their primary caregiver). This particular area of Tanzania was selected over alternate study locations due to its 1) track record and infrastructure for conducting high-quality early childhood research, and 2) similarity to the broader population of Tanzania with regard to its high prevalence of poverty and

115 malnutrition, mix of peri-urban and rural settings, and cultural diversity. Newborns were eligible
116 for the original vitamin A study if they were able to feed orally, were born within the past 72
117 hours, were not already enrolled in other clinical trials, their family intended to reside in the
118 study area for at least six months post-delivery, and their caregivers provided written informed
119 consent. Notably, results of the original vitamin A trial revealed no detectable impacts on
120 children's developmental outcomes [19] , suggesting that randomization in the original study
121 should not have affected the results of the present analysis.
122 For the original trial, a total of 20,104 randomly selected children living in Morogoro
123 region were enrolled. For the follow-up study, sampling was restricted to children from the
124 original trial living within the Ifakara Demographic Surveillance Site (IHI DSS). No other
125 exclusion criteria (e.g., based on disability or health status) were applied. Given this, the sample
126 is representative of the greater Ifakara area, with all eligible children in Ifakara town and the
127 surrounding villages being equally likely to be selected for participation. In keeping with the aim
128 of the study to validate the CREDI for children 18-36 months, only those within this age range
129 were selected, with the specific age of the child varying non-systematically based on the timing
130 of initial recruitment to the vitamin A study and the timing of the CREDI assessment (38% 18-
131 23mo, 25% 24-29mo, and 38% 30-36mo). Children in the present sample were found to be
132 comparable to those sampled from the 2015-2016 Tanzanian national Demographic Health
133 Survey (DHS) in rates of stunting (43.3% vs. 43.8%, respectively; [20]. Compared to the
134 Tanzanian average, mothers in this sample were more likely to have attended primary school
135 than those in the DHS (87.9% vs. 61.9%, respectively), but less likely to have completed
136 secondary school or higher (7.3% versus 23.4%).
137 **Ethics**

All study protocols were approved by institutional review boards (IRBs) at the Harvard School of Public Health, the National Institute of Medical Research of Tanzania, and the Ifakara Health Institute. Caregivers provided written consent for their own participation and the participation of their children after a field worker read the consent out loud and answered any questions. All study staff were trained and monitored in IRB-approved procedures for identifying participant needs and, as necessary, providing referrals to local physical and mental health services.

Item development phase

Multiple steps were taken to develop the ECD items analyzed in this study. First, we reviewed the ECD measurement literature to help us to define 1) the purpose of the scale, 2) the age-appropriate developmental domains and constructs to be covered by the scale, and 3) the validation plan. Second, and based on the literature review, we built an inventory of existing measurement tools from high-, middle-, and low-income country contexts (see Appendix A), and identified gaps in their coverage of our age-specific domains and constructs. Third, we selected, adapted, and/or created an initial set of items based on the following criteria:

Each item must:

1) have evidence for face, construct, and/or criterion validity for representing one of the core ECD domains[1]
2) be developmentally appropriate for children 18 to 36 months[2]

[1] Although not representative of the core ECD domains, one item – whether the child was frequently too sick to play – was borrowed from the MICS ECDI and included alongside the motor items to test its utility in the <3 age group. This health item was tested in the qualitative and quantitative pilots but not included in reliability or validity analyses of the total CREDI scale.

[2] In several cases, items appropriate for younger ages (down to 12 months) were included to assess their suitability for older children living in an under-studied, at-risk sample.

3) be reportable by a primary caregiver on a yes/no response scale (i.e., the item cannot be task-based, cannot be rated on a continuous scale[3], and must be sufficiently concrete that a caregiver would already be familiar with the specified behavior/skill in the child)
4) be simple in wording to allow for easy translation and comprehension by caregivers with minimal formal education
5) have the potential to discriminate between individuals (i.e., indicate a high likelihood of variability in response)
6) not be subject to severe social desirability (i.e., a caregiver will not feel compelled to respond in a particular way in order to please the assessor or avoid shame/embarrassment)
7) be culturally neutral (i.e., involve skills, behaviors, objects, ideas, or terminology that are common across contexts)

Each of these three phases was led by the study authors, with results reviewed by a group of advisory team members who represented multiple backgrounds (e.g., research, practice, policy), fields (e.g., health, nutrition, psychology, education), and geographical contexts (e.g., United States, sub-Saharan Africa, Asia, Latin America). Advisory group members provided oral and written feedback on study procedures and materials via bi-monthly conference calls, formal surveys, and informal communications (e.g., emails, one-on-one meetings).

Finally, all items were translated and back-translated to/from Swahili by bilingual Tanzanian and American study staff. Discrepancies in translation were resolved based on the

[3] In the case of items of child behavior traditionally measured using Likert response scales (e.g., never/sometimes/often/always), we integrated "frequency anchors" into the questions themselves to indicate the prevalence of behavior necessary to achieve a "yes" versus a "no" response (e.g., "Does the child get along well with other children *most of the time*?"). These frequency anchors were selected to discriminate between adaptive vs. non-adaptive behavior for each item.

consensus of a committee comprised of CREDI team members, local staff, and bilingual Tanzanian community members.

Qualitative pilot phase

To provide preliminary feedback on the initial set of items, we conducted a series of "cognitive" (qualitative) interviews in December of 2013 with 10 caregiver-child pairs in and around Ifakara, Tanzania (mean age of children = 28.2 months, range = 20-35 months). A local female research scientist with a Master's degree in human development was recruited based on her previous experience conducting qualitative research in the study community. The interviewer conducted interviews one-on-one with caregivers in children's homes using a semi-structured interview protocol designed to elaborate each item's acceptability, clarity, and applicability, as well as the comprehensiveness and redundancy of the scale as a whole [21, 22]. Specifically, the interviewer asked the caregiver (all of whom happened to have been mothers) to respond to each item based on her child's ability or behavior. The interviewer then asked one or more in a series of seven follow-up questions designed to elicit the caregiver's perceptions of the item, her thought process in responding to the item, and/or her suggestions for improving the item. At the end of each interview, the caregiver was also asked to give her general impressions of what positive ECD means to her, the acceptability of the scale, and whether she had any suggestions for improving the scale. (For the full interview protocol, contact the first author.) The results of these interviews were used to provide preliminary information regarding the overall acceptability of the scale, as well as to identify items that required further adaptation or elimination prior to larger-scale quantitative testing.

Quantitative pilot phase

198 Following the qualitative phase, we conducted a full quantitative pilot from January to
199 October of 2014 in 2,481 caregiver-child pairs, of which 2,320 (93.5%) included mothers, 68
200 (2.7%) included fathers, and 93 (3.8%) included other family members (e.g., grandparents,
201 aunts). Child-caregiver pairs who participated in the qualitative pilot portion of the validation
202 study were excluded from participation in the full quantitative pilot phase. Of the 4,356 children
203 randomly selected for a home visit, 2,481 (57.0%) completed the visit, 558 (12.8%) were
204 temporarily away, 1,204 (27.6%) had permanently moved, 60 (1.4%) had died, and 53 (1.2%)
205 had caregivers who refused to participate. The characteristics of those who completed the home
206 visit versus those who were invited but did not complete the home visit are shown in Appendix B
207 and indicate relative similarity across the groups. Each caregiver-child pair was visited in their
208 home, invited and consented to participate, and interviewed using all items on the CREDI.
209 Caregivers also reported on cognitive stimulation using six items from UNICEF's Multiple
210 Indicator Cluster Survey ECD module capturing adult-child interactions [14] and children's
211 physical and mental disability using six items from the Ten Questions screener [23]. Stimulation
212 items reflected whether an adult household member had engaged the child in six different
213 activities (e.g., reading, counting, playing, singing) over the preceding three days. Children were
214 grouped into low (0-2 activities), moderate (3-4 activities), and high (5-6 activities) stimulation
215 categories for analyses. Disability items reflected children's difficulty with seeing, hearing,
216 moving, and learning. Children were considered to have a disability if their caregiver answered
217 "yes" to any of the six screening items.
218 Home visits were completed by eight male, secondary school-educated field workers with
219 previous experience conducting field-based research with families and children in the local area.
220 Field workers were selected based on their performance as data collectors in the original vitamin

A study and participated in a two-day training on the CREDI and other study visit procedures. All workers were also monitored by the study coordinator in the field on a bi-weekly basis to ensure continued adherence to study protocols. During the home visits, field workers rated their perceptions of caregivers' understanding of and honesty in responding to the CREDI items. They also recorded any questions or concerns stated by the caregivers during the interview. At the end of the visit, field workers measured children's height to the nearest 0.1 cm. Children less than 24 months were measured using a Seca length board, whereas those 24 months or older were measured using a portable Seca stadiometer. Field workers measured height twice in a row, and if the two values differed by more than 0.2 cm, they repeated the measurement a third time, taking an average of the two closest values. Table 1 shows descriptive statistics for this sample.

Table 1. Descriptive characteristics of the quantitative pilot sample

	N	Mean/%	SD	Min	Max
CREDI[a]					
Total Score (n=44 items)	2481	0.64	0.17	0.07	0.98
Motor (n=5 items)	2481	0.63	0.24	0.00	1.00
Cognitive (n=19 items)	2481	0.64	0.29	0.00	1.00
Socioemotional (n=20 items)	2481	0.64	0.15	0.10	1.00
Bayley Scales of Infant Development-III					
BSID Cognitive	959	60.50	8.67	30	81
BSID Receptive Communication	950	25.78	7.00	5	42
BSID Expressive Communication	947	30.06	8.58	3	46
BSID Fine Motor	955	40.56	6.55	12	62
BSID Gross Motor	960	57.00	5.56	34	70
BSID BOI - Caregiver	1033	1.51	0.35	0	2
BSID BOI - Assessor	1033	1.56	0.25	0	2
Child and Family Characteristics					
Child female	2481	45.6%			
Child age (months)	2481	27.07	6.08	17.03	37.08
Child height-for-age z-score	2177	-1.82	1.28	-5.99	4.94
Child stunted (HAZ<-2)	2177	43.3%			

Child any disability	2481	1.9%			
Proportion of stimulation activities conducted (out of 6)	2480	0.49	0.16	0.00	1.00
Maternal educ - No school	2481	4.6%			
Maternal educ - Primary school	2481	86.2%			
Maternal educ - Secondary school	2481	7.3%			

Note: [a] CREDI mean scores represent proportion of correct responses on the scale or sub-scale. Scores calculated based on the final set of 44 items only.

Approximately 60 percent of caregiver-child pairs were selected by a computer-generated random number draw before their home visit to be invited to an additional clinic visit, which occurred one to six days after the home visit. Of the 1,478 children randomly selected for a clinic visit, 1,037 (70.2%) completed the visit, 224 (15.2%) agreed to the visit but did not show up, 57 (3.9%) refused the visit, and the remainder (10.8%) were not scheduled due to logistical reasons (e.g., caregiver or child was ill, no clinic appointments were available). The characteristics of home visit participants who completed the clinic visit versus those who did not complete the clinic visit are shown in Appendix C and indicate relative similarity across the groups. During the clinic visit, a female nurse with training in child development and research re-administered a subset of 11 CREDI items (selected for their conceptual diversity) and conducted an adapted and translated version of the Bayley Scales of Infant Development (BSID-III) [24] with the child, including all direct assessment subscales as well as the Behavior Observation Inventory (BOI). The BSID-III was chosen as the comparison metric for the present study due to its acceptance as a "gold standard" clinical assessment with strong reliability and validity, its complementary direct assessment format, and its previous use by our team in Tanzanian ECD research [25-27].

Because the BSID-III was originally developed in the United States, field and research staff completed a detailed adaptation process over the period of several weeks to improve its

applicability within the Tanzanian context. Details of the training, adaptation, and psychometric properties of the BSID-III can be found in Sudfeld et al. [28]. Briefly, six nurses were trained to administer the BSID-III by two American PhD-level psychologists over a three-week period, after which four nurses were selected as study staff based on quantitative ratings of their performance and knowledge. Study nurses were each monitored by the local study coordinator on a biweekly basis to ensure quality and to avoid assessor drift. To enhance cultural applicability, unfamiliar images and terminology within 13% of BSID-III items (n=30) were replaced using more culturally relevant stimuli (e.g., changing a picture of an apple to a banana) based on local expert consensus. To maintain functional equivalence, replacement stimuli were selected to be of similar size, style, and complexity to original stimuli. Raw scores were used for analyses due to lack of Tanzania-specific age-norms. At the end of the clinic visit, nurses recorded mothers' questions and any problems that may have precluded full completion of the visit (e.g., child was sick or uncooperative).

Data from the quantitative pilot phase were used at the item level to understand individual items' distributional properties, including pass/fail rates and levels of non-response (i.e., "don't know" answers). Test-retest reliability was assessed for the 11 CREDI items tested in both the home and clinic visit. Additional tests of reliability and validity were performed for items that were identified to have sufficient variability (i.e., that did not show evidence for floor or ceiling effects). Specifically, internal consistency was captured within each of the three CREDI domains/subscales using Cronbach's alpha. Discriminant validity was assessed by comparing CREDI total and subscale scores across a set of child and family characteristics, including child age, gender, stunting status (height-for-age z-score of <2SDs below the WHO standard) [29], caregiver-reported disability, caregiver-reported cognitive stimulation in the

home, and maternal education (which was collected at children's births as part of the original vitamin A study). Finally, concurrent validity was assessed by correlating each CREDI subscale score with the corresponding BSID-III raw score. Psychological field standards (e.g., Cicchetti, 1994) [30] were used as the basis for determining acceptability of the items' and subscales' reliability and validity.

Field staff interviews

At the end of the quantitative pilot phase, 10 qualitative "exit" interviews were conducted with field staff (including 6 field workers, 3 nurses, and 1 field supervisor) to identify areas of confusion, difficulty, or lack of clarity in the CREDI based on their experiences over nine months of data collection.

Results

Item development & qualitative interviews

Review of the literature and consultation with ECD experts resulted in the identification of three primary domains – motor, cognitive/language, and socioemotional skills – and 12 constructs or subdomains for inclusion in the CREDI (see Table 2). Based on a review of existing ECD measurement tools (see Appendix A) and the process of identifying conceptual gaps, an initial set of items was developed by the core research team. Whereas many of these items were highly similar to questions from existing ECD assessments, a substantial number – particularly from the socioemotional domain, where the largest conceptual gaps were identified – were completely novel. Following a round of revisions to the items by the ECD expert team, a total of 92 items were submitted for initial qualitative pilot testing. Following qualitative interviews, 22 items (n=7 from motor, n=8 from cognitive, and n=7 from socioemotional) were dropped from the CREDI for the following reasons: the item was too easy/hard for children of

this age group (n=10), the item was redundant with another item (n=8), the item was confusing and could not be easily clarified (n=3), and the item was culturally inappropriate and could not be easily adapted (n=1). Of the remaining 70 items, 15 (n=1 for motor, n=7 for cognitive, and n=7 for socioemotional) were adapted prior to the quantitative pilot based on suggestions from cognitive interview participants and additional consultation with local experts. These adaptations primarily involved the addition of examples to improve item clarity, such as changing "Does the child know any numbers?" to "Does the child know any numbers (e.g., one, two, three)?" In several instances, words relating to culturally specific objects (e.g., toys) were removed or replaced.

Table 2. Domains and constructs of the CREDI

DOMAINS	Motor	Cognitive	Socioemotional
CONSTRUCTS	1) Fine 2) Gross	1) Expressive language 2) Receptive language 3) Preacademic skills/knowledge 4) Reasoning & problem solving	1) Early executive function & effortful control 2) Emotion regulation 3) Externalizing symptoms 4) Internalizing symptoms 5) Reactivity & soothability 6) Social competence

Acceptability

Cognitive interviews revealed that 10/10 caregivers were cooperative with and felt pleased by the items, and 9/10 felt that "there were no right or wrong answers." (One mother of a 20-month-old child reported, "I was uncomfortable when you asked me things which my child cannot do, as she is too young.") Field workers' average ratings of whether the caregivers understood the questions during the quantitative pilot was 3.85 (SD=0.28) and whether they appeared to answer truthfully was 3.77 (SD=0.36) on a scale of 1 (No, not at all) to 4 (Yes, all

questions). In addition, exit interviews of field staff identified no problems with items' demand characteristics, with the exception of a socioemotional item capturing whether the child "gets along well with other children most of the time" that was reported by 5 of the 11 field workers as eliciting problems with social desirability.

Item analysis

Results of item analyses to understand the completeness, distribution, and relative difficulty of each item as measured during the quantitative pilot home visit can be found in Appendix B. Results revealed that 25 of the 70 items (n=10 for motor, n=8 for cognitive, n=7 for socioemotional) showed evidence of ceiling effects, with pass rates of >95%. In general, these items tended to represent more basic developmental skills that may be more appropriate for children <18 months (e.g., walking, achieving object permanence, saying one word, showing affection). These items were removed from the final subscales used for reliability and validity analyses. Figure 1 summarizes the item selection process. Figures 2 to 4 show score distributions by age.

<< Figures 1-4 here >>

"Don't know" responses were infrequent across the CREDI, with an average of 1.8% of the sample responding "don't know" for any given item during the home visit. In comparison, among 1,037 BSID-III assessments, 9.9% were incomplete and an additional 10.1% were flagged by nurses as challenging or unreliable due to children's illness, injury, uncooperativeness, or distraction. Of the items that were most frequently answered as "don't know," the majority were also acknowledged as unclear in the qualitative interviews due translation difficulties (e.g., inability to find an equivalent word or set of words for "distracted"

in Swahili) or lack of a concrete behavioral marker (e.g., ambiguity of what it means to show sympathy or concern).

Reliability

A total of 26 items were excluded from the original 70-item set due to ceiling effects (n=25) and the lack of conceptual fit with a specific developmental domain (n=1, "too sick to play"). Cronbach's alpha coefficients calculated in the final set of 44 items suggested acceptable internal consistency / inter-item reliability for motor (α=.68), cognitive (α=.90), and socioemotional (α=.68) items. Kappa coefficients were used to capture the reliability of responses from the same caregiver to 11 items administered at both the home and clinic visits (see Table 3). It should be highlighted that the Kappa statistic was originally developed as a measure of inter-rater reliability, where two raters directly observe or assess the same individual at the same time. In the case of the present study, our Kappas capture both test-retest reliability (with an average time between study visits of 3.17 days [SD=2.11]) and inter-rater reliability (between male home visitors and female clinic nurses). Given this, they represent both true variation in children's skills over time, as well as multiple potential sources of measurement error. As such, we might expect our Kappas to be lower than those used simply to capture inter-rater reliability. Indeed, results indicate differential reliability, with 2 items showing moderate reliability (Kappa≥0.40), 6 items showing fair reliability (Kappa≥0.20), 2 items showing slight reliability (Kappa≥0.00), and 1 item showing poor reliability (Kappa<0.00). Additional analyses revealed no consistent evidence for systematic differences in mean scores across home and clinic visits (see Table 3) or for substantial differences in Kappa values based on the time delay between the home and clinic visit (contact first author for detailed results).

Table 3. Test-retest reliability of 11 select items across data collection contexts with same caregiver reporter (n=962)

	Mean Score		Difference in means	% Agreement	Kappa
	Home Interview	Clinic Interview			
Does the child walk several steps without the support of a person or object (e.g., wall or furniture)?	0.99	0.99	0.00	99.3%	0.663
Does the child know the names of at least two body parts (e.g., arm, eye, or nose)?	0.67	0.65	0.02	80.4%	0.563
Does the child say five or more words (e.g., names like Mama or objects like cup)?	0.91	0.93	-0.02	90.4%	0.335
When asked what common objects (like a cup or a knife) are for, does the child explain correctly?	0.34	0.20	0.14	74.4%	0.363
Does the ever child kick, bite, or hit other children or adults?[r]	0.45	0.31	0.14	68.0%	0.333
Does the child pick up a small object like a rock with just his/her thumb and a finger?	0.78	0.84	-0.06	78.0%	0.276
Does the child get along well with other children most of the time?	0.92	0.97	-0.05	92.2%	0.241
When the child is upset, is he/she able to calm down by him/herself?	0.45	0.59	-0.14	60.0%	0.213
Does the child pay attention when someone is talking to him/her?	0.90	0.95	-0.05	87.8%	0.156
Does the child follow simple directions (e.g., "Stand up or Come here")?	0.98	0.99	-0.01	97.7%	0.144
Is the child sometimes impatient or unwilling to wait or hold still when you ask him/her to?[r]	0.38	0.48	-0.10	49.4%	-0.020
Average	*0.71*	*0.72*	*-0.01*	*79.8%*	*0.297*

Notes: CREDI mean scores represent proportion of correct responses on the item. [r] indicates item that was reverse coded.

Validity

Table 4 shows the results of tests of discriminant validity for CREDI scores based on child and family characteristics. These results show significantly higher total CREDI scores for children who were older, non-stunted, non-disabled, and from high-stimulation households at the

time of data collection. Effect sizes for these differences ranged from small (d≈0.20SD) for stunting, to large (d>0.50SD) for age, disability, and stimulation. No significant (p<.05) differences were observed for CREDI scores across gender or maternal education with the exception of socioemotional scores, which were highest for children of non-educated mothers (d≈0.20SD).

Table 4. CREDI mean scores (SE) by subgroup (n=2,481)

	Total	Motor	Cognitive	Socioemotional
Child age				
18-24 mo[a] (n=934)	0.50 (0.005)	0.45 (0.008)	0.44 (0.007)	0.57 (0.005)
>24-30 mo[b] (n=614)	0.67 (0.005)	0.67 (0.010)	0.69 (0.008)	0.65 (0.005)
>30-36 mo[c] (n=933)	0.76 (0.004)	0.82 (0.007)	0.79 (0.005)	0.71 (0.004)
dif	F(2, 2478)=965.00**	F(2, 2478)=560.63**	F(2, 2478)=860.93**	F(2, 2478)=234.26**
	a<b: d=1.00**	a<b: d=0.78**	a<b: d=1.03**	a<b: d=0.52**
	b<c: d=0.51**	b<c: d=0.50**	b<c: d=0.41**	b<c: d=0.39**
	a<c: d=1.51**	a<c: d=1.28**	a<c: d=1.44**	a<c: d=0.92**
Child gender				
Male (n=1,349)	0.64 (0.005)	0.64 (0.008)	0.63 (0.007)	0.64 (0.004)
Female (n=1,132)	0.64 (0.005)	0.64 (0.009)	0.64 (0.007)	0.64 (0.004)
dif	t(2479)=-0.70	t(2479)=0.08	t(2479)=-1.34	t(2479)=0.31
	d=0.03	d=-0.00	d=0.06	d=-0.01
Child stunting				
Non-Stunted (n=1,222)	0.66 (0.005)	0.66 (0.008)	0.66 (0.007)	0.65 (0.004)
Stunted (n=955)	0.62 (0.006)	0.62 (0.010)	0.60 (0.008)	0.65 (0.005)
dif	t(2175)=4.79**	t(2175)=3.46**	t(2175)=6.09**	t(2175)=0.86
	d=-0.19**	d=-0.13**	d=-0.25**	d=-0.04
Child disability				
No disability (n=2,434)	0.64 (0.003)	0.65 (0.006)	0.64 (0.005)	0.64 (0.003)
Any disability (n=47)	0.53 (0.032)	0.45 (0.050)	0.48 (0.047)	0.60 (0.025)
dif	t(2479)=4.43**	t(2479)=4.73**	t(2479)=4.45**	t(2479)=1.81+
	d=-0.65**	d=-0.69**	d=-0.65**	d=-0.27+
Stimulation				
Low stimulation[a] (n=847)	0.63 (0.005)	0.62 (0.10)	0.63 (0.008)	0.64 (0.005)
Mod stimulation[b] (n=1,498)	0.63 (0.005)	0.65 (0.08)	0.62 (0.006)	0.64 (0.004)
High stimulation[c] (n=135)	0.75 (0.011)	0.77 (0.021)	0.81 (0.014)	0.68 (0.014)
dif	F(2, 2477)=29.57**	F(2, 2477)=17.41**	F(2, 2477)=41.00**	F(2, 2477)=3.65*
	a<b: d=0.02	a<b: d=0.09+	a<b: d=-0.02	a<b: d=0.03
	b<c: d=0.67**	b<c: d=0.45**	b<c: d=0.80**	b<c: d=0.22*
	a<c: d=0.68**	a<c: d=0.54**	a<c: d=0.77**	a<c: d=0.25*
Maternal education				

No education[a] (n=113)	0.63 (0.016)	0.61 (0.029)	0.61 (0.022)	0.67 (0.015)
Primary school[b] (n=2,138)	0.64 (0.004)	0.65 (0.026)	0.64 (0.005)	0.65 (0.003)
Secondary school[c] (n=181)	0.61 (0.013)	0.63 (0.022)	0.61 (0.019)	0.62 (0.012)
dif	$F(2, 2429)=2.63+$ a<b: $d=0.14$ b<c: $d=-0.18+$ a<c: $d=-0.04$	$F(2, 2429)=1.36$ a<b: $d=0.20$ b<c: $d=-0.08$ a<c: $d=0.12$	$F(2, 2429)=2.29+$ a<b: $d=0.19$ b<c: $d=-0.14$ a<c: $d=0.05$	$F(2, 2429)=4.36**$ a<b: $d=-0.04$ b<c: $d=-0.19*$ a<c: $d=-0.24*$

Note: ** $p<.01$, * $p<.05$, + $p<.10$; CREDI mean scores represent proportion of correct responses on the scale or sub-scale; *d* indicates effect size of standardized mean differences as represented by Cohen's *d*

Figure 5 shows the correlations between the CREDI and BSID-III subscales. Linear bivariate correlations between the CREDI motor items and the BSID-III fine and gross motor subscales were $r=.50$ and $r=.51$, respectively. Correlations between the CREDI cognitive items and the BSID-III cognitive, receptive communication, and expressive communication subscales were $r=.68$, $r=.69$, and $r=.73$, respectively. All of these correlations were significant at the $p<.001$ level. Correlations between the CREDI socioemotional items and the BSID-III BOI were much smaller, at $r=.16$ ($p<.001$) for the caregiver-reported BOI and $r=.09$ ($p<.01$) for the examiner-reported BOI.

<< Figure 5 here >>

Discussion

The primary aim of the present study was to describe initial evidence for the acceptability, reliability, and validity of the newly developed CREDI as a measure of ECD designed for feasible use within standard household surveys in low-resourced settings. Results of our initial validation effort in Tanzania suggest that the CREDI tool may provide a valid method for capturing young children's development across motor, cognitive, and socioemotional domains. In particular, the CREDI was able to clearly discriminate between the skills of younger versus older children, children with adequate versus low nutritional status, disabled versus non-

410 disabled children, and children from more versus less cognitively stimulating households, while
411 showing evidence for equality across gender and maternal education within a large quantitative
412 sample. Collectively, the items also showed adequate criterion validity with the BSID-III motor,
413 cognitive, and communication subscales, which are "gold standard" direct assessments of
414 children's early developmental status often used in clinical settings by highly trained staff.
415 In addition to showing positive evidence for validity, the CREDI was found to be an
416 acceptable tool for use in low-resourced settings. It was well understood by respondents and
417 quick to implement (taking an average of 20 minutes to administer in total) by trained field staff
418 with the equivalent of a secondary education. Furthermore, initial findings suggest that the
419 caregiver report format may be advantageous for use with young children in low-resourced
420 settings in order to avoid problems with non-compliance (e.g., due to unfamiliarity with testing
421 situations, fear of unfamiliar adults, child illness, etc.) that were found to affect the quality and
422 completeness of nearly 20 percent of BSID-III direct assessments.
423 Although the CREDI as a whole shows promise as an acceptable and valid measurement
424 tool, test-retest reliability was also low for many individual items, and particularly for those that
425 qualitative interview respondents noted were difficult to translate or lacking in examples,
426 benchmarks, or behavioral markers. Given that no systematic differences were found based on
427 the interviewer, setting, or time between visits, these results suggest that further adaptation is
428 needed to make items as concrete as possible and reduce respondent "guessing." Additional
429 reliability testing, qualitative work, and empirical analysis (e.g., item response theory) are also
430 warranted in future work to ensure that items' interpretation is occurring similarly across time,
431 context, respondent, and assessor.

In addition, these results revealed a relatively weak correspondence between the socioemotional items and the BSID-III BOI. This low correlation was not particularly surprising given that 1) the BOI was not designed as a measure of socioemotional functioning, per se, and 2) our aim in developing the socioemotional items was to capture a large breadth of important but potentially non-overlapping developmental constructs. Our review of the literature and consultation with ECD experts revealed that the vast majority of previous measurement tools (like the BSID-III) have focused on young children's motor and cognitive development, with far fewer options for capturing social, emotional, and higher-order cognitive processes like self-regulation and executive function that are increasingly being shown by the literature to predict later life outcomes [31-33, 10, 12]. Given that our socioemotional items showed adequate reliability and validity in other ways (e.g., internal consistency, discrimination by age, caregiver-reported stimulation, etc.), we are confident that their inclusion represents an important advance over previous work in this age group. At the same time, we acknowledge the need for further validation against alternative socioemotional measurement approaches (e.g., the Ages and Stages personal-social and socio-emotional scales, observer ratings of child behavior during assessment) and clinical diagnoses, as well as examinations of predictive validity over time in diverse settings, particularly given a lack of understanding of these early skills cross-culturally. Additional research is also needed to explore the somewhat counterintuitive finding that less educated caregivers report the highest levels of socioemotional development for their children.

Despite the strengths of this study, the research presented also has several important limitations that must be addressed through future work. First, and most importantly, our focus on a single geographic context substantially limits the generalizability of these results. Second, the number of qualitative interviews conducted in this study was quite small, and focused only

455 on mothers. Third, as is noted previously, our lack of a "gold standard" metric against which to
456 compare our socioemotional items limits our understanding of their concurrent validity. Fourth,
457 our additional measures of context and disability were limited and coarse, and may not have been
458 suitable for fully describing the risks and challenges faced by children. Finally, the cross-
459 sectional nature of our data collection effort precludes our ability to draw conclusions about the
460 CREDI's long-term predictive validity. To address these limitations, we plan to continue
461 validation of the CREDI using 1) a large number of geographically, linguistically, and culturally
462 diverse contexts, 2) different types of caregivers, 3) a wider range of locally-generated
463 comparison and diagnostic metrics, and 4) longitudinal data. In particular, additional qualitative
464 and quantitative work is currently underway in multiple countries to improve the clarity and
465 objectivity of items in an attempt to improve test-retest reliability. Based upon the results of
466 these ongoing and future efforts, we hope to finalize and disseminate the CREDI as an open-
467 source tool for governments, agencies, and organizations to quantify developmental status at a
468 population level and track progress in alleviating ECD-related disparities around the world.

469 **Conclusions**

470 Given growing justification for and investment in the promotion of positive development
471 in the first 1000 days of life, providing a tool for quantifying and monitoring early
472 developmental outcomes – particularly for the 89 percent of children under five globally who
473 live in low- and middle-income country contexts – is critically important [34]. Designed as a
474 comprehensive, caregiver-reported assessment of ECD for children under three, the aim of the
475 CREDI is to provide low-cost, large-scale data that will facilitate decision making regarding
476 intervention and resource allocation, and track global progress in alleviating early developmental
477 disparities. The results of the present study suggest that overall, the CREDI worked well for

capturing ECD behaviors and skills in 18- to 36-month-old children within Tanzania. Additional research in diverse linguistic and cultural contexts and younger age groups is needed to ensure the CREDI's utility prior to full dissemination.

Abbreviations

BOI: Behavior Observation Inventory

BSID-III: Bayley Scales of Infant Development (III)

ECD: Early Childhood Development

CREDI: Caregiver-Reported Early Development Index

Declarations

Ethics approval and consent to participate: Primary caregivers provided informed consent for their own participation, as well as the participation of their children. All study protocols were approved by institutional review boards (IRBs) at the Harvard School of Public Health, the National Institute of Medical Research of Tanzania, and the Ifakara Health Institute.

Consent for publication: Not applicable

Availability of data and material: The datasets generated and analyzed during the current study are not publicly available due to their ongoing analysis but are available from the corresponding author on reasonable request.

Competing interests: The authors declare that they have no competing interests.

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

Author contributions: DM conceptualized and designed the study, developed the data collection instruments, conducted the preliminary analyses, drafted the initial manuscript, and approved the final manuscript as submitted. CS assisted in the conceptualization and design of the study, developed portions of the data collection instruments, reviewed and revised the manuscript, and approved the final manuscript as submitted. DB led the cultural adaptation of the BSID-III and training of data collectors, reviewed and revised the manuscript, and approved the final manuscript as submitted. AM assisted in the development of the data collection instruments, translated study materials, coordinated and supervised data collection, critically reviewed the manuscript, and approved the final manuscript as submitted. GA assisted in the development of the data collection instruments, translated study materials, coordinated and supervised data collection, critically reviewed the manuscript, and approved the final manuscript as submitted. TW conducted the interviews with field staff, analyzed, translated and summarized the resulting qualitative data, critically reviewed the manuscript, and approved the final manuscript as submitted. WF oversaw the design and conceptualization of the study, reviewed and edited the data collection instruments, critically reviewed the manuscript, and approved the final manuscript as submitted. GF assisted in the design and conceptualization of the study, developed portions of the data collection instruments, led the study sampling, conducted the final analyses, reviewed and revised the manuscript, and approved the final manuscript as submitted.

Acknowledgements: The authors would also like to express gratitude to the field staff and study participants who made this research possible.

References

1) Fink G, Peet E, Andrews K, McCoy DC, Sudfeld CR, Danaei G, Ezzati M, Fawzi WW. Schooling and wage income losses due to early childhood developmental delays in low- and middle-income countries: National, regional and global estimates. Am. J. Clin. Nutr. 2015;104:104-112.
2) Heckman JJ. Skill formation and the economics of investing in disadvantaged children. Science. 2006;312:1900-2.
3) Nores M, Barnett WS. Benefits of early childhood interventions across the world: (Under) Investing in the very young. Econ Educ Rev. 2010;29:271-82.
4) Shonkoff JP. Building a new biodevelopmental framework to guide the future of early childhood policy. Child Dev. 2010;81:357-67.
5) Snow CE, Van Hemel SB. Early childhood assessment: Why, what, and how. Washington: The National Academies Press; 2008.
6) Fernald LC, Kariger P, Engle P, Raikes A. Examining Early Child Development in Low-Income Countries: A Toolkit for the Assessment of Children in the First Five Years of Life. In: The International Bank for Reconstruction and Development/The World Bank. 2009. http://siteresources.worldbank.org/EXTAFRREGTOPEDUCATION/Resources/444707-1291071725351/ExaminingECDtoolkitFULL.pdf. Accessed 15 Nov 2016.
7) Hui CH, Triandis HC. Measurement in cross-cultural psychology a review and comparison of strategies. J. Cross-Cult. Psychol. 1985;16: 131-152.
8) Peña ED. Lost in translation: Methodological considerations in cross‐cultural research. Child Dev. 2007;78:1255-64.

9) Keller H, Yovsi R, Borke J, Kärtner J, Jensen H, Papaligoura Z. Developmental consequences of early parenting experiences: Self‐recognition and self‐regulation in three cultural communities. Child Dev. 2004;75:1745-60.
10) Mischel W, Ayduk O, Berman MG, Casey BJ, Gotlib IH, Jonides J, Shoda Y. 'Willpower' over the life span: Decomposing self-regulation. Soc Cogn Affect Neurosci. 2010;6:252-6.
11) Molden DC, Dweck CS. Finding" meaning" in psychology: a lay theories approach to self-regulation, social perception, and social development. Am. Psychol. 2006;61:192.
12) Moffitt TE, Arseneault L, Belsky D, Dickson N, Hancox RJ, Harrington H, Caspi A. A gradient of childhood self-control predicts health, wealth, and public safety. Proc. Natl. Acad. Sci. USA. 2011;108:2693-8.
13) Raver CC. Placing emotional self-regulation in sociocultural and socioeconomic contexts. Child Dev. 2004;75:346-353.
14) Multiple Indicator Cluster Surveys (MICS). In: United Nations Children's Fund (UNICEF). 2009-2012. http://mics.unicef.org/surveys. Accessed 15 Nov 2016.
15) Janus M, Offord DR. Development and psychometric properties of the Early Development Instrument (EDI): A measure of children's school readiness. Can. J. Behav. Sci. 2007;39:1.
16) Shonkoff JP, Boyce WT, McEwen BS. Neuroscience, molecular biology, and the childhood roots of health disparities: building a new framework for health promotion and disease prevention. JAMA. 2009; 301:2252-9.
17) Yoshikawa H. Public Policy and Interventions Designed to Reduce Poverty and Inequality - Proposing and implementing a post-2015 global early childhood

development goal with quality at scale. In: Workshop on Poverty, Inequality, and Education/Stanford Graduate School of Education. 2014. http://edpov.stanford.edu/events/hirokazu-yoshikawa-public-policy-and-interventions-designed-reduce-poverty-and-inequality. Accessed 15 Nov 2016.

18) Sabanathan S, Wills B, Gladstone M. Child development assessment tools in low-income and middle-income countries: how can we use them more appropriately? Arch. Dis. Child. 2015;100:482-8.

19) Masanja H, Smith ER, Muhihi A, Briegleb C, Mshamu S, Ruben J, Neovita Tanzania Study Group. Effect of neonatal vitamin A supplementation on mortality in infants in Tanzania (Neovita): A randomised, double-blind, placebo-controlled trial. Lancet. 2015; 385:1324-32.

20) Ministry of Health, Community Development, Gender, Elderly and Children [Tanzania], Ministry of Health [Zanzibar], National Bureau of Statistics, Office of the Chief Government Statistician, & ICF International. Tanzania Demographic and Health Survey and Malaria Indicator Survey 2015-2016: Key Indicators. In: The Demographic and Health Surveys (DHS) Program. 2016. https://dhsprogram.com/pubs/pdf/PR74/PR74.pdf. Accessed 16 Nov 2016.

21) Bradburn NM, Sudman S, Wansink B. Asking questions: The definitive guide to questionnaire design--for market research, political polls, and social and health questionnaires. 1st ed. San Francisco: John Wiley & Sons; 2004.

22) Desimone LM, Le Floch KC. Are we asking the right questions? Using cognitive interviews to improve surveys in education research. Educ. Eval. Policy Anal. 2004; 26:1-22.

590 23) Durkin MS, Davidson LL, Desai P, Hasan ZM, Khan N, Shrout PE, Zaman SS. Validity
591 of the Ten Questions Screen for childhood disability: results from population-based
592 studies in Bangladesh, Jamaica, and Pakistan. Epidemiol. 1994;5:283-289.
593 24) Bayley N. Bayley Scales of Infant and Toddler Development. San Antonio: The
594 Psychological Corporation; 2006.
595 25) Albers CA, Grieve AJ. Review of Bayley Scales of Infant and Toddler Development—
596 Third Edition. J. Psychoeduc. Assess. 2007; 25:180-190.
597 26) McGrath N, Fawzi WW, Bellinger D, Robins J, Msamanga GI, Manji K, Tronick E. The
598 timing of mother-to-child transmission of human immunodeficiency virus infection and
599 the neurodevelopment of children in Tanzania. Pediatr. Infect. Dis. J. 2006;25:47-52.
600 27) McDonald CM, Manji KP, Kupka R, Bellinger DC, Spiegelman D, Kisenge R, Duggan
601 CP. Stunting and wasting are associated with poorer psychomotor and mental
602 development in HIV-exposed Tanzanian infants. J. Nutr. 2013;143:204-214.
603 28) Sudfeld CR, McCoy DC, Fink G, Muhihi A, Bellinger DC, Masanja H, Fawzi WW.
604 Malnutrition and its determinants are associated with suboptimal cognitive,
605 communication, and motor development among Tanzanian children. J Nutr.
606 2015;145:2705-14.
607 29) WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length,
608 weight-for-height and body mass index-for-age: Methods and development. In: World
609 Health Organization. 2006.
610 http://www.who.int/childgrowth/standards/technical_report/en/. Accessed 15 Nov 2016.
611 30) Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and
612 standardized assessment instruments in psychology. Psychol. Assess. 1994; 6:284.

31) Blair C, Razza RP. Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. Child Dev. 2007; 78:647-663.

32) Caspi A, Moffitt TE, Newman DL, Silva PA. Behavioral observations at age 3 years predict adult psychiatric disorders: Longitudinal evidence from a birth cohort. Arch. Gen Psychiatry. 1996;53:1033-9.

33) Diamond A. Executive functions. Annu Rev Psychol. 2013;64:135-168.

34) State of the World's Children 2014 In Numbers: Every Child Counts Revealing disparities, advancing children's rights. In: United Nations Children's Fund (UNICEF). 2014. https://www.unicef.org/sowc2014/numbers/documents/english/SOWC2014_In%20Numbers_28%20Jan.pdf. Accessed 15 Nov 2016.

Figure Captions

Figure 1. Item selection tree.

Figure 2. Proportion children passing each motor item, by age (n=2,481)

Figure 3. Proportion children passing each cognitive item, by age (n=2,481)

Figure 4. Proportion children passing each socioemotional item, by age (n=2,481)

Figure 5. Histogram of CREDI distribution and local polynomial graph of the relation between CREDI and BSID-III subscale scores (line w/ 95% CI)