






The D-score: a metric for interpreting the early development of infants and toddlers across global settings

Ann M Weber ,^{1,2} Marta Rubio-Codina ,³ Susan P Walker,⁴ Stef van Buuren,^{5,6} Iris Eekhout,⁵ Sally M Grantham-McGregor,⁷ Maria Caridad Araujo,³ Susan M Chang,⁴ Lia CH Fernald,⁸ Jena Derakhshani Hamadani,⁹ Charlotte Hanlon ,^{10,11} Simone M Karam,¹² Betsy Lozoff,¹³ Lisy Ratsifandrihamanana,¹⁴ Linda Richter ,¹⁵ Maureen M Black ,^{16,17} Global Child Development Group collaborators

To cite: Weber AM, Rubio-Codina M, Walker SP, *et al*. The D-score: a metric for interpreting the early development of infants and toddlers across global settings. *BMJ Global Health* 2019;4:e001724. doi:10.1136/bmjgh-2019-001724

Handling editor Seye Abimbola

► Additional material is published online only. To view please visit the journal online (<http://dx.doi.org/10.1136/bmjgh-2019-001724>).

Received 20 May 2019
Revised 28 August 2019
Accepted 30 August 2019



© Author(s) (or their employer(s)) 2019. Re-use permitted under CC BY. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to
Dr Ann M Weber;
annweber@unr.edu

ABSTRACT

Introduction Early childhood development can be described by an underlying latent construct. Global comparisons of children's development are hindered by the lack of a validated metric that is comparable across cultures and contexts, especially for children under age 3 years. We constructed and validated a new metric, the Developmental Score (D-score), using existing data from 16 longitudinal studies.

Methods Studies had item-level developmental assessment data for children 0–48 months and longitudinal outcomes at ages >4–18 years, including measures of IQ and receptive vocabulary. Existing data from 11 low-income, middle-income and high-income countries were merged for >36 000 children. Item mapping produced 95 'equate groups' of same-skill items across 12 different assessment instruments. A statistical model was built using the Rasch model with item difficulties constrained to be equal in a subset of equate groups, linking instruments to a common scale, the D-score, a continuous metric with interval-scale properties. D-score-for-age z-scores (DAZ) were evaluated for discriminant, concurrent and predictive validity to outcomes in middle childhood to adolescence.

Results Concurrent validity of DAZ with original instruments was strong (average $r=0.71$), with few exceptions. In approximately 70% of data rounds collected across studies, DAZ discriminated between children above/below cut-points for low birth weight (<2500 g) and stunting (–2 SD below median height-for-age). DAZ increased significantly with maternal education in 55% of data rounds. Predictive correlations of DAZ with outcomes obtained 2–16 years later were generally between 0.20 and 0.40. Correlations equalled or exceeded those obtained with original instruments despite using an average of 55% fewer items to estimate the D-score.

Conclusion The D-score metric enables quantitative comparisons of early childhood development across ages and sets the stage for creating simple, low-cost, global-use instruments to facilitate valid cross-national comparisons of early childhood development.

Summary

What is already known?

- Theories of infant development and empirical evidence support both a universal biological unfolding of stage-based skills as well as individual differences due to genetic, environmental and cultural influences.
- Despite the availability of multiple measures, a common and easily interpretable metric does not exist for making valid international comparisons of children's development from birth to 3 years.

What are the new findings?

- Existing data from 16 longitudinal studies and 11 countries were mathematically linked with an innovative statistical model to construct a common metric, the Developmental Score (D-score), that represents a latent construct for early childhood development.
- The D-score, estimated with an average of 55% fewer items than the original instruments, demonstrated discriminant and concurrent validity and was predictive of outcomes during middle childhood through adolescence.

What do the findings imply?

- The D-score's interval-scale property, with a common unit of measurement across ages, allows for the depiction of a developmental trajectory with increasing age, which can be interpreted similarly to growth trajectories for height and weight.
- The statistical model enables both the estimation of D-scores for existing datasets and the derivation of new instruments, which will allow for valid international comparisons and future construction of global standards for the development of children 0–3 years.

INTRODUCTION

Theories of infant development support both a universal biological unfolding of stage-based skills as well as individual differences due to

varying genetic, environmental and cultural influences.^{1,2} Empirical evidence validates these theories by demonstrating that, on average, infants and toddlers achieve major neurodevelopmental milestones in a consistent and ordinal pattern during the first few years of life, regardless of country of origin, while demonstrating within country variability conditional on parental and household disparities.^{3,4} Therefore, from both theoretical and empirical perspectives, childhood development in the first years of life can be described by an underlying latent construct that is relatively invariant across countries, progresses in a predictable sequence and represents domains of motor, language, cognitive and personal-social development. However, we lack a valid and easily interpretable metric that represents a latent construct of early childhood development and would enable global comparisons of child development, just as growth trajectories for height and weight facilitate global comparisons of children's nutritional status.

There is a long history of testing the emergent developmental skills of infants and toddlers, through direct observation, child's response to specific tasks and situations, or by caregiver report. As a result, multiple assessment instruments incorporating similar tasks have been developed, most of which are standardised for high-income country populations.⁵ Although some have been used globally, instruments adapted in one setting may not measure the same construct as originally designed, or as adapted in other settings, and may not perform equivalently across countries. As such, global comparisons of scores obtained from adapted instruments may be misleading.

Our goal was to develop and evaluate a metric representing a universal latent construct of early childhood development by leveraging existing data from 16 longitudinal cohorts from 11 countries, gathered using 12 existing instruments. In this paper, we describe the construction of a statistical model using these data to produce the Developmental Score (D-score), an interval-scale metric to express children's development with a common numerical unit. The D-score facilitates interpretation of children's abilities across different ages (just as centimetres are used for height), and an age-standardised D-score enables comparisons of children's development both within and between countries.⁶ We examine discriminant, concurrent and predictive validity of model-derived D-scores for children living in diverse cultural settings. We conclude with a discussion of how the validated D-score metric and model can be used to convert existing data from disparate settings to a common metric and to construct new instruments for global use.

METHODS

Country and study cohorts

Longitudinal data from 16 cohorts of children ($n > 36\,000$) in 11 countries were previously collected as birth cohort studies (Brazil 1 and Brazil 2,^{7,8} Chile 2,⁹ Ethiopia,^{10,11} Netherlands 2,¹² and South Africa¹³), instrument

validation studies (the Netherlands 1¹⁴ and Colombia 2¹⁵), and programme evaluations focused on low-income or undernourished children (Bangladesh,¹⁶ Chile 1,¹⁷ China,¹⁸ Colombia 1,¹⁹ Ecuador,²⁰ Jamaica 1²¹ and Jamaica 2,²² and Madagascar²³). Brazil 2 (Pelotas), Netherlands 2 (Maastricht) and South Africa (Johannesburg-Soweto) cohorts were representative of a city in each country; the Colombia 2 cohort was representative of low-income and low-middle-income groups in Bogota; the Ethiopia cohort was representative of a rural district; and the Chile 2 cohort was representative of the country. Although initially representative of the city of Pelotas, the Brazil 1 data included here were obtained from all low birth weight (< 2500 g) children in the cohort and a systematic sample of the remaining cohort members. An Advisory Board was formed that included an investigator from each study with in-depth knowledge about the local context and data collection. Information on these studies, validity and primary analyses were published previously.⁶⁻²¹ Table 1 shows an overview of the cohorts.

Study cohorts were purposively selected for inclusion in this study if children were assessed with a direct assessment instrument at least once during early childhood (< 48 months, Time 1) and again during middle childhood to adolescence (Time 2) when they were ages $> 4-18$ years. Availability of item-level assessment data at Time 1 was also a requirement. We included item-level data for children ≥ 36 months at Time 1 to ensure items were included that high-performing 3-year-old children would fail. In some cohorts, multiple instruments were used and/or multiple rounds of data were collected at Time 1 (eg, at 6, 12, 18, and 24 months) and Time 2. All data from Time 1 were included in the D-score model building process (see below). Available data for children 48-58 months at Time 1 were excluded from validity tests as our aim was to create a metric for young children that would be predictive of their skills at ages over 4 years.

Item instrument mapping

Instruments used in each study were internationally recognised and locally adapted for assessing development of young children using multiple items (table 1 and online supplementary appendix table A1 with associated references). Instruments were primarily direct assessment, with two caregiver-report instruments (Ages and Stages Questionnaire or ASQ and Vineland Social Maturity Scale). A list of instruments and corresponding citations are provided in the supplementary materials. Although published separately, these instruments incorporate many similar items designed to assess the same developmental skills, a critical feature required for linking across disparate datasets.

Advisory Board members created a master spreadsheet of > 1500 items administered with instruments at Time 1, organised across five developmental domains: fine motor, gross motor, receptive language, expressive language, and cognition. Personal-social development was not included, as measures of this domain were inconsistently

Table 1 Study cohorts: sample sizes, age range and instruments by time of measurement

	Time 1: Early childhood				Time 2: Middle childhood				Time 2: Adolescence			
	Round	N	Age (months)	Instrument*	Age (years;months)	Instrument†	Age (years;months)	Instrument†	Age (years;months)	Instrument†	Type of study	
Bangladesh ¹⁶	1	1862	18	Bayley-II	5;3–5;8	WPPSI	18;6	WAIS			Program evaluation	
Brazil ¹⁷	1‡	644	3	Denver-II							Birth cohort	
	2	1412	6	Denver-II								
	3	1362	12	Denver-II								
Brazil ²⁸	1‡	3907	12	BDI-2							Birth cohort	
	2‡	3869	24	BDI-2								
Chile ¹⁷	1	128	6	Bayley-I	5;6–5;8	WPPSI					Program evaluation	
	2	1732	12	Bayley-I								
	3	279	18	Bayley-I								
Chile ²⁹	1‡	4869	7–23	BDI-2							Birth cohort	
	1§	9201	24–58	Tepsi	4;1–6;2	TVIP					Program evaluation	
China ¹⁸	1	990	18	Bayley-III							Program evaluation	
Colombia ^{119,42}	1	704	12–24	Bayley-III	4;6–5;8	TVIP					Program evaluation	
	2	631	24–42	Bayley-III								
Colombia ²¹⁵	1	1311	6–42	Bayley-III	6;0–8;11	WISC-V, TVIP					Instrument validation	
	1	658	6–42	Denver-II								
	1	658	6–42	ASQ-3								
	1‡	635	6–42	BDI-2 Screener								
Ecuador ²⁰	1	667	0–35	Barrera	5;7–8;8	TVIP	9;2–12;2	TVIP			Program evaluation	
Ethiopia ^{10,11}	1	193	12	Bayley-III			9;10–10;10	PPVT			Birth cohort	
	2	440	30	Bayley-III								
	3	456	42	Bayley-III								
Jamaica ¹²¹	1	225	15	Griffiths	6;6–7;1	WPPSI, PPVT					Program evaluation	
	2	218	24	Griffiths								
Jamaica ²²²	1	159	9–24	Griffiths	7;0–8;3	SB-4, Raven's, PPVT	16;9–18;0	WAIS			Program evaluation	
	2	159	21–36	Griffiths								
	3	159	33–48	Griffiths								
Madagascar ²³	1	205	34–42	SB-5	6;10–7;10	SB-5, PPVT					Program evaluation	

Continued

Table 1 Continued

	Time 1: Early childhood			Time 2: Middle childhood			Time 2: Adolescence		
	Round	N	Age (months)	Instrument*	Age (years;months)	Instrument†	Age (years;months)	Instrument‡	Type of study
Netherlands 1 ¹⁴	1	1985	1	DDI	4;7-6;1	UKKI			Instrument validation
	2	1807	2	DDI					
	3	1963	3	DDI					
	4	1919	6	DDI					
	5	1881	9	DDI					
	6	1802	12	DDI					
	7	1776	15	DDI					
	8	1787	18	DDI					
	9	1815	24	DDI					
Netherlands 2 ¹²	1	1016	24	DDI				Birth cohort	
	2	995	30	DDI					
	3	1592	36	DDI					
	4	1534	42	DDI					
	5	1034	48	DDI					
South Africa ¹³	1	485	6	Bayley-I, Griffiths	4;10-5;6	Denver-II		Birth cohort	
	2	275	12	Bayley-I, Griffiths	7;0-8;6	Raven's (Coloured)			
	3	1802	24	Vineland					
	4†§	1614	48	Vineland					

*Early childhood testing instruments were the ASQ-3; Barrera; BDI-2 and BDI-2 Screener; Bayley-I, II and III; Denver-II, DDI; Griffiths; SB-5; Tepsi; and the Vineland. References for instruments are included with online supplementary appendix table A1.

†Middle childhood and adolescence testing instruments were the Denver-II; PPVT and the Spanish version, TVIP; Raven's and Raven's (Coloured); SB-4 and SB-5; UKKI; WAIS; WISC-V; and WPPSI. References for instruments are included with online supplementary appendix table A1.

‡These rounds of data that were not used in the final 565-item D-score model.
§Excludes children >48 months at Time 1 from validity tests.

ASQ-3; Ages & Stages Questionnaires; Barrera; Barrera Moncada; Bayley I, II and III, Bayley Scales for Infant and Toddler Development; BDI-2, Battelle Development Inventory and Screener-2; DDI, Van Wiechenschema, referred to as the Dutch Developmental Instrument; Denver-II, Denver Developmental Screening Test; D-Score, Developmental Score; Griffiths, Griffiths Mental Development Scales; PPVT, Peabody Picture Vocabulary Test; Raven's, Raven's Progressive Matrices; SB-4 and SB-5, Stanford Binet Intelligence Scales; Tepsi, Test de Desarrollio Psicomotor; TVIP, Test de Vocabulario en Imagenes Peabody; UKKI, Utrechtse Korte Kleuter Intelligentietest; Vineland, Vineland Social Maturity Scale; WAIS, Wechsler Adult Intelligence Scale; WISC-V, Wechsler Intelligence Scale for Children - Revised; WPPSI, Wechsler Preschool and Primary Scale of Intelligence.

used across the cohorts. Although early personal-social development facilitates rich child-caregiver interactions, the expression and interpretation of personal-social development vary across cultures.²⁴

Within each of the five domains, individual items from each instrument (eg, Denver Developmental Screening Test or Griffiths Mental Development Scales) were mapped to same-skill items in the Bayley Scales of Infant and Toddler Development, third edition (Bayley-III), which was the most frequently administered instrument. Equivalency of skills between items was determined by referring to manuals, item descriptions and extensive hands-on testing experience by Board members. We also mapped groups of same-skill items across other instruments that did not map onto Bayley-III items. Caregiver-report items were mapped to direct assessment items if the skill assessed was considered equivalent. The mapping exercise resulted in 95 groups of items from different scales measuring the same skill termed ‘equate groups’, each containing at least two same-skill items from different instruments (eg, item ‘stacks 2 cubes’ in Instrument A=item ‘builds 2 block tower’ in Instrument B).

Data harmonisation

The master spreadsheet of Time 1 items formed the basis for combining the data from the 16 cohorts into a single database, with equate groups identified to link items across instruments and cohorts. All items were coded as 0 (fail), 1 (pass) or missing. In the Battelle Developmental Inventory, items were originally scored as 0 (fail) with passing scores of 1 or 2 depending on the level of skill demonstrated or time taken to complete the task. For all Battelle items, 2 was recoded as 1. For six Battelle items, a score of 1 was recoded as 0 because these items were mapped to Bayley-III items that were more difficult. Similarly, ASQ items were originally scored as 0 (not yet), 5 (sometimes) and 10 (succeeds); both 5 and 10 were recoded as 1. Harmonisation resulted in a matrix with 71 403 rows (child-round observations) and 1572 columns (items) collected from 36 345 unique children. Since each cohort and round of data collection yielded information on a subset of items, by design, the matrix included many empty cells.

Model building and active equate groups

A unidimensional statistical model for the D-score was built using the Rasch model, a simple logistic model for which an observed response is a function of the difference between person ability and item difficulty.²⁵ In the Rasch model, when a person’s ability is equal to the item difficulty, there is a 50–50 chance of passing that item. Probability of passing is above 50% when ability is greater than the item difficulty and below 50% when ability is lower than the item difficulty. To convert scores from different instruments to a common scale, we applied psychometric equating methods typically used in educational testing.²⁶ Instrument equating in our application required the identification of equate groups with

comparable psychometric performance for all items in the group, across instruments and cohort origins, such that group items could be statistically constrained to have the same difficulty.²⁷ We defined these equate groups as ‘active’ as they mathematically bridge instruments and cohorts, linking them to a common scale. Children with the same underlying developmental ability should have the same probability of passing active equate items.

Model building was a multistep process. We removed 233 items with fewer than 10 observations in the least populated response category (ie, pass or fail), leaving 1339 items. Next, we evaluated progressively refined Rasch models that varied along two dimensions: (1) the subset of active equate groups and (2) the cut-points for item fit statistics (ie, residual (outfit) and weighted (infit) mean square fit) used to exclude poorly fitting items from the model. To optimise measurement properties, we limited activation of equate groups to those that performed very well across instruments and cohorts, rather than activating ones with variable performance. Also, we sought active equate groups representative of the five developmental domains of interest and of abilities of children across the age range 0–47 months. Finally, we aimed for active equate groups to connect instruments by at least three items.

In the final Rasch model, items were retained if they were part of an active equate group or included as independent items if both their infit and outfit statistic were <1. Items from equate groups that were not activated (ie, passive equate groups) were not constrained to the same difficulty and treated as independent items. Independent items from a single instrument administered in more than one country were statistically constrained to a single difficulty (eg, Bayley-III items administered in China, Colombia and Ethiopia) if children with the same latent ability (but not necessarily of the same age) were found to have the same probability of passing these items regardless of country of origin. By constraining the difficulty of same-instrument items in the model, we gain additional links to the common scale for cohorts from different countries who were administered the same instrument. Independent items also improve the precision of estimated D-scores. Items with poor fit to the Rasch model or demonstrating differential performance by country were excluded from the final model and analyses of validity.

D-score and DAZ estimates

For each child at each round, a D-score was estimated from the final model by the expected a posteriori (EAP) method.²⁸ To establish the numerical range for the scale, we anchored the D-score relative to two indicators that are used widely in different instruments and are easy to measure and minimally sensitive to cultural variation: ‘lifts head to 45 degrees in prone position’ and ‘sits in stable position without support’. Fixed item difficulties of 20 and 40 D-score units were used for these items, respectively, based on previous analyses of the Netherlands 1 cohort data.²⁹ These values were chosen such that D-scores start near

zero at age 1 month. In the first year of life, a one D-score unit increase corresponds to approximately 1 week difference in age. In the second year of life, a one unit increase corresponds to approximately 1 month. Regardless of age, a 10-unit increase in the D-score corresponds to a change from children being very likely to fail (>90%) an item to very likely to pass (>90%).

We modelled the age-conditional distribution of D-scores across country cohorts with the Lambda-Mu-Sigma (LMS) method,³⁰ an accepted approach for fitting growth curves, to generate a D-score-for-age z-score (DAZ) for each child.

Validation

DAZ estimates for children aged <48 months at Time 1 were used to examine discriminant, concurrent and predictive validity of the D-score metric. There were 35 rounds of data collection across the 16 cohorts (referred to henceforth as data rounds). Discriminant validity was examined by comparing mean DAZ by three predictors of early child development: low birth weight (<2.5 kg), stunting (height-for-age < -2 SD of median WHO Growth Standards for same age and sex children)³² and maternal education (no education, any primary, any secondary and above secondary education). The maternal education classification chosen could be consistently applied across the available studies, with categories in some cohorts having small samples. Household wealth was captured in all studies, but wealth was estimated in ways that were not comparable across settings. Other predictors of early development, such as gestational age and nurturing care indicators,³³ were considered but were not generally available across studies. We used t-tests for low birth weight and stunting, and analysis of variance F-tests for maternal education, to evaluate whether DAZ was sensitive to differences in child ability across these established risk/protective factors for early childhood development,³¹ with significance set at $p < 0.05$. Scores for categories with fewer than 10 observations were excluded from tests of significance.

For concurrent validity, we calculated pairwise Pearson correlations of DAZ with age-standardised scores for the original instruments. When available, we used standardised scores based on external standards. Otherwise, we generated age-adjusted z-scores for a given cohort (internal standardisation) using non-parametric methods.¹⁵ For the Netherlands 1 cohort, the 9 rounds of data collection were collapsed into three 12-month age intervals, which did not change results.

For predictive validity, we correlated DAZ by data collection round at Time 1 with standardised test scores acquired at Time 2 in middle childhood (>4–9 years) and adolescence (>9–18 years). Time 2 data were included in prediction analyses if ≥ 2 years had passed since Time 1 data collection. Because initial age of testing affects prediction of later outcomes,^{34 35} cross-sectional data covering a wide age range at Time 1 in Chile 2, Colombia 2 and Ecuador were grouped into 12-month age intervals.

The data collection rounds of the Netherlands 1 cohort were collapsed as explained above. Although originally planned for the analysis, China Time 2 data were not ready to be shared for this project. Time 2 assessments (see table 1) include tests of IQ (eg, Wechsler Preschool and Primary Scale of Intelligence), matrix reasoning (Raven's Coloured Progressive Matrices) and receptive vocabulary (Peabody Picture Vocabulary Test).

For both concurrent and predictive validity, we classified correlations as low ($r = 0.20$ – 0.39), moderate ($r = 0.40$ – 0.59), strong ($r = 0.60$ – 0.79) or very strong ($r = 0.80$ – 1).³⁶

Software

All model fitting and evaluation of items and equate groups was done with R. We extended the function `sirt::rasch.pairwise.itemcluster`³⁷ with an option to constrain the solution by equate groups. See Eekhout, Weber and van Buuren (under review)²⁷ for more details. Tests of validation were performed in R or Stata V 14.

Role of the funding source

The Bill and Melinda Gates Foundation (BMGF) approved the study design as part of funding approval, but had no role in data collection, analysis, interpretation or write-up of results or in the decision to submit the paper for publication.

Ethical considerations

The study involved secondary data analyses of deidentified data. Investigators signed a data sharing agreement stating that they had approval to use these data for this project from study collaborators and/or institutions. Approval for the secondary analyses was obtained from the Netherlands Organization for Applied Research (TNO) and the ethical review board at Stanford University.

Patient and public involvement

This non-clinical research was performed using deidentified data from completed studies without patient or public involvement. No new participants were recruited and no new data were collected.

RESULTS

The final model used to estimate D-scores contained 565 items originating from 11 instruments and included 18 active equate groups. The number of items administered for any given child varied considerably across and within cohorts, with an overall average of 27 items per child used to estimate their D-score (country per child averages ranged from 3.5 items in Ecuador to 59.5 items in Bogota where multiple instruments were used, see online supplementary Table A2). Items from the Battelle Developmental Inventory performed poorly in the model and were removed from further analysis, resulting in the loss of one cohort (Brazil 2) as well as Battelle data from Colombia 2 and Chile 2.

The plots in figure 1 show the distribution of D-score estimates by age and cohort applying the final model to all

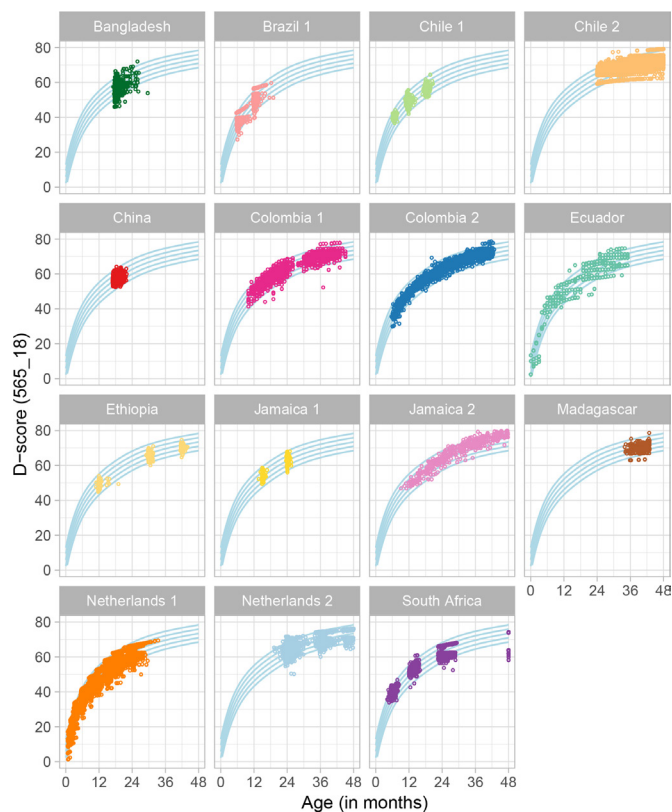


Figure 1 Distribution of the D-score by age and cohort with the final model (565 items and 18 equate groups). D-score, Developmental Score.

rounds of Time 1 data. The blue curved lines represent the age-conditional distribution of the combined dataset for all cohorts and are driven by the large Colombia 2 and Chile 2 samples. Average D-score trajectories from the Netherlands 1 and Colombia cohorts follow the age-conditional distribution of the combined dataset across a ≥ 2 year age range. Distributions of scores in the other cohorts reflect study sampling and data availability, but generally fall within the age-specific percentiles developed for the full dataset. For example, the China cohort was assessed at 18 months at Time 1 such that all D-scores are grouped together around that age. In contrast, the Ethiopia cohort was assessed at 12, 30 and 42 months and the plot shows three groupings of scores that increase on average with age.

Discriminant validity

The overall mean DAZ for all cohorts combined is 0 and the range is from -7 to $+4.5$ SD units. The mean DAZ and SDs by birth weight, stunting and maternal education are shown in table 2 for cohort rounds with available data. Children above the low birth weight cut-point demonstrated significantly higher mean DAZ scores than those below the cut-point in 18 of 26 data rounds (69%) with available data. Non-stunted children had significantly higher scores than stunted children in 21 of 28 (75%) data rounds. DAZ scores increased significantly with maternal education in 17 of 31 (55%) data rounds. In another six data rounds, mean DAZ increased with maternal education, but differences were not statistically significant.

Concurrent validity

Moderate to strong concurrent validity was anticipated as the D-score is computed from subsets of items from original instruments (table 3). The proportion of items from the original instrument used to estimate the D-score for each cohort averaged 0.61 and ranged from 0.13 to 1.0. The average concurrent validity of the DAZ with standardised scores from the original instruments was strong ($r=0.71$), ranging from 0.24 to 0.96. Results were robust to the use of externally and internally standardised scores in the Colombia 1 and 2 cohorts, which allowed for both methods of standardisation (not shown).

Predictive validity

The figure 2A and B presents predictive validity to measures of IQ and receptive vocabulary in middle childhood (Time 2 for ages >4 –9 years) and adolescence (Time 2 for ages >9 –18 years), respectively, of both DAZ and the original instruments. When multiple scores were available for original instruments, we included the cognitive score (eg, over language or motor) or the Bayley-III score (eg, over the Denver-II in Colombia 2). Detailed tables are included in the online supplementary appendix, table A3a for DAZ and online supplementary appendix, table A3b for scores obtained from original instruments.

The average predictive correlation of the DAZ with IQ and receptive vocabulary scores in middle childhood was 0.29 (range 0.07–0.54) and 0.31 (range 0.008–0.54), respectively. Predicting to adolescence, the average correlation of the DAZ with IQ and receptive vocabulary was 0.37 (range 0.17–0.56) and 0.14 (range 0.07–0.23), respectively. The DAZ performed, as well as occasionally outperformed, single dimension scores from the original instruments. For example, in Colombia 1, the correlation of the 10–26 month DAZ with later receptive language (0.368) was comparable or slightly larger than correlations of age-standardised cognition, language and motor subscale scores from the Bayley-III with the later measure (0.277, 0.322 and 0.278, respectively). In Brazil 1, correlations of age-standardised cognitive and language subscale scores from the Denver-II with IQ at age 18 years were only 0.051 and 0.127, whereas correlations for the DAZ and the composite measure from the Denver-II were similar (0.187 and 0.189). In general, the correlation of the DAZ with Time 2 measures increased with age at Time 1 within a given cohort. However, the age trend was not consistent across cohorts.

Simulation of a new instrument

Although the number of items administered to each child varied considerably between cohorts, children's D-score was estimated from an average of 55% fewer items than the average 55 items used in the original instruments. For example, in the Jamaica 1 cohort, information was available for, on average, 87 items per child, and yet the D-score was calculated, on average, from 43 items per child. The ability to reduce the items needed to estimate the D-score suggests the feasibility of creating a relatively short instrument for future field work. We simulated this by obtaining

Table 2 Discriminant validity of DAZ with birth weight, nutritional status and maternal education

Country	Round	Nutritional status†				Maternal education				P value
		Low	Normal	Stunted	Non-stunted	No education	Any primary	Any secondary	Above secondary	
Bangladesh	1	-0.71 (0.04) 510	-0.51 (0.03) 1317	-0.73 (0.03) 820	-0.42 (0.03) 1007	-0.73 (0.04) 625	-0.67 (0.05) 397	-0.41 (0.04) 783	0.75 (0.25) 22	<0.001
		0.23 (0.86) 415	0.73 (0.79) 994	0.09 (0.9) 218	0.67 (0.8) 1194	0.26 (0.77) 35	0.55 (0.85) 1031	0.63 (0.81) 242	*	0.041
		0.37 (0.88) 401	0.87 (0.87) 959	0.18 (1.01) 184	0.81 (0.85) 1177	0.2 (0.91) 32	0.69 (0.87) 988	0.79 (1) 240	*	0.002
Chile 1	1	*	0.51 (0.46) 128	*	0.5 (0.47) 126	*	0.57 (0.4) 48	0.48 (0.51) 69	0.42 (0.42) 10	0.511
		*	0.16 (0.53) 1611	0.01 (0.46) 12	0.16 (0.53) 1409	*	0.14 (0.49) 601	0.16 (0.55) 838	0.26 (0.52) 163	0.037
		*	-0.27 (0.7) 278	N/A	-0.27 (0.7) 225	N/A	-0.36 (0.68) 109	-0.21 (0.72) 147	-0.15 (0.65) 21	0.212
Chile 2	1	-0.12 (0.99) 327	-0.02 (0.91) 8071	-0.14 (0.99) 292	-0.03 (0.91) 8383	-0.01 (0.99) 100	-0.35 (0.92) 1717	-0.05 (0.89) 5245	0.24 (0.90) 2139	<0.001
		0.06 (0.56) 9	-0.08 (0.58) 968	-0.2 (0.47) 29	-0.08 (0.59) 943	*	-0.31 (0.57) 34	-0.1 (0.56) 742	0.06 (0.64) 178	<0.001
Colombia 1	1	-0.11 (0.78) 47	0.29 (0.96) 593	0.001 (1.07) 99	0.28 (0.92) 591	-0.23 (0.92) 27	0.1 (0.96) 238	0.33 (0.93) 375	*	<0.001
		0.09 (0.93) 43	0.2 (0.93) 535	-0.18 (1.31) 59	0.21 (0.86) 569	-0.29 (1.01) 21	-0.06 (0.93) 217	0.29 (0.87) 337	*	<0.001
Colombia 2	1	-0.1 (0.77) 128	0.14 (0.78) 1062	-0.03 (0.78) 229	0.13 (0.8) 1080	*	-0.23 (0.82) 140	0.06 (0.78) 744	0.32 (0.78) 393	<0.001
		0.5 (1.05) 12	0.38 (1.48) 133	-0.17 (1.45) 89	0.44 (1.49) 293	-0.27 (2.04) 14	0.19 (1.45) 433	0.65 (1.32) 197	0.70 (1.56) 11	<0.001

Continued

Table 2 Continued

Country	Birth weight†	Nutritional status†					Maternal education					
		Round	Low	Normal	Stunted	Non-stunted	P value	No education	Any primary	Any secondary	Above secondary	P value
Ethiopia	1	-0.26 (0.45) 16	-0.08 (0.6) 174	0.152	-0.13 (0.64) 84	-0.08 (0.56) 103	0.583	-0.07 (0.6) 168	-0.29 (0.5) 22	*	*	0.102
	2	-0.59 (0.78) 33	-0.34 (0.56) 407	0.088	-0.4 (0.58) 328	-0.27 (0.59) 111	0.042	-0.39 (0.6) 385	-0.19 (0.46) 54	*	*	0.017
	3	-0.8 (0.5) 35	-0.58 (0.46) 421	0.017	-0.65 (0.47) 345	-0.44 (0.42) 109	<0.001	-0.62 (0.46) 402	-0.45 (0.47) 53	*	*	0.016
Jamaica 1	1	0.04 (0.53) 131	0.22 (0.49) 94	0.01	-0.17 (0.45) 14	0.14 (0.52) 210	0.026	*	*	0.12 (0.5) 216	*	N/A
	2	0.35 (0.78) 130	0.55 (0.68) 88	0.044	*	0.44 (0.74) 215	N/A	*	*	0.46 (0.71) 208	*	N/A
Jamaica 2	1	ND	ND	ND	-0.23 (0.76) 122	0.55 (0.57) 37	<0.001	*	-0.03 (0.77) 138	-0.15 (0.88) 21	*	0.54
	2	ND	ND	ND	0.78 (1.02) 62	0.97 (0.87) 97	0.213	*	0.93 (0.96) 138	0.73 (0.75) 21	*	0.36
	3	ND	ND	ND	1.61 (0.99) 48	1.86 (0.98) 111	0.141	*	1.81 (0.98) 138	1.66 (1.08) 21	*	0.52
Madagascar	1	ND	ND	ND	-0.39 (0.94) 113	-0.23 (0.9) 89	0.205	-0.53 (0.84) 51	-0.35 (0.99) 118	-0.03 (0.89) 35	*	0.054

Continued

Table 2 Continued

Country	Birth weight†	Nutritional status‡					Maternal education																
		Round	Low	Normal	Stunted	Non-stunted	P value	No education	Any primary	Any secondary	Above secondary	P value											
Netherlands 1	1	-1.02	(1.02)	-0.23	(1.02)	-0.87	(1.04)	96	-0.17	(1)	1264	<0.001	*	-0.22	(1.01)	602	-0.28	(1.05)	1008	-0.32	(1.06)	328	0.335
		-0.48	(1.18)	0.06	(0.9)	-0.42	(1.32)	75	0.07	(0.92)	1149	0.003	*	0.02	(0.94)	520	0.08	(0.88)	901	-0.03	(0.99)	299	0.157
		-0.83	(1.1)	-0.03	(0.9)	-0.46	(1.07)	120	-0.05	(0.9)	1793	<0.001	*	-0.09	(0.94)	593	-0.1	(0.92)	994	0.02	(0.94)	325	0.156
		-0.59	(1.08)	-0.05	(0.92)	-0.41	(0.93)	49	-0.07	(0.93)	1830	0.011	*	-0.15	(0.97)	579	-0.09	(0.94)	973	0.1	(0.84)	320	<0.001
		-1.01	(1.1)	-0.25	(0.95)	-0.83	(1.33)	43	-0.28	(0.96)	1804	0.011	*	-0.3	(1.03)	566	-0.34	(0.97)	965	-0.14	(0.91)	306	0.006
		-0.92	(1.11)	-0.37	(1.07)	-0.99	(1.5)	30	-0.39	(1.06)	1739	0.002	*	-0.43	(1.12)	540	-0.42	(1.06)	924	-0.28	(1.03)	292	0.104
		-0.77	(1.14)	-0.37	(1.06)	-1	(1.7)	33	-0.38	(1.05)	1706	0.001	*	-0.47	(1.11)	528	-0.39	(1.04)	918	-0.3	(1.06)	286	0.079
		-0.68	(1.07)	-0.22	(1.1)	*			-0.4	(1.09)	40	N/A	*	-0.36	(1.09)	508	-0.24	(1.11)	871	-0.05	(1.1)	269	<0.001
		-0.25	(1.25)	-0.02	(1.12)	-1.15	(1.84)	16	-0.01	(1.12)	1746	0.026	*	-0.3	(1.16)	557	0.04	(1.1)	925	0.32	(1.04)	293	<0.001
South Africa	1	0.69	(0.55)	1.01	(0.64)	ND	ND	ND	ND	ND	<0.001	*	0.98	(0.74)	55	1.01	(0.62)	376	0.71	(0.6)	41	0.020	
		0.48	(0.82)	0.5	(0.7)	0.06	(1.02)	12	0.56	(0.71)	0.026	*	0.33	(0.59)	33	0.49	(0.74)	206	0.71	(0.52)	32	0.096	
		-0.24	(1.04)	-0.02	(1.03)	-0.11	(1.06)	299	0.08	(1.02)	1020	0.006	-0.02	(1.01)	224	-0.05	(1.02)	1337	0.08	(1.01)	215	0.229	

* <10 observations, N/A=Not applicable, ND=No data available. Differences significant <0.05 are in bold.
 † Low birth weight is defined as <2.5 kg and stunted is height-for-age z-score of <-2 SD of the median WHO Growth Standards for same-age and same-sex children.³²
 ‡ Subgroup data are mean (SD) n.

Table 3 Concurrent correlation of DAZ in children under 48 months with measures from original instruments

Cohort	Age range (months)	% items in D-score from original instrument*	Bayley-I, II, III†			Other measures		
			Cognition		Language	Motor	Total Score	Measure
			MDI	PSDI	PDI			
Bangladesh	18	0.13 (16/121)	0.797	0.503				
Brazil 1	5–11 11–19	1 (18/18) 0.84 (16/19)					0.859 0.926	Denver-II
Chile 1	6 12 18	0.67 (31/46) 0.55 (60/109) 0.45 (26/58)	0.861 0.880 0.835	0.438 0.361 0.249				
Chile 2‡	24–35 36–47	0.53 (33/62)					0.768 0.855	Tepsi
China	18	0.53 (40/76)	0.541	–	0.458			
Colombia 1	10–26 28–45	0.41 (104/254) 0.45 (96/212)	0.710 0.742	0.809 0.840	0.775 0.672			
Colombia 2	6–17 18–29 30–42	0.32 (200/631)	0.386 0.671 0.649	0.333 0.837 0.811	0.675 0.651 0.620		0.758 0.642 0.795	Denver-II
Ecuador	0–11 12–23 24–35	0.68 (15/22)					0.791 0.815 0.768	Barrera
Ethiopia	11–12 29–32 41–44	0.48 (73/151) 0.46 (83/181) 0.42 (61/146)	0.614 0.737 0.631	0.560 0.814 0.723	0.915 0.808 0.696			
Jamaica 1	15 24	0.47 (69/148) 0.44 (68/155)					0.930 0.862	Griffiths DOQ
Jamaica 2	9–25 21–37 33–48	0.44 (94/212) 0.33 (66/200) 0.29 (52/181)					0.574 0.888 0.864	Griffiths DOQ
Madagascar	34–42	0.24 (10/41)					0.452	SB-5
Netherlands 1¶	0–11 12–23 24–34	0.57 (4/7) to 1 (13/13) 0.92 (12/13) 0.71 (10/14)					0.949 0.958 0.486	DDI

Continued

Table 3 Continued

Cohort	Age range (months)	% items in D-score from original instrument*	Bayley-I, II, III†			Other measures	
			Cognition	Language	Motor	Total Score	Measure
South Africa	6	0.48 (80/166)	0.791	0.775	0.868	Griffiths DQ§	
	12	0.5 (101/202)	0.763	0.659	0.725		
	24	0.41 (12/29)			0.729	Vineland	

*Not all items from the original instrument were included in the final model. Items from multiple instruments were included in D-score estimation in Colombia 2 and South Africa.

†Bayley-I, II, III scores are externally standardised, except for Ethiopia, which is internally age-standardised. Externally standardised scores are cognitive, language and motor composite scores for the Bayley-III; and the MDI and PDI for the Bayley-I and Bayley-II.

‡Excludes children >48 months at Time 1.

§Griffiths DQ includes motor development items.

¶Data collection rounds for Netherlands 1 were collapsed into 1 year age increments (0–11 m, 12–23 m, 24–34 m). Range of % items from original instrument used in D-score varies by round, but concurrent correlation is ≥0.9 for all collapsed rounds. 32 months was the maximum age in completed months for the test of predictive validity.

Barraera, Barrera Moncada; Bayley-I, II, III, Bayley Scales for Infant and Toddler Development; DDI, Van Wiechenschema, referred to as the Dutch Developmental Instrument; Denver-II, Denver Developmental Screening Test; DQ, Developmental Quotient; D-Score, Developmental Score; Griffiths, Griffiths Mental Development Scales; MDI, Mental Development Index; PDI, Psychomotor Development Index; SB-5, Stanford Binet Intelligence Scales; Tepsi, Test de Desarrollo Psicomotor; Vineland, Vineland Social Maturity Scale.

estimated D-scores on a subset of items included in the final D-score model. First, final model items were sorted by age equivalence (ages at which 10% pass, 50% pass and 90% of children pass each item) and reviewed by Advisory Board members to retain items that were non-duplicative of a skill, easy to train and administer, feasible for use in the field, and likely to demonstrate cross-cultural validity. The subset of 165 items comprised approximately 20–25 items per 6-month age group. The simulation showed that D-score estimates from this subset were very strongly correlated (r=0.999) with the full 565-item model.

DISCUSSION

The development of the D-score was driven by the need for a valid and easily interpretable metric for an underlying latent construct of infant and toddler development that is comparable across cultures and contexts. A statistical model for the D-score was constructed that mathematically bridges data from multiple internationally recognised and commonly used instruments, using a set of linking items that performed equivalently across countries and cohorts. By leveraging existing longitudinal data for >36 000 children from 11 low-income, middle-income and high-income countries, we produced a common metric of early childhood development with acceptable discriminant and concurrent validity. Children from diverse countries were shown to have similar developmental profiles with increasing age, supporting theories of a universal unfolding of stage-based skills in the first few years of life that is responsive to environmental and cultural variation.

A primary strength of this study was the use of existing longitudinal data from early childhood (<4 years) and again during middle childhood and adolescence (>4–18 years), circumventing the high cost and time associated with obtaining new data prospectively. Critically, the interval-scale property of the D-score enables quantitative comparisons across ages, which in turn will allow for the construction of international standards for children’s healthy development in the future. Using the D-score, depictions of children’s developmental trajectories with age are easy to interpret, unlike scores obtained from conventional instruments that employ age-based standardisation.

In further contrast to conventional instruments, which are typically designed and validated in a single country or region, data for this study encompassed cohorts from multiple countries and contexts, reflecting children’s development across a diverse global sample. Although representation from high income countries was limited to one country, an innovative feature of the statistical model is that it enables the estimation of D-scores for other item-level datasets not included in this project. Such use of the model will enable external validation in new contexts. A user-friendly open-source platform and algorithm that allows users to generate D-scores from item-level data obtained in their sites is under development (preliminary access to the algorithm is available at <https://github.com/stefvanbuuren/dscore>).

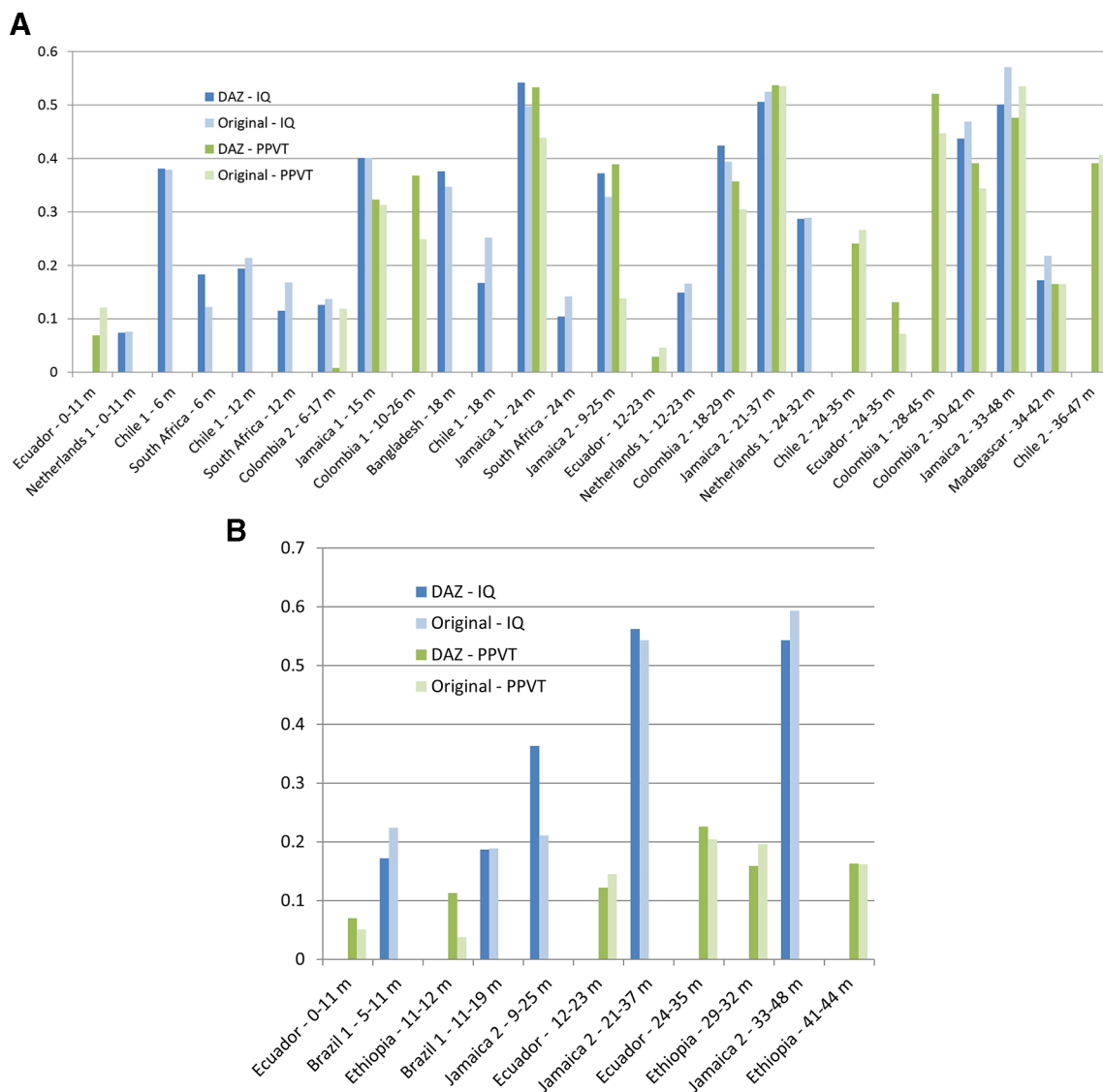


Figure 2 (A) Correlations of DAZ and age-adjusted original measures of early childhood development in children under 48 months with IQ and receptive vocabulary measures at Time 2 for ages >4–9 years, arranged by age at Time 1. For the original instruments, Bayley-I and Bayley-II, we used the MDI in the correlations (Bangladesh, Chile 1 and South Africa); for Bayley-III, we used the measure from the cognition domain (Colombia 1 and Colombia 2). (B) Correlations of DAZ and age-adjusted original measures of early childhood development in children under 48 months with IQ and receptive vocabulary measures at Time 2 for ages >9–18 years, arranged by age at Time 1. For the original instrument, Bayley-III, we used the measure from the cognition domain (Ethiopia). IQ measures are Denver-II, Raven’s and Raven’s (Coloured), SB-4 and SB-5, UKKI, WAIS, WISC-V, and WPPSI. Receptive language measures are the PPVT and its Spanish version, TVIP. Bayley-I,II and III, Bayley Scales for Infant and Toddler Development; Denver-II, DenverDevelopmental Screening Test; DAZ, D-score-for-age z-scores; D-Score, Developmental Score; IQ, Intelligence Quotient; MDI, Mental Development Index; PDI, Psychomotor Development Index; PPVT, Peabody Picture Vocabulary Test; Raven’s, Raven’s Progressive Matrices; SB-4 and SB-5, Stanford Binet Intelligence Scales; TVIP, Test de Vocabulario en Imagenes Peabody; UKKI, Utrechtse Korte Kleuter Intelligentietest; WAIS, Wechsler Adult Intelligence Scale - Revised; WISC-V, Wechsler Intelligence Scale for Children; WPPSI, Wechsler Preschool and Primary Scale of Intelligence

Although a strength, the use of existing data also represents one of the study’s limitations: validation results were affected by differences in sampling strategies across studies (eg, inclusion criteria for low-income and low-middle-income families in Bogota or selection based on children’s stunting status in Jamaica). Nonetheless, predictive validity of the D-score metric to later IQ and language outcomes was comparable to that obtained with the original instruments from which the metric

was derived. It improved with increasing age at Time 1, consistent with other reports, including those using the Bayley.^{34 35} Unexpectedly high correlations of 6-month age group children in the Chile 1 cohort may be a function of the study sampling children with and without iron-deficient anaemia, thus widening the distribution of scores across the whole sample. Similarly, high correlations in the Brazil 1 and Jamaica cohorts, even to 18 years, may be related to sampling groups of normal and low birth weight

(Brazil 1 and Jamaica 1) and stunted and non-stunted (Jamaica 2) children.

Predictive correlations were low in some samples. In rural Africa, these may be explained either by low variability in the samples or from an education bias resulting in poor performance of the school-age instruments. In addition, cohorts in Ethiopia and Madagascar were assessed at Time 2 with adaptations of a receptive vocabulary test that is subject to item bias in countries with multiple languages or dialects.³⁸ The low predictive correlation in Ecuador may be a function of measurement error due to the small number of items used in estimating the D-score in that cohort (as few as four items per child). Finally, the Dutch instrument was designed to screen children for developmental delay such that the high end of the D-score distribution was less well-represented than the low end.

We speculate that the poor performance of the Battelle in the model was due to its original 3-level item scoring, which made it difficult to map Battelle items precisely to items from the other instruments scored as pass/fail. Although some recoded Battelle items had reasonable fit to the Rasch model, in general, they did not equate well with other instruments or demonstrated differential performance by country.

The D-score metric and model set the stage for constructing new instruments with test items that are likely to demonstrate global validity. As we demonstrated with the simulation exercise, fewer items may be necessary than those included in existing conventional instruments, some of which are challenging to adapt to local languages and contexts. Furthermore, by relying on the D-score model's predicted probability for successfully completing each item, we have the opportunity to incorporate adaptive or tailored testing with instruments based on the D-score. Model-based adaptive testing tailors the test to the child's ability level by administering items based on success (or failure) in passing previously administered items. This approach allows for the rapid assessment of a child's development with, for example, 10 or fewer items, while maintaining validity of the metric. Items are selected from the larger pool of items and targeted to the child's age and individual pattern of passing items (ie, children of the same age may be administered different items depending on ability).

The D-score is currently being used by the Global Scale of Early Development (GSED) project to construct two new instruments. The first is intended as a population-level instrument for large-scale surveys, such as the Demographic Health Surveys or UNICEF's Multiple Indicator Cluster Surveys, and will trade off precision in favour of speed and administrative simplicity (ie, using few items and caregiver-report). The second instrument will be for evaluations of small and large-scale programmes and policies.³⁹ The programme evaluation instrument will be longer for better precision and will incorporate both caregiver-report and direct assessment, which takes longer and requires more administrative expertise, but avoids reporting bias, particularly when evaluating parenting programmes.

Instruments based on the D-score, such as the GSED, will allow for the new data collection necessary to develop standards from healthy populations and track country progress towards global goals of early childhood development. Although tracking progress can inform programmes and policies, the history of test score mis-use⁴⁰ and the possibility of invalid and unfair conclusions drawn from cross-national comparisons should be acknowledged. Future examination of D-score trajectories will be most useful in highlighting environmental variations within and across countries, particularly in relation to poverty, education, nurturing care, and nutrition.

CONCLUSION

With the recognition that critical building blocks for adult health and well-being are established early in life,¹ countries throughout the world are instituting policies and programmes to ensure that all children reach their developmental potential. However, evaluating progress has been hampered by the lack of a validated metric of early childhood development across cultures, especially for children 0–3 years living in low- and middle-income countries (LMICs).⁴¹

The D-score metric and model aim to overcome this obstacle in two important ways. First, the D-score model can be used to convert existing data collected from multiple instruments across multiple settings to a common metric of early child development, advancing external validity. Second, the D-score can inform the selection of a subset of items from the larger pool of validated items in the model for constructing culturally-neutral, simple, fast and low-cost instruments, as with the GSED project. The inclusion of instruments based on a common metric in global surveys can ultimately lead to the data collection necessary to establish global standards for early childhood development.

Author affiliations

¹School of Community Health Sciences, University of Nevada Reno, Reno, Nevada, USA

²Department of Pediatrics, Stanford University School of Medicine, Stanford, California, USA

³Inter-American Development Bank, Washington, District of Columbia, USA

⁴Caribbean Institute for Health Research, University of the West Indies, Kingston, Jamaica

⁵Netherlands Organization for Applied Scientific Research TNO, Leiden, Netherlands

⁶Methodology & Statistics, Utrecht University, Utrecht, Netherlands

⁷Institute of Child Health, University College London, London, UK

⁸School of Public Health, University of California Berkeley, Berkeley, California, USA

⁹Maternal and Child Health Division, icddr,b, Dhaka, Bangladesh

¹⁰Institute of Psychiatry, Psychology and Neuroscience, Health Service and Population Research Department, Centre for Global Mental Health, King's College London, London, UK

¹¹Department of Psychiatry, WHO Collaborating Centre for Mental Health Research and Capacity Building, School of Medicine, and Centre for Innovative Drug Development and Therapeutic Trials for Africa (CDT-Africa), College of Health Sciences, Addis Ababa University, Addis Ababa, Ethiopia

¹²Department of Pediatrics, Federal University of Rio Grande, Rio Grande, Brazil

¹³Center for Human Growth and Development, University of Michigan, Ann Arbor, Michigan, USA

¹⁴Centre Médico-Educatif "Les Orchidées Blanches", Antananarivo, Madagascar

¹⁵Centre of Excellence in Human Development, University of the Witwatersrand, Johannesburg, South Africa

¹⁶Department of Pediatrics, University of Maryland School of Medicine, Baltimore, Maryland, USA

¹⁷International Education, RTI International, Research Triangle Park, North Carolina, USA

Acknowledgements We would like to thank the Global Child Development Group collaborators for their data contributions and support of the project. We are also grateful to the many people involved in gathering the data that made this study possible.

Global Child Development Group collaborators Orazio Attanasio; Gary L. Darmstadt; Bernice M. Doove; Emanuela Galasso; Pamela Jervis; Girmay Medhin; Ana M. B. Menezes; Helen Pitchik; Sarah Reynolds; Norbert Schady.

Contributors All authors contributed to item mapping and to analysis decisions during three investigator meetings. SvB and IE conducted the data harmonisation and analyses to derive the model and estimate D-score and DAZ values. AMW and MRC conducted the validation analyses. AMW led the drafting of the paper with guidance from SPW, SGM, MRC, SvB, IE, and MMB. SPW and MMB obtained funding. All authors reviewed the manuscript, provided critical input, and approved submission.

Funding The Global Child Development Group (<https://www.globalchilddevelopment.org/>) was funded by the Bill and Melinda Gates Foundation, OPP1138517, to perform this study. The Bernard van Leer Foundation supported the initial meeting of investigators to establish the Advisory Board and conduct the instrument mapping. CH (King's College London and AAU) is funded by the National Institute of Health Research (NIHR) Global Health Research Unit on Health System Strengthening in Sub-Saharan Africa, King's College London (GHRU 16/136/54) using UK aid from the UK Government. The views expressed in this publication are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care, or of the Inter-American Development Bank, their Board of Directors, or the countries they represent. CH additionally receives support from the African Mental Health Research Initiative (AMARI) as part of the DELTAS Africa Initiative [DEL-15-01]. The original data collected in Ethiopia was funded by the Wellcome Trust (project grant 093559).

Competing interests CH receives support from the African Mental Health Research Initiative (AMARI) as part of the Wellcome Trust-funded DELTAS Africa Initiative [DEL-15-01]. The original data collected in Ethiopia was funded by the Wellcome Trust (project grant 093559).

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>.

ORCID iDs

Ann M Weber <http://orcid.org/0000-0001-8130-5858>

Marta Rubio-Codina <https://orcid.org/0000-0002-1286-7918>

Charlotte Hanlon <http://orcid.org/0000-0002-7937-3226>

Linda Richter <http://orcid.org/0000-0002-3654-3192>

Maureen M Black <http://orcid.org/0000-0002-6427-4639>

REFERENCES

- Phillips DA, Shonkoff JP. *From neurons to neighborhoods: the science of early childhood development*. National Academies Press, 2000.
- Sameroff A. *The transactional model*. American Psychological Association, 2009.
- Ertem IO, Krishnamurthy V, Mulaudzi MC, et al. Similarities and differences in child development from birth to age 3 years by sex and across four countries: a cross-sectional, observational study. *Lancet Glob Health* 2018;6:e279-91.
- Villar J, Fernandes M, Purwar M, et al. Neurodevelopmental milestones and associated behaviours are similar among healthy children across diverse geographical locations. *Nat Commun* 2019;10:511.
- Fernald LC, Prado E, Kariger P, et al. *A toolkit for measuring early childhood development in low and middle-income countries*, 2017.
- Jacobusse G, van Buuren S, Verkerk PH. An interval scale for development of children aged 0-2 years 2006;25:2272-83.
- Victora CG, Araújo CLP, Menezes AMB, et al. Methodological aspects of the 1993 Pelotas (Brazil) birth cohort study. *Rev. Saúde Pública* 2006;40:39-46.
- Moura DR, Costa JC, Santos IS, et al. Natural history of suspected developmental delay between 12 and 24 months of age in the 2004 Pelotas birth cohort. *J Paediatr Child Health* 2010;46:329-36.
- Contreras D, González S. Determinants of early child development in Chile: health, cognitive and demographic factors. *Int J Educ Dev* 2015;40:217-30.
- Hanlon C, Medhin G, Alem A, et al. Impact of antenatal common mental disorders upon perinatal outcomes in Ethiopia: the P-MaMiE population-based cohort study. *Trop Med Int Health* 2009;14:156-66.
- Hanlon C, Medhin G, Worku B, et al. Adapting the Bayley scales of infant and toddler development in Ethiopia: evaluation of reliability and validity: measuring child development in Ethiopia. *Child: Care, Health and Development* 2016;42:699-708.
- Doove BM. Ontwikkeling kinderen in Maastricht en Heuvelland (mom), Evaluatie integraal kindvolgsysteem voor signalering in de Jeugdgezondheidszorg: MOMknowsbest. Maastricht, the Netherlands, 2010. Available: <https://academischewerkplaat.slimburg.nl/wp-content/uploads/170310-Mom-knows-best.pdf> [Accessed 3 Dec 2018].
- Richter L, Norris S, Pettifor J, et al. Cohort profile: Mandela's children: the 1990 birth to twenty study in South Africa. *Int J Epidemiol* 2007;36:504-11.
- Herngreen WP, Reerink JD, van Noord-Zaadstra BM, et al. SMOCC: design of a representative Cohort-study of Live-born infants in the Netherlands. *The European Journal of Public Health* 1992;2:117-22.
- Rubio-Codina M, Araujo MC, Attanasio O, et al. Concurrent validity and feasibility of short tests currently used to measure early childhood development in large scale studies. *PLoS One* 2016;11:e0160962.
- Tofail F, Persson Lars Åke, El Arifeen S, et al. Effects of prenatal food and micronutrient supplementation on infant development: a randomized trial from the maternal and infant nutrition interventions, Matlab (MINIMat) study. *Am J Clin Nutr* 2008;87:704-11.
- Lozoff B, Andraca D I, Castillo M, et al. Behavioral and developmental effects of preventing iron-deficiency anemia in healthy full-term infants. *Pediatrics* 2003;112:846-54.
- Lozoff B, Jiang Y, Li X, et al. Low-Dose iron supplementation in infancy modestly increases infant iron status at 9 Mo without decreasing growth or increasing illness in a randomized clinical trial in rural China. *J Nutr* 2016;146:612-21.
- Attanasio OP, Fernández C, Fitzsimons EOA, et al. Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in Colombia: cluster randomized controlled trial. *BMJ* 2014;349:g5785.
- Paxson C, Schady N. Does money matter? the effects of cash transfers on child development in rural Ecuador. *Econ Dev Cult Change* 2010;59:187-229.
- Walker SP, Chang SM, Powell CA, et al. Psychosocial intervention improves the development of term low-birth-weight infants. *J Nutr* 2004;134:1417-23.
- Grantham-McGregor SM, Powell CA, Walker SP, et al. Nutritional supplementation, psychosocial stimulation, and mental development of stunted children: the Jamaican study. *The Lancet* 1991;338:1-5.
- Fernald LCH, Weber A, Galasso E, et al. Socioeconomic gradients and child development in a very low income population: evidence from Madagascar. *Dev Sci* 2011;14:832-47.
- Tsai JL, Knutson B, Fung HH. Cultural variation in affect valuation. *J Pers Soc Psychol* 2006;90:288-307.
- Wilson M, Allen DD, Li JC. Improving measurement in health education and health behavior research using item response modeling: introducing item response modeling. *Health Educ Res* 2006;21(Suppl 1):i4-18.
- Kolen MJ, Brennan RL, equating T. *Test equating, scaling, and linking*, 2004.
- Eekhout I, Weber AM, van Buuren S. Equate groups: an innovative method to link instruments across cohorts and contexts. *Applied Psychological Measurement*. Under review.
- Bock RD, Mislevy RJ. Adaptive EAP estimation of ability in a microcomputer environment. *Appl Psychol Meas* 1982;6:431-44.
- van Buuren S. Growth charts of human development. *Stat Methods Med Res* 2014;23:346-68.

- 30 Cole TJ, Green PJ. Smoothing reference centile curves: the lms method and penalized likelihood. *Stat Med* 1992;11:1305–19.
- 31 Walker SP, Wachs TD, Meeks Gardner J, *et al*. Child development: risk factors for adverse outcomes in developing countries. *The Lancet* 2007;369:145–57.
- 32 World Health Organization. The who child growth standards, 2011. Available: <http://www.who.int/childgrowth/en/> [Accessed 14 Sept 2015].
- 33 Black MM, Walker SP, Fernald LCH, *et al*. Early childhood development coming of age: science through the life course. *The Lancet* 2017;389:77–90.
- 34 Bruce AB. Creating the optimal preschool testing situation. in: psychoeducational assessment of preschool children. *Routledge* 2017:137–54.
- 35 Snow CE, Van Hemel SB, National Research Council of the National Academies, National Research Council, National Academies. *Early childhood assessment: why, what, and how*. Citeseer, 2008.
- 36 Evans JD. *Straightforward statistics for the behavioral sciences*. Pacific Grove: Thomson Brooks/Cole Publishing Co, 1996.
- 37 Robitzsch A. *Sirt: supplementary item response theory models, R package version 2.6-9*, 2018.
- 38 Weber AM, Fernald LCH, Galasso E, *et al*. Performance of a receptive language test among young children in Madagascar. *PLoS One* 2015;10:e0121767.
- 39 The GSED team. The global scale for early development (GSED). *Early childhood matters* 2019;14:80–4.
- 40 Tucker WH. *The Cattell controversy: race, science, and ideology*. University of Illinois Press, 2010.
- 41 Chan M. Linking child survival and child development for health, equity, and sustainable development. *Lancet* 2013;381:1514–5.
- 42 Andrew A, Attanasio O, Fitzsimons E, *et al*. Impacts 2 years after a scalable early childhood development intervention to increase psychosocial stimulation in the home: a follow-up of a cluster randomised controlled trial in Colombia. *PLoS Med* 2018;15:e1002556.