Global Scales for Early Development v1.0

Technical report

World Health Organization

Images ©WHO
Design and layout: büro svenja

Global Scales for Early Development v1.0

Technical report

World Health Organization

WHO/MSD/GSED package v1.0/2023.1

Global Scales for Early Development v1.0 Technical report – Global Scales for Early Development v1.0 Short Form (caregiver-reported) – Global Scales for Early Development v1.0 Short Form (caregiver-reported). Item guide – Global Scales for Early Development v1.0 Short Form (caregiver-reported). User manual – Global Scales for Early Development v 1.0 Long Form (directly administered) – Global Scales for Early Development v 1.0 Long Form (directly administered). Item guide – Global Scales for Early Development v1.0 Long Form (directly administered). User manual – Global Scales for Early Development v1.0 Scoring guide – Global Scales for Early Development v1.0 Adaptation and translation guide

Selected questions and descriptions for the GSED measures have been reproduced or adapted from the following tools/assessments: Ages and Stages Questionnaire, third edition (ASQ-3); Bayley Scales of Infant Development (Bayley); Bayley Scales of Infant Development, second edition (Bayley II); Caregiver-Reported Early Development Instruments (CREDI); Denver Developmental Screening Test (DDST); Denver Developmental Screening Test, second edition (DDST II); Developmental Milestones Checklist (DMC); Developmental Milestones Checklist II (DMC II); Dutch Development Instrument (DDI); Griffiths Mental Development Scales (GMDS); Griffiths Mental Development Scales – South African version (GMDS-SA); Kilifi Developmental Inventory (KDI); Malawi Developmental Assessment Tool (MDAT); Preschool Pediatric Symptoms Checklist (PPSC); Saving Brains Early Childhood Development Scale (SBECD); Stanford-Binet Intelligence Scales, fifth edition (SBIS-5); Test de Desarrollo Psicomotor [Psychomotor Development Test] (TEPSI); and Vineland Adaptive Behavior Scales (Vineland) (see Bibliography for details).

Contents

Acknowledgements

Vision and conceptualization

The Global Scales for Early Development (GSED) package v1.0 was developed under the overall guidance of and conceptualization by Tarun Dua and Dévora Kestel of the Department of Mental Health and Substance Use of the World Health Organization (WHO).

Project coordination

The development of the GSED package was coordinated by Vanessa Cavallera, Brain Health Unit, Department of Mental Health and Substance Use, WHO.

Writing of the Technical Report

The writing team included (in alphabetical order): Salahuddin Ahmed (Projahnmo Research Foundation, Bangladesh), Romuald Kouadio E Anago (Innovations for Poverty Action [IPA], Côte d'Ivoire), Abdullah H Baqui (Johns Hopkins University, United States of America [USA]), Maureen Black (RTI International and University of Maryland School of Medicine, USA), Alexandra Brentani (University of São Paulo Medical School, Brazil), Kieran Bromley (Keele University, United Kingdom), Stef van Buuren (Netherlands Organisation for Applied Scientific Research [TNO] and Utrecht University, Netherlands), Vanessa Cavallera (WHO headquarters, Switzerland), Symone Detmar (TNO, Netherlands), Tarun Dua (WHO headquarters, Switzerland), Arup Dutta (Center for Public Health Kinetics [CPHK], United Republic of Tanzania), Iris Eekhout (TNO, Netherlands), Melissa Gladstone (University of Liverpool, United Kingdom), Katelyn Hepworth (University of Nebraska, USA), Andreas Holzinger (IPA, Côte d'Ivoire), Magdalena Janus (McMaster University, Canada), Fyezah Jehan (Aga Khan University [AKU], Pakistan), Fan Jiang (Shanghai Jiao Tong University School of Medicine, China), Patricia Kariger (Independent Consultant, USA), Raghbir Kaur (WHO headquarters, Switzerland), Rasheda Khanam (Johns Hopkins University, USA), Gillian Lancaster (Keele University, United Kingdom), Dana McCoy (Harvard Graduate School of Education, USA), Gareth McCray (Keele University, United Kingdom), Imran Nisar (AKU, Pakistan), Ambreen Nizar (AKU, Pakistan), Mariana Pacifico Mercadante (University of São Paulo Medical School, Brazil), Michelle Pérez Maillard (WHO headquarters, Switzerland), Abbie Raikes (University of Nebraska Medical Center, USA), Arunangshu Dutta Roy (Projahnmo Research Foundation, Bangladesh), Marta Rubio-Codina (Inter-American Development Bank, USA), Sunil Sazawal (CPHK, United Republic of Tanzania), Yvonne Schönbeck (TNO, Netherlands), Jonathan Seiden (Harvard Graduate School of Education, USA), Fahmida Tofail (International Centre for Diarrhoeal Disease Research, Bangladesh), Marcus Waldman (University of Nebraska Medical Center, USA), Ann M Weber (University of Nevada, USA), Yunting Zhang (Shanghai Jiao Tong University School of Medicine, China), Arsène Zongo (IPA, Côte d'Ivoire).

Technical contribution and review

Valuable input was received from WHO staff at headquarters and in regional and country offices, as well as from many international experts and programme implementers. These technical and/or data contributions were central to the development of the GSED package. Those who provided input include (in alphabetical order): Amina Abubakar (AKU, Kenya), Fahad Aftab (Center for Public Health Kinetics, United Republic of Tanzania), Claudia Regina Lindgren Alves (Universidade Federal de Minas Gerais, Brazil), Omar Ali Ali (WHO Country Office, United Republic of Tanzania), Elisa Altafim (Universidade de São Paulo, Brazil), Maria Caridad Araujo (Inter-American Development Bank, USA), Orazio Attanasio (Institute for Fiscal Studies, United Kingdom), Farzana Begum (AKU, Pakistan), Florence Baingana (WHO Regional Office for Africa, Congo), Molly Biel (WHO headquarters, Switzerland), Geoffrey Bisoborwa (WHO Regional Office for Africa, Congo), Andrea Bruni (WHO Regional Office for South-East Asia, India), Betzabe Butron Riveros (WHO Regional Office for the Americas, USA), Claudia Cappa (UNICEF, USA), Andreana Castellanos (Afinidata, Guatemala), Susan M Chang (University of the West Indies, Jamaica), Alexandra Chen (Harvard University, USA), Anne Marie Chomat (McGill

University, Canada), Marie-Helene Cloutier (The World Bank, USA), Bernadette Daelmans (WHO headquarters, Switzerland), Sandra Dao (WHO Country Office, Côte d'Ivoire), Teshome Desta Woldehanna (WHO Country Office, Kenya), Erinna Dia (UNICEF, USA), Bernice M Doove (Maastricht University, Netherlands), Dan Fang (WHO Country Office, China), Wafaie Fawzi (Harvard University, USA), Lia CH Fernald (University of California, USA), Emanuela Galasso (The World Bank, USA), Sally M Grantham-McGregor (Emeritus Professor of International Child Health at University College London, United Kingdom), Shuchita Gupta (WHO Regional Office for South-East Asia, India), Cristina Gutierrez de Piñeres (United Way, Colombia), Jena Derakhshani Hamadani (International Centre for Diarrheal Disease Research, Bangladesh), Charlotte Hanlon (Addis Ababa University, Ethiopia), Natalia Henao (Genesis Foundation, Colombia), Alaka Holla (The World Bank, USA), Sidra Inayat (AKU, Pakistan), Matias Irarrazaval (WHO Regional Office for the Americas, USA), Nemes Joseph Iriya (WHO Country Office, United Republic of Tanzania), Charles Newton (Kenya Medical Research Institute, Kenya), Ledia Lazeri (WHO Regional Office for Europe, Denmark), Zhao Li (WHO Regional Office for the Western Pacific, Philippines), Pamela Jervis (Universidad de Chile, Chile), Codie Kane (Temple University, USA), Simone M Karam (Federal University of Rio Grande Brazil), Janet Kayita (WHO Regional Office for Africa, Congo), Jamila Khalfan (Public Health Laboratory-IdC, United Republic of Tanzania), Neema Kileo (WHO Country Office, United Republic of Tanzania), Betsy Lozoff (University of Michigan, USA), Diego Luna (The World Bank, USA), Raquel Dulce Maguele (WHO Country Office, Mozambique), Limbika Agatha Maliwichi (University of Malawi, Malawi), Sheila Manji (WHO headquarters, Switzerland), Susanne Martin Herz (University of California San Francisco, USA), Jeffrey Measelle (University of Oregon, USA), Girmay Medhin (Aklilu Lemma Institute of Pathobiology, Ethiopia), Patricia Medrano (University of Chile, Chile), Junaid Mehmood (AKU, Pakistan), Rajesh Mehta (WHO Consultant, India), Ana MB Menezes (Universidade Federal de Pelotas, Brazil), Assumpta Muriithi (WHO Regional Office for Africa, Congo), Néllia Mutisse (WHO Country Office, Mozambique), Alphoncina Nanai (WHO Country Office, United Republic of Tanzania), Renato Oliveira e Souza (WHO Regional Office for the Americas, USA), Nicole Petrowski (UNICEF, USA), Lauren Pisani (Save the Children, United Kingdom), Helen Pitchik (University of California at Berkeley School of Public Health, USA), Chemba Raghavan (UNICEF, USA), Muneera Rasheed (AKU, Pakistan), Lisy Ratsifandrihamanana (Centre Médico-Educatif, Madagascar), Sarah Reynolds (University of California at Berkeley School of Public Health, USA), Linda Richter (University of the Witwatersrand, South Africa), Peter C Rockers (Boston University School of Public Health, USA), Marta Rubio-Codina (Inter-American Development Bank, USA), Norbert Schady (The World Bank, Washington), Khalid Saeed (WHO Regional Office for the Eastern Mediterranean, Egypt), Makeba Shiroya (WHO Country Office, Kenya), Kathleen Louise Strong (WHO headquarters, Switzerland), Christopher R Sudfeld (Harvard University, USA), Edwin Exaud Swai (WHO Country Office, United Republic of Tanzania), Safila Telatela (WHO Country Office, United Republic of Tanzania), Daisy Trovada (WHO Country Office, Mozambique), Martin Vandendyck (WHO Regional Office for the Western Pacific, Philippines), Paul H Verkerk (TNO, Netherlands), Susan P Walker (Caribbean Institute for Health Research, Jamaica), Christine Wong (Hong Kong University, China), Dorianne Wright (Oregon Health & Science University, USA) and Aisha K Yousafzai (Harvard University, USA).

Information technology programming

We acknowledge UniversalDoctors (Jordi Serrano Pons, Fernando Vaquero, Jeannine Lemaire and Montse Garcia) for conceptualization and initial information technology support to the GSED application creation, and the CPHK (Arup Dutta, Vishi Saxena, Waseem Ali, Poonam Rathore) for further conceptual development and operationalization of the GSED App as well as data management throughout implementation of the project.

Financial support

Support for this work was received from (in alphabetical order): the Bernard van Leer Foundation, Bill & Melinda Gates Foundation, Children's Investment Fund Foundation, Jacobs Foundation, King Baudouin Foundation United States and the United States Agency for International Development.

Abbreviations

2PL	two-parameter logistic
AKU	Aga Khan University
ASQ	Ages and Stages Questionnaire
Bayley-III	Bayley Scales of Infant and Toddler Development, Third Edition
BRS	Brief Resilience Scale
CAT	computerized adaptive tests
CB	combined format
CI	confidence interval
CPAS	Childhood Psychosocial Adversity Scale
CPHK	Center for Public Health Kinetics
CREDI	Caregiver Reported Early Development Instruments
D-score	Developmental score
DAZ	Development-for-Age z-score
DDI	Dutch Development Instrument
DHS	Demographic Health Survey
ECD	early childhood development
ECDI2030	Early Childhood Development Index 2030
EDC	electronic data capture
FCI	Family Care Indicator
FGD	focus group discussion
FSS	Family Support Scale
GMDS	Griffiths Mental Development Scales
GSED	Global Scales for Early Development
HAZ	height-for-age z-score
HF	Household Form
HOME	Home Observation Measurement of the Environment
ICC	intraclass correlation coefficient
IPA	Innovations for Poverty Action
IRT	item response theory
IYCD	WHO Indicators of Infant and Young Child Development
LF	Long Form
ODK	Open Data Kit
PF	Psychosocial Form
PHQ9	Patient Health Questionnaire-9
PRIDI	Regional Project on Child Development Indicators
SD	standard deviation
SDG	Sustainable Development Goal
SEE	standard error of estimation
SES	socioeconomic status
SF	Short Form
SME	subject matter expert
TNO	Netherlands Organisation for Applied Scientific Research
TPP	Target Product Profile
WAZ	weight-for-age z-score
WHO	World Health Organization

© WHO / Alasdair Bell

> "The GSED package v1.0 includes open-access measures that provide a standardized method for measuring the development of children up to 36 months of age across diverse cultures and contexts."

Executive summary

This *Technical Report* documents the development and validation of the Global Scales for Early Development (GSED). The GSED package v1.0 includes open-access measures that provide a standardized method for measuring the development of children up to 36 months of age across diverse cultures and contexts. It has been created to serve as a population-level assessment of early childhood development (ECD) (up to 36 months) for the global community that may be used for comparisons across countries. In contrast to growth, which is measured by changes in children's weight in grams and height in centimetres, there has been no uniform scale for children's early development. The GSED uses an innovative metric, the Developmental score (D-score), a scale with interval properties, to measure children's development.

The package includes the GSED measures v1.0 as well as accompanying materials to facilitate their implementation and use. The GSED measures are meant to collect population-level data on ECD and are designed to be used primarily for research and programmatic evaluations.
They comprise a:

Global Scales for Early Development v1.0
Short Form (caregiver-reported)
World Health Organization

Short Form (SF), a caregiver-reported measure; and

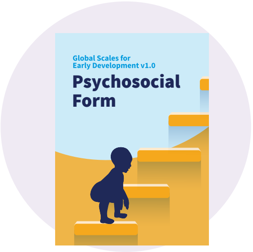
Global Scales for Early Development v1.0
Long Form (directly administered)
World Health Organization

Long Form (LF), comprised of items administered directly to children by a trained assessor.

Current evidence indicates that the psychometric properties of the GSED SF and LF are comparable. The choice of one or the other, or the two together, to measure child development should be dictated by: i) the purpose of the evaluation and/or specific research question (e.g. type of intervention); ii) preferred modality of administration (caregiver report versus direct administration); and iii) the capacity and expertise of the team. A combined format (CB) of GSED SF together with GSED LF may be used to increase measurement precision. Further evidence on sensitivity to potential changes after interventions and increase in precision of the GSED CB is currently being tested and will be made available in the near future.

The GSED measures also include a:[1]

Household Form (HF), a caregiver-reported measure, designed to be integrated into large-scale and national-level surveys for monitoring child development; and

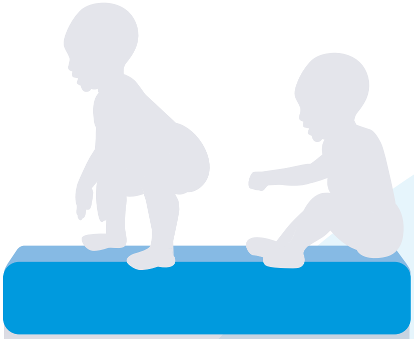


Psychosocial Form (PF), a caregiver-reported measure of children's psychosocial behaviours.

The GSED measures were created through an integrated empirical-conceptual approach. The empirical approach included statistical modelling of a data set of 100 153 observations. The conceptual approach included subject matter experts (SMEs) identifying conceptually relevant and globally feasible items of child development for children under 36 months of age.

A rigorous and standardized method has been used to evaluate the psychometric properties of the newly-created GSED measures in seven countries. A prospective cross-sectional design was implemented, including a six-month longitudinal follow up, with an age- and sex-stratified sample of children.

This interim package (v1.0) is based on validation data from three countries: Bangladesh, Pakistan and the United Republic of Tanzania. The results demonstrate statistically significant reliability and validity of the D-score to measure child development under 36 months of age at the population level. Specifically, convergent validity measured through contextual measures likely to be related to child development (e.g. socioeconomic status [SES] and exposure to adversity) and concurrent validity with the Bayley Scales of Infant and Toddler Development (Bayley-III) were statistically significant. Additional analyses have shown that GSED measures are culturally neutral, have good content validity and are easy to implement at scale. Revisions to the package are planned for 2024 when data are available from four additional countries: Brazil, China, Côte d'Ivoire and the Netherlands.



[1] Not part of this Technical Report in detail as they are being further tested, but can be made available on request.

1. Introduction

The pathways to adult health and well-being begin in childhood are often measured by children's growth and development. Both are products of children's specific genetic blueprint and influenced by environmental factors that begin prenatally *(1)*. Children in the early years and, in particular, the first 1000 days (from conception to age 24 months) are highly sensitive to environmental conditions due to the rapid brain development that occurs during this period. Monitoring of child development at this time is important to track progress toward global and national policy goals for children and provides a critical reference to plan and evaluate services to support healthy development. However, there have been no universal measures designed to quantify children's development during the earliest years at population level *(2)*. Without such measures, countries are unable to monitor children's progress and determine how to allocate resources to provide the necessary support for children to reach their developmental potential.

Measures of stunting and severe poverty have been effective indicators of the proportion of children at risk of not reaching their developmental potential *(3)* and have contributed to advances in global policies and programmes for young children *(4)*. However, they have limited predictive ability *(5)* and lack sensitivity, and as such they are not suitable to measure change in response to interventions or environmental conditions. Countries have traditionally used physical growth as a proxy based on findings that chronic linear growth faltering (i.e. stunting: height-for-age < 2 standard deviations [SDs] below WHO growth standards) is associated with fewer years of adult schooling, poorer economic indices and greater likelihood of experiencing poverty *(6,7)*. Reductions in the proportion of childhood stunting is frequently used as a national and global target, often also as a proxy for developmental outcomes improvement.

© WHO / Ala Kheir

The urgent need for a comprehensive measure and reporting system for child development has been reinforced by the United Nations Sustainable Development Goals (SDGs). Specifically, SDG4 (education) calls for ensuring inclusive and equitable quality education and promoting lifelong learning opportunities. SDG4 includes an early childhood-specific indicator, 4.2.1, which mandates measurement of the proportion of children under 5 years of age who are developmentally on track in health, learning and psychosocial well-being, by sex *(8)*. WHO's Multicentre Growth Reference Study has shown that under optimal conditions, children's early growth (up to age 2 years) is comparable across countries *(9)*. These findings form the basis for global growth standards that have been adopted by 125 countries. Similarly, emerging evidence suggests universality in multiple domains of children's development across countries during the first 2 to 3 years of life *(10,11)*. Additionally, in contrast to growth, which is measured universally by changes in children's weight in grams and height in centimetres, there is no unit (or metric) for measuring child development.

To address the lack of a population or programmatic measure or metric of ECD, WHO assembled an interdisciplinary and multi-country team to develop the GSED. These measures of ECD for children up to 36 months of age provide a metric of child development (the D-score) at both the population and programmatic level as well as a system for interpreting scores.

This Technical Report summarizes the process of creating the GSED and related measures, and describes the validation methodology and psychometric properties in three countries (Bangladesh, Pakistan and the United Republic of Tanzania). Data from additional countries (Brazil, China, Côte d'Ivoire and the Netherlands) (currently being collected) will address broader global validity and inform potential revisions of the measures. The report concludes by describing the release of the package v1.0, the next steps for GSED, and the process of dissemination and continuous feedback to enable the use of GSED to monitor progress towards global goals and inform programmes and policy to promote the health and development of young children globally.

> “Monitoring of child development in the early years is important to track progress toward global and national policy goals for children and provides a critical reference to plan and evaluate services to support children's healthy development.”

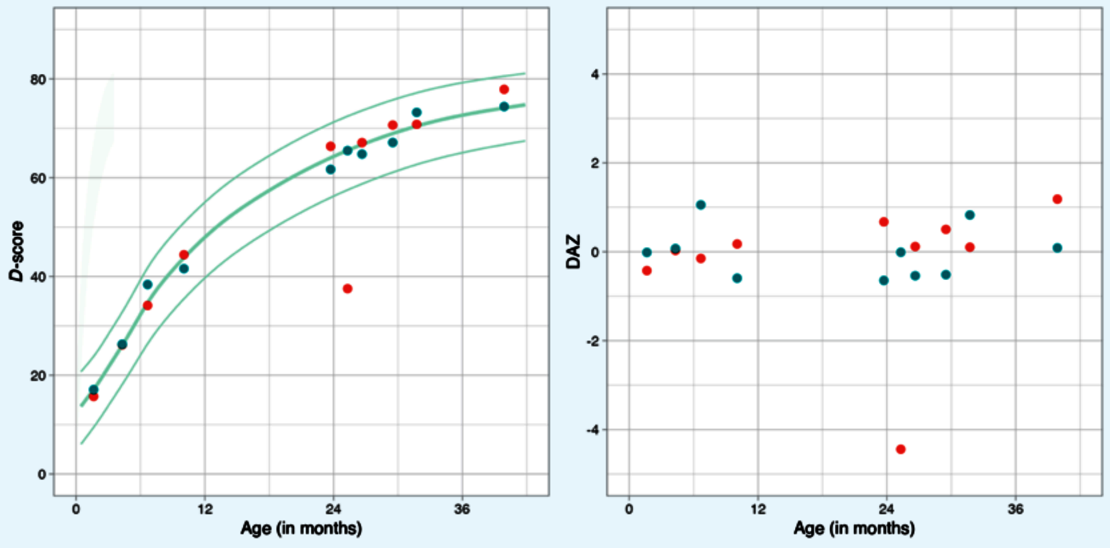
2. GSED: overview

The GSED is an open-access package specifically designed to provide a standardized method for measuring development of children up to 36 months of age at population level globally. The GSED meets that objective by incorporating all domains of ECD through a common scale for measurement translated into a single score, the D-score, that represents holistic development. This section of the report provides an overview of the D-score approach and the GSED measures.

The D-score

The D-score is a unit of measurement with an interval scale representing child development by a single number *(12,13)*. As height (in centimetres) and weight (in grams) change over time with the growth of the child, development (measured in D-score units) also increases with age as the child acquires more skills. The D-score is calculated from Yes/No responses on a set of age-appropriate developmental items (e.g. “Can the child stack two blocks”, or “Does the child use two-word sentences?”). Conceptually, a child’s D-score falls along a developmental continuum, beginning with simple skills and behaviours that the child is able to perform and progressing through the child’s repertoire until reaching skills and behaviours that the child has yet to acquire (see Figure 1). It is calculated as the mean of the posterior distribution conditional on the responses, the items’ difficulty and the child’s age. The D-score may also be transformed into the Development-for-Age z-score (DAZ). The DAZ is age-independent and is scaled such that at each age, the distribution of scores is normally distributed with a mean of 0 and a variance of 1. Since DAZ adjusts for the natural increase in D-score with age, it helps ease the comparison between samples from different ages or countries. Similar to height-for-age z-score (HAZ) and weight-for-age z-score (WAZ), the DAZ is calculated relative to a reference population.

FIGURE 1. D-SCORE AND DAZ



> “The GSED incorporates all domains of ECD through a common scale for measurement of holistic development. The D-score is a unit of measurement with an interval scale representing child development by a single number.”

The GSED measures

The GSED SF and LF measures were created as a response to the need for a population-based measure of ECD up to 36 months of age. Criteria for the measure were that it should be reliable, valid globally, easy to administer in field conditions, require limited training, be easy to interpret and openly accessible. The GSED measures capture the underlying construct of child development which includes the motor, language, cognitive and socio-emotional domains. The measures comprise a caregiver report SF and a direct administration LF. Each GSED measure (SF or LF) can be stand-alone, or used in combination with each other for a more comprehensive and precise assessment of child development, depending on the requirements of the measurement effort (e.g. research, programmatic evaluation, etc.).

The GSED measures are designed for use at population and programmatic level. Even though they are collected for individual children, the results are not validated to be interpreted for any specific child. The GSED measures have not yet been tested within the context of clinical use and should therefore not be used for screening individual children for developmental delays or impairments nor for diagnosis.

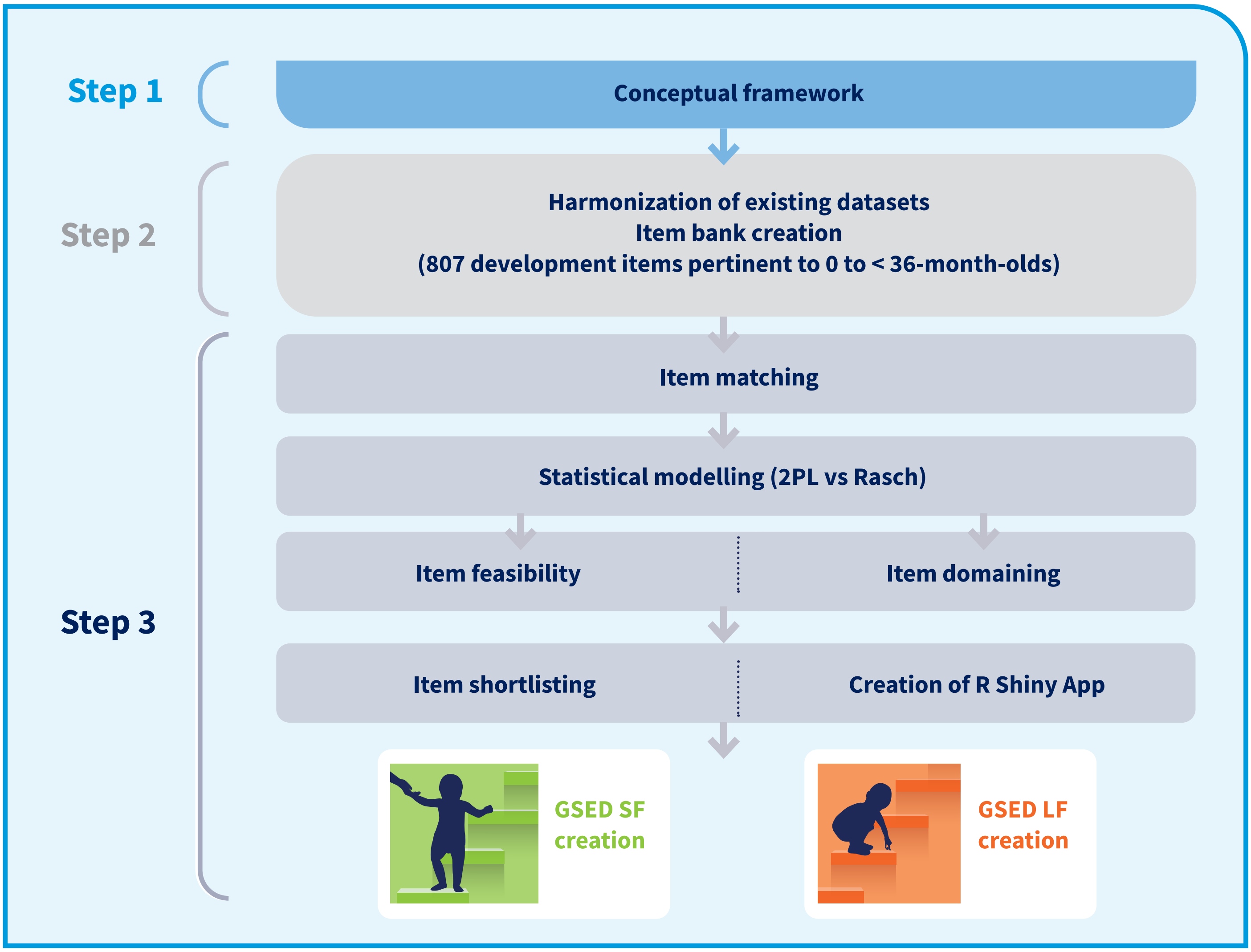
Given the need for an easy-to-interpret score, as well as evidence that in the first years of life development is integrated, overlapping and linked across multiple areas *(14,15)*, developmental domain sub-scores are not provided. Rather an aggregate score comprises the essence of each domain at each point in time. The intended use includes the evaluation of the impact of interventions applied to groups, and the comparison of groups of children (for example, for monitoring or research purposes). The GSED measures may be administered using the GSED App on a tablet or as a paper version. The development and validation process of the GSED measures as well as a detailed description of each measure are covered in Sections 3 and 4 of this report.

Two additional caregiver report forms are being further developed and tested. The GSED HF is being proposed to be used for inclusion in large-scale data collection efforts, including existing multi-topic household surveys (such as Multiple Indicator Cluster Surveys or Demographic Health Surveys [DHS]). The GSED PF aims to capture psychosocial skills and behaviours or non-normative developmental patterns that may indicate challenges in children’s mental health. The GSED PF’s items are not age-ordinal, and hence their results are not scored on the D-score scale. The creation and validation of the GSED HF and PF across different cultural contexts are ongoing and are part of forthcoming work (see Section 6 and Boxes 4 and 5 for an overview).

3. Development of GSED v1.0 SF and LF measures

This section describes the development process of the GSED SF and LF measures, from conceptualization through statistical analysis and expert consensus, to the prototypes for validation *(16)*. The GSED development followed a rigorous methodological process which required a multi-step approach as summarized in Figure 2.

FIGURE 2. GSED MEASURES CREATION FLOWCHART



Step 1. Conceptual framework: Target Product Profile (TPP)

A conceptual framework was developed by a group of international ECD and measurement experts and researchers *(13, 17-18)* through consensus around the assumption that ECD refers to various skills and behaviours that emerge sequentially early in life and are influenced by gene-environment interactions occurring between conception and 3 years of age *(19,20)*. Therefore, the first step in the GSED measures development process focused on identifying target properties and implementation characteristics that the package of measures needed to achieve. The output of this process was the TPP listing minimum and optimal target characteristics.

Step 2. Harmonization of existing datasets on child development

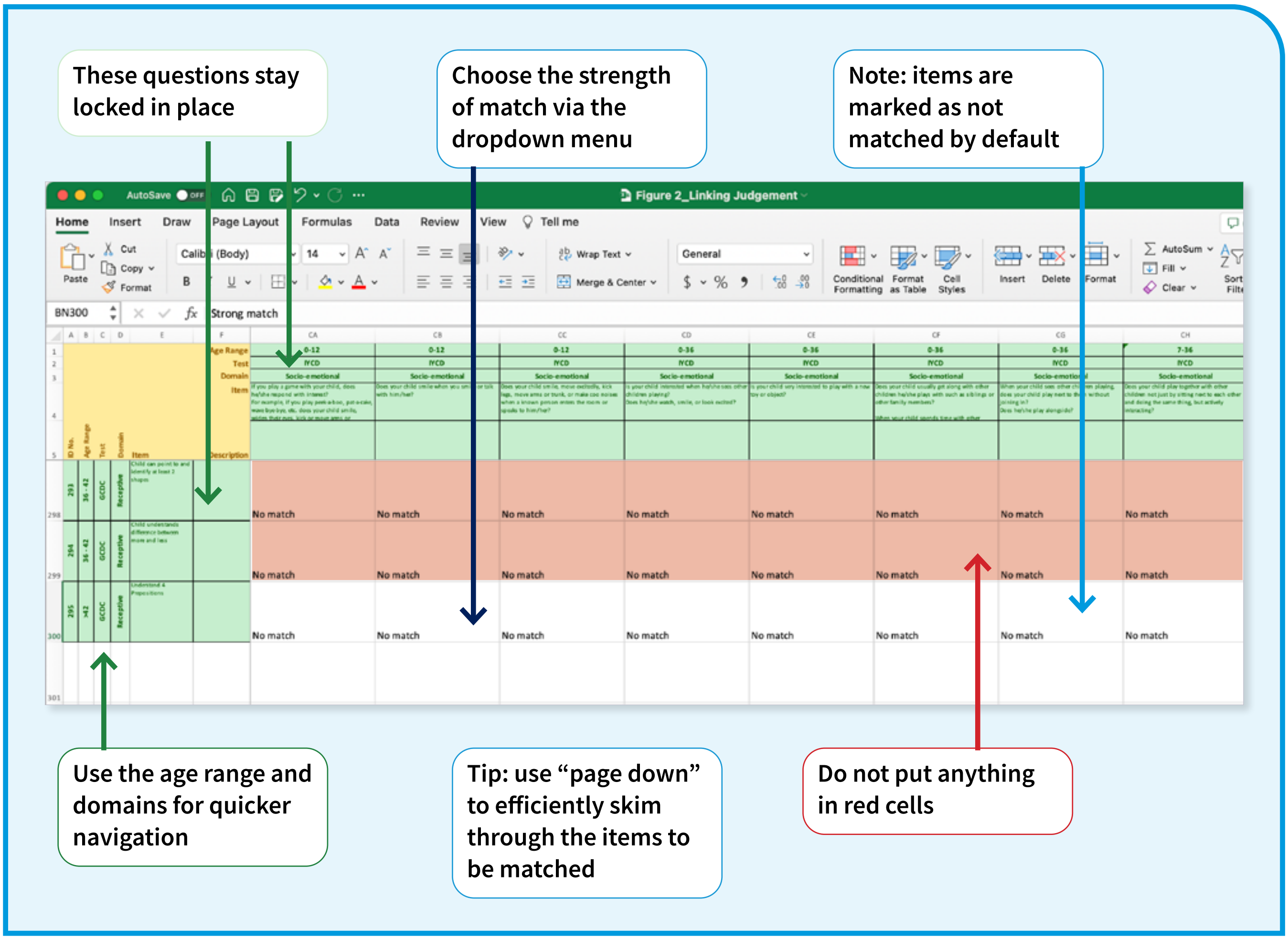
Once the TPP was in place, an extensive collection of ECD datasets was brought together into an item bank. The common dataset included data from 18 instruments, 51 cohorts, 32 countries (of which 30 are low- or middle-income), 66 075 unique children, 100 153 observations (child/age combinations) and 4 314 146 scores (see Annex 1 for details). The total number of items was 2221. Sixteen studies included longitudinal data. Most instruments collected dichotomous Yes/No scores, but a few used additional categories, such as ”Sometimes” or ”Not yet”. In consultation with SMEs, these responses were re-coded into binary 0/1 scores.

Step 3. Creation of the GSED SF and GSED LF measures

Step 3.1 Item matching

The item-matching component was designed to establish the strength of the relations between pairs of items across different instruments in the GSED item bank that measure the same or very similar skills and behaviours in slightly different ways. SMEs in child development classified such similar items from different instruments into clusters or equate groups. A subset of these groups was used to connect every instrument to a shared latent factor (child development construct). Additionally, these clusters helped link instruments and avoided duplication in the final measures (see Figure 3 for the data collection matrix).

FIGURE 3. DATA COLLECTION MATRIX AND CRITERIA FOR MATCHING EXERCISE



These questions stay locked in place

Choose the strength of match via the dropdown menu

Note: items are marked as not matched by default

Use the age range and domains for quicker navigation

Tip: use "page down" to efficiently skim through the items to be matched

Do not put anything in red cells

						CA	CB	CC	CD	CE	CF	CG	CH
1					Age Range	0-12	0-12	0-12	0-36	0-36	0-36	0-36	7-36
2					Test	IYCD	IYCD	IYCD	IYCD	IYCD	IYCD	IYCD	IYCD
3					Domain	Socio-emotional	Socio-emotional	Socio-emotional	Socio-emotional	Socio-emotional	Socio-emotional	Socio-emotional	Socio-emotional
4					Item	If you play a game with your child, does he/she respond with interest? For example, if you play peek-a-boo, pat-a-cake, wave bye-bye, etc. does your child smile, widen their eyes, kick or move arms or	Does your child smile when you smile or talk with him/her?	Does your child smile, move excitedly, kick legs, move arms or trunk, or make coo noises when a known person enters the room or speaks to him/her?	Is your child interested when he/she sees other children playing? Does he/she watch, smile, or look excited?	Is your child very interested to play with a new toy or object?	Does your child usually get along with other children he/she plays with such as siblings or other family members? When your child spends time with other	When your child sees other children playing, does your child play next to them without joining in? Does he/she play alongside?	Does your child play together with other children not just by sitting next to each other and doing the same thing, but actually interacting?
5	ID No.	Age Range	Test	Domain	Item	Description							
298	293	36 - 42	GCDC	Receptive	Child can point to and identify at least 2 shapes		No match	No match	No match	No match	No match	No match	No match	No match
299	294	36 - 42	GCDC	Receptive	Child understands difference between more and less		No match	No match	No match	No match	No match	No match	No match	No match
300	295	>42	GCDC	Receptive	Understand 4 Prepositions		No match	No match	No match	No match	No match	No match	No match	No match
301														

Step 3.2 Statistical modelling

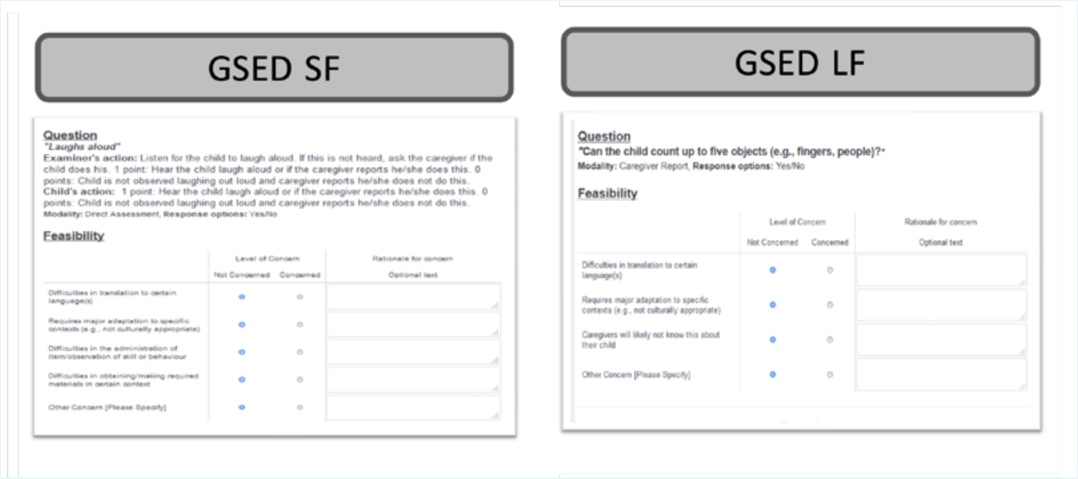
The Rasch model *(21)* was used to place children and items on the same scale. Rasch constructed a method to map and scale the responses on different, overlapping attainment tests given to different school classes. This method uses one parameter to quantify the child's ability at a given age and one parameter to quantify the difficulty of a test item. Rasch placed both parameters on the same scale and showed that this scale adheres to the principle of invariant comparison. Briefly, the principle implies that the comparison between two individuals, or between the same individual at two time points, should be independent of the test administered. The Rasch model has been highly influential in educational research and is slowly being adopted in health, agricultural and market research fields. The approach for child development data is very similar to Rasch's application in the education field, which makes it a natural fit *(22)*.

The difficulty parameters were estimated for a subset of 818 items that fitted the Rasch model as judged by infit and outfit[1] criteria using 17 equate groups[2] that anchored all instruments to the D-score scale. Scores extracted from the Rasch model were compared to scores from a more general two-parameter logistic (2PL) model, which uses one additional discrimination parameter per item. Extremely strong correlations between scores from the two models were observed (r = 0.97), and so the more parsimonious Rasch model was selected over the 2PL.[3]

Step 3.3 Item feasibility

The item feasibility component was designed to provide judgement data by SMEs on the appropriateness of each item for capturing development across various geographic, cultural and language contexts. The data were used to identify for removal any items that were in the item bank but were judged to be unsuitable for the diverse contexts in which the GSED instruments would be used (see Figure 4 for the survey on feasibility).

FIGURE 4. SURVEY ON FEASIBILITY INFORMATION



GSED SF

Question
"Laughs aloud"
Examiner's action: Listen for the child to laugh aloud. If this is not heard, ask the caregiver if the child does his. 1 point: Hear the child laugh aloud or if the caregiver reports he/she does this. 0 points: Child is not observed laughing out loud and caregiver reports he/she does not do this.
Child's action: 1 point: Hear the child laugh aloud or if the caregiver reports he/she does this. 0 points: Child is not observed laughing out loud and caregiver reports he/she does not do this.
Modality: Direct Assessment, **Response options:** Yes/No

Feasibility

	Level of Concern		Rationale for concern
	Not Concerned	Concerned	Optional text
Difficulties in translation to certain language(s)	●	○	
Requires major adaptation to specific contexts (e.g., not culturally appropriate)	●	○	
Difficulties in the administration of item/observation of skill or behaviour	●	○	
Difficulties in obtaining/making required materials in certain context	●	○	
Other Concern [Please Specify]	●	○	

GSED LF

Question
"Can the child count up to five objects (e.g., fingers, people)?"
Modality: Caregiver Report, **Response options:** Yes/No

Feasibility

	Level of Concern		Rationale for concern
	Not Concerned	Concerned	Optional text
Difficulties in translation to certain language(s)	●	○	
Requires major adaptation to specific contexts (e.g., not culturally appropriate)	●	○	
Caregivers will likely not know this about their child	●	○	
Other Concern [Please Specify]	●	○	

[1] Infit and outfit are "fit" statistics. In a Rasch context they indicate how accurately or predictably data fit the model.

[2] Equate refers to groups of items that were held to have a constant difficulty (tau) across tools, based on evidence-led analysis of item similarity. The equate groups permitted the different instruments used for different children to be on the same scale.

[3] For a deeper discussion on model choice, see the AERA 1992 debate between Hambleton and Wright (https://www.rasch.org/rmt/rmt61a.htm and https://www.rasch.org/rmt/rmt62d.htm).

Step 3.4 Item domaining

The domaining exercise was designed to provide expert judgement on precisely which domains of development each item measures. The taxonomy of domains was adapted from the Caregiver Reported Early Development Instruments (CREDI) project *(23)*. See Table 1 for the taxonomy of domains.

TABLE 1. TAXONOMY OF DEVELOPMENTAL DOMAINS

Primary domain	Sub-domain
Motor	• Gross • Fine
Language	• Receptive • Expressive • Problem-solving/reasoning
Cognition	• Executive function (e.g. attention, memory, inhibition) • Pre-academic knowledge (e.g. letters, numbers, colours)
Socio-emotional	• Emotional and behavioural self-regulation (e.g. controlling emotions and behaviours) • Emotion knowledge (e.g. identifying emotions) • Social competence (e.g. getting along with others) • Behaviour challenges/issues – internalizing (e.g. withdrawal) • Behaviour challenges/issues – externalizing (e.g. hitting, kicking, biting)
Adaptive	Life skills (e.g. using toilet, dressing)

© WHO / Sergey Volkov

Step 3.5 Item shortlisting

As a guiding principle and to avoid burden on respondents, two separate forms were created, one fully caregiver reported and one fully directly administered, that could potentially be combined for assessment depending on the target use. Additionally, for each of the forms, it was decided that it would be preferable to test more items during the validation process to allow a better-informed item selection after field-testing in multiple contexts and comparison of which administration modality (caregiver reported versus directly administered) would better capture specific skills and behaviours (to inform the final item selection for the additional CB, avoiding repetition).

Step 3.6 Creation of R Shiny App for information display

All items were initially collated in spreadsheets and then combined with the psychometric data available on each item from existing datasets and their estimated difficulties (taus) onto an interactive dashboard created via an R Shiny App *(24)* for interactive visualizations. Shiny is an R package that enables building interactive applications that execute R-code on the back end, for example stand-alone applications on a webpage with interactive charts or dynamic dashboards. The interactive dashboard facilitated the SMEs' final item selection, as it provided real-time evaluations of domain coverage, reliability statistics and measurement performance by age group for any selected items.

Step 3.7 Item selection for each measure

The creation of the GSED measures for validation comprised four steps, which were all processed in the interactive R Shiny App dashboard: i) selection of a single item from a group of matched items for measuring a given behaviour (which included choosing the item that was performing best across countries and wording of the item in the simplest and most culturally-neutral way possible); ii) allocation of items to GSED SF or GSED LF (and potential adaptation of the item from caregiver reported to directly administered and vice versa as needed); iii) evaluation of the age and domain coverage of the list of items selected for each measure; and iv) evaluation of the psychometric properties of the measure based on available data and simulations. SMEs iterated over item selection and evaluation until a suitable final set of items was identified to contribute to each form.

Items in the GSED SF were ordered by level of difficulty, based on data available on each item reflecting children's emerging skills. The order was slightly revised by SMEs to ensure it was consistent with child development theory. Items in the GSED LF were first grouped in three streams to facilitate administration flow during direct administration to the child and then ordered following the same process described for the GSED SF. Each stream uses specific materials to facilitate administration. Stream A includes items related to physical activity and movement, Stream B uses tablet-based images and Stream C uses materials from the GSED LF Kit (see below).

One hundred thirty-nine items were shortlisted for the GSED (caregiver response options of Yes, No and Don't know) and 155 items for GSEF LF that either observed incidentally or elicited by the assessor or both (response options are binary, skill observed or not). "Start" and "stop" rules are used for both forms' administration based on child's age and performance.

Step 3.8 GSED SF and LF measures finalization

Once the items and sequences were finalized for each form, the SMEs reviewed the items to identify where audio or visual prompts might be beneficial to facilitate training, caregiver understanding (for GSED SF) and administration of items for assessors (for GSED LF). All media files were created to accurately capture the skill or behaviour described in the items as well as to be as culturally neutral as possible. Lastly, GSED LF items were reviewed to identify those requiring physical materials for administration (e.g. stacking blocks, responding to rattle, drawing on paper). A detailed description of the materials needed to form the GSED LF Kit was created (including objects such as a timer, small 2 cm square blocks, a spoon, a plate, a cup, a crayon/pencil, etc.) and guided by the principle that the kit should be assembled locally and at a low cost. For the GSED LF items requiring children to look at images, efforts were made to include drawings and pictures that could be displayed on a tablet. The principle was to reduce the need for additional materials and printing and to streamline and standardize administration. The GSED measures were tested and revised during the training and feasibility phases of the validation study (see Section 4).

Step 4. GSED App development

To facilitate administration of the GSED measures, which require "start" and "stop" rules based on age and performance, limit data entry errors, and standardize data collection, an electronic data capture (EDC) system was designed. The EDC system uses Open Data Kit (ODK), a free and open-source software used widely for collecting, managing and using data in resource-constrained environments. The ODK Collect App (see Figure 5), an Android-based data collection app, was customized to create a grid-based user-friendly interface for administering the GSED LF. The GSED App was programmed to automatically determine the age-based starting question, and all the required "start" criteria and "stop" rules. For the GSED LF, the app enabled the assessors to score items during the direct observation of the child's performance while also providing administration instructions. The GSED measures were created using XLSForms which were then converted to ODK XForm. For the specific purpose of data collection in the validation sites (see Section 4), in addition to GSED measures, all other study data collection instruments were designed and incorporated in the same app. ODK aggregate with MySQL database was used as an aggregator. In addition, a separate dedicated data management and monitoring tool was designed enabling the study team to effectively manage, monitor and generate analysable output files in a standardized manner.

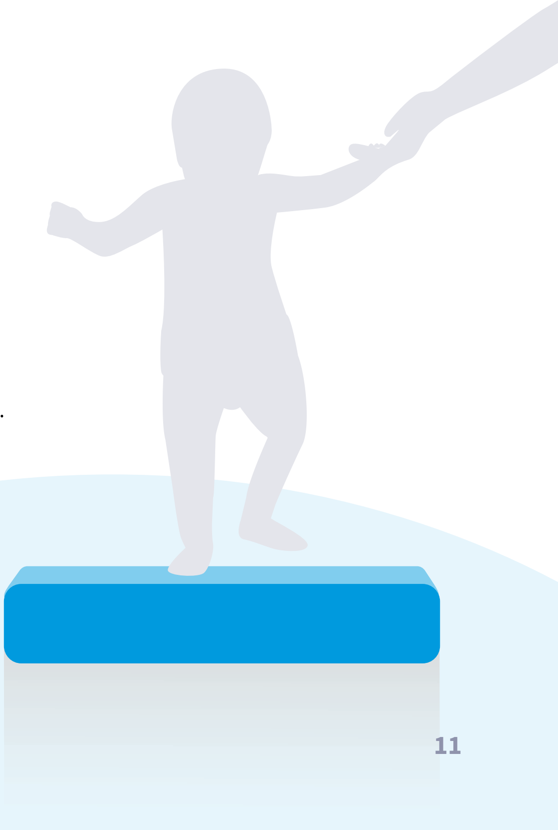
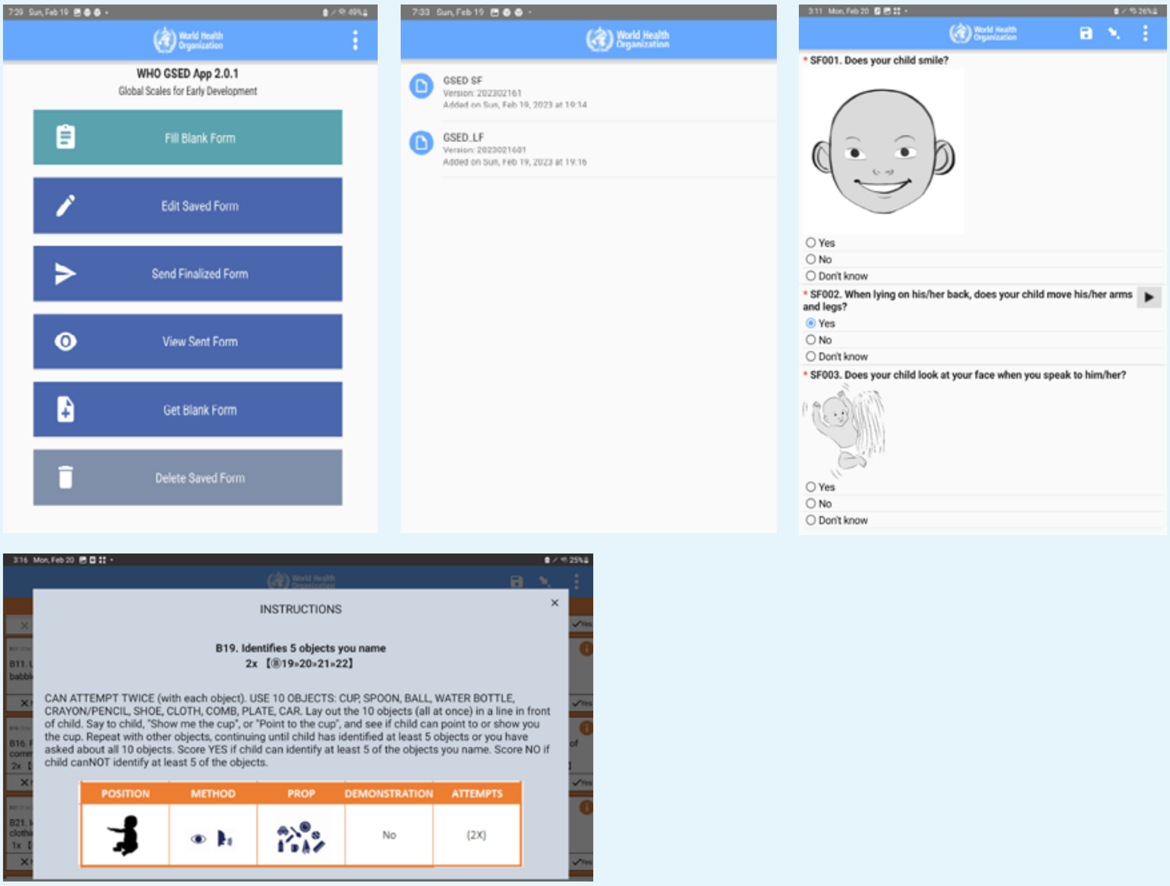


FIGURE 5. GSED APP

WHO GSED

The GSED App prototype was pre-tested through Google Play with two rounds of feedback. The key feedback received focused on the visual interface, colours and fonts, number of questions per screen, and the functioning of the administration rules as intended, as well as the ease of using the media files (GSED SF) and in-built administration instructions support (GSED LF). The GSED App was revised and tested in the feasibility phase of the validation (see Section 4).



WHO GSED App 2.0.1
Global Scales for Early Development

GSED SF
Version: 202302161
Added on Sun, Feb 19, 2023 at 19:14

GSED LF
Version: 2023021601
Added on Sun, Feb 19, 2023 at 19:16

* SF001. Does your child smile?
○ Yes
○ No
○ Don't know

* SF002. When lying on his/her back, does your child move his/her arms and legs?
● Yes
○ No
○ Don't know

* SF003. Does your child look at your face when you speak to him/her?
○ Yes
○ No
○ Don't know

INSTRUCTIONS

B19. Identifies 5 objects you name
2x 【®19»20»21»22】

CAN ATTEMPT TWICE (with each object). USE 10 OBJECTS: CUP, SPOON, BALL, WATER BOTTLE, CRAYON/PENCIL, SHOE, CLOTH, COMB, PLATE, CAR. Lay out the 10 objects (all at once) in a line in front of child. Say to child, "Show me the cup", or "Point to the cup", and see if child can point to or show you the cup. Repeat with other objects, continuing until child has identified at least 5 objects or you have asked about all 10 objects. Score YES if child can identify at least 5 of the objects you name. Score NO if child canNOT identify at least 5 of the objects.

POSITION	METHOD	PROP	DEMONSTRATION	ATTEMPTS
			No	(2X)

4. GSED validation

This section describes the methodology undertaken for the validation of the GSED measures. The GSED validation study was planned in seven countries varying in geography, language, culture and income, and implemented in two rounds (since funding was received in two stages): Round 1 included Bangladesh, Pakistan and the United Republic of Tanzania (validation completed and results informed this technical report); and Round 2 including Brazil, China, Côte d'Ivoire and the Netherlands (data collection for the validation ongoing). Data from Round 2 countries will further expand the generalizability of the results for global use and is expected to be published in early 2024. Figure 6 shows the GSED validation partners in each country.

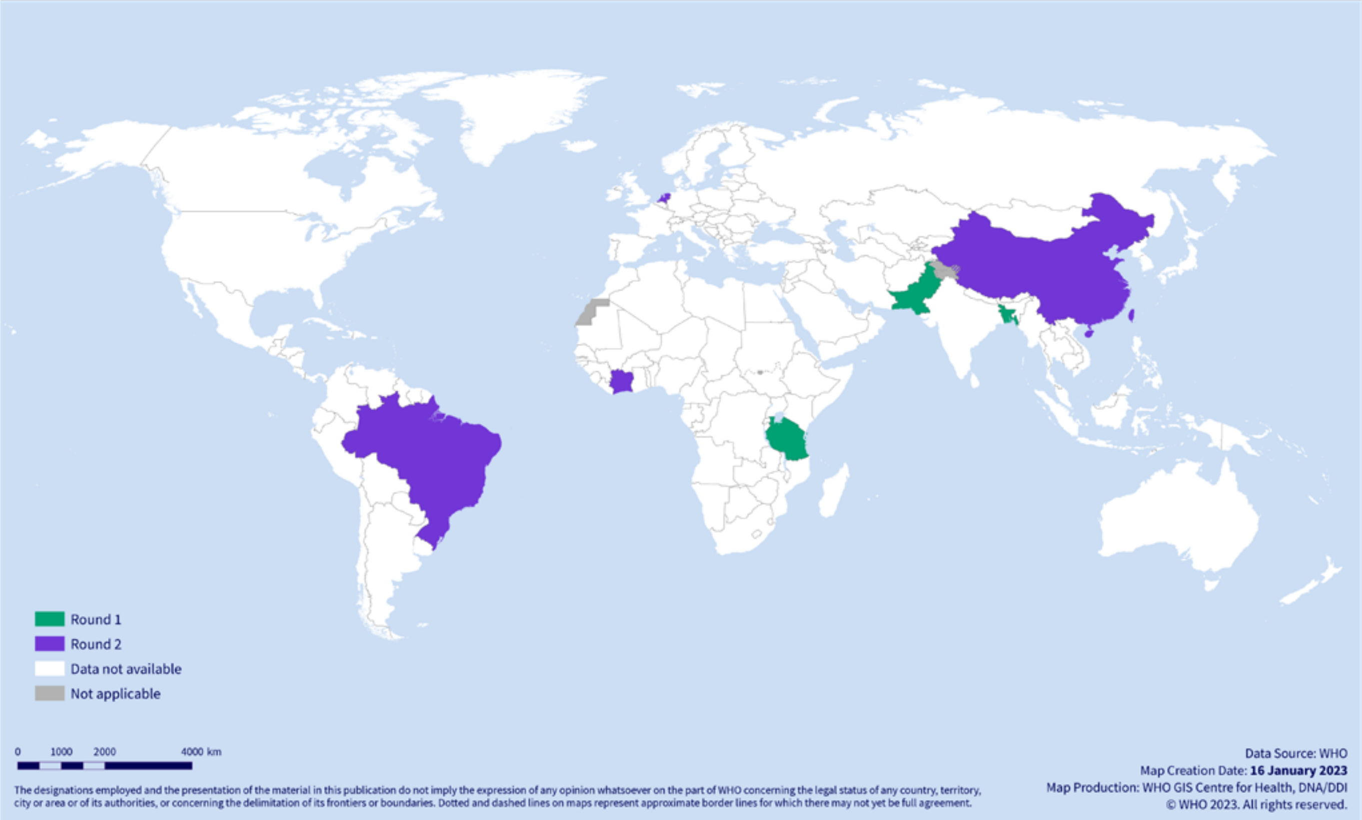
The study utilized a rigorously standardized protocol across countries with a mixed qualitative and quantitative methods approach combining cross-sectional and longitudinal approaches to evaluate the psychometric properties of the GSED SF and LF: reliability (inter-rater and test-retest), concurrent validity, convergent validity, short-term predictive validity at six months and responsiveness (*25*). This study received ethical approval from WHO (protocol GSED validation 004583 20.04.2020) and approval in each site.

The work included a preparation and feasibility phase and the main validation data collection phase. The methodology of both phases is described below. The results presented in this report are limited to Round 1 countries where data collection has been completed.

In addition to the main validation effort, external research groups have proposed supporting generation of further GSED validation data through the inclusion of the GSED package as secondary research outcomes in ongoing studies (see Box 1).

© WHO / NOOR / Sebastian Liste

FIGURE 6. GSED VALIDATION STUDY SITES AND PARTNERS



The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of WHO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

Data Source: WHO
Map Creation Date: **16 January 2023**
Map Production: WHO GIS Centre for Health, DNA/DDI
© WHO 2023. All rights reserved.

	Country	Partner Institution
Round 1 [completed]	Bangladesh	Projahnmo Research Foundation, Dhaka, Bangladesh Johns Hopkins Bloomberg School of Public Health, Baltimore, USA
	Pakistan	AKU, Karachi, Pakistan
	United Republic of Tanzania	CPHK Global, Zanzibar, United Republic of Tanzania Public Health Laboratory – Ivo de Carneri Pemba, Zanzibar, United Republic of Tanzania
Round 2 [ongoing]	Brazil	University of São Paulo Medical School, São Paulo, Brazil
	China	National Children’s Medical Center/Shanghai Children’s Medical Center, Shanghai, China
	Côte d’Ivoire	IPA, Abidjan, Côte d’Ivoire
	Netherlands	TNO, Sylviusweg, Leiden, the Netherlands

BOX 1. TESTING OF GSED THROUGH ONGOING STUDIES

The inclusion of GSED in external studies is a form of testing the GSED measures (either SF and/or LF and/or PF) at sites of ongoing or newly-launched research projects with the aim of producing further validation and feasibility evidence on GSED that is otherwise not covered by WHO's main validation efforts. This approach provides an opportunity to leverage existing resources by partnering with external research groups to access a wider range of countries and populations, and to broaden the range of entry points and settings for administering the GSED. Specifically, this work aims at generating additional GSED validation and feasibility data for:

i. global relevance (increasing generalizability);

ii. sensitivity to change (programmatic evaluation);

iii. biological markers convergence;

iv. predictive validity;

v. feasibility of ongoing survey integration;

vi. developmental trajectories exploration.

Depending on the original study design and population, the additional data generated contribute to one or more of the objectives above. Through the inclusion of GSED in external research protocols, additional evidence is expected to be collected for the use of GSED for specific purposes in diverse contexts that should increase its global uptake.

Preparation and feasibility phase

The preparation and feasibility phase included translation and local adaptation of study measures, defining and writing standard operating procedures, recruitment and training of local teams, and testing of implementation processes to ensure feasibility. This section describes the activities conducted as well as the results and related decisions for main validation.

Translations and adaptation

The GSED measures followed a rigorous and standardized translation and adaptation process to capture linguistic and cultural nuances, without losing the essential conceptual focus of the items and forms. In brief, each GSED measure was translated into the local languages (Bangla for Bangladesh, Urdu and Sindhi for Pakistan and Swahili for the United Republic of Tanzania), by two independent and qualified translators. These translations were reviewed by the local study team to agree on a translation by consensus. This agreed-upon local language translation was then back-translated into English by two other independent

translators. This back-translation was then shared with the WHO team, who initiated an iterative process of revisions with the local team before approval. During the preparation and feasibility phase the approved translations were tested (see below) so that final modifications, if necessary, could be made before data collection. This process led to adaptation of the GSED measures to include cultural nuances for language and context; few items were revised with minor rephrasing in each country.

Training

One joint in-person training of trainers course was conducted for the study teams. The training covered implementation processes and all data collection forms. The training included workshops on child development principles, detailed review of administration processes and item-by-item review. It included practical sessions with live demonstrations, role-plays, and practice in pairs and with both caregivers and children. Participants in the training were certified as local master trainers and were then responsible for training the field assessors in their country. For the local assessors' training, a structured two-week training programme was designed by the local master trainers, in consultation with the WHO team, with thematic/didactic sessions in classrooms and practice sessions. To be able to collect data, field assessors completed a certification process, which entailed achieving an agreement of 90% on the forms' scoring between the assessor and the local master trainer while administering the GSED forms.

Feasibility of implementation processes

Data were collected from a minimum of 32 children per country, stratified by age group and sex, to test feasibility and acceptability of administration of the study instruments, as well as to finalize implementation processes, such as visit schedules including time required, and the use of tablets. The sample adequately enabled all implementation procedures to be tested such as for reliability testing processes, checks carried out on completeness of data collection and whether any items from the measures, both GSED and contextual, had a significant number of missing or otherwise invalid responses in any of the countries. Some of the key findings from the preparatory and feasibility phase are provided below.

Visit schedules Data were collected in two visits (in Bangladesh both a one-visit and two-visit option were tested) with the sample divided into subgroups to test the feasibility of different visit schedules in terms of order of administration of study measures and settings for the visits. The first visit was done at home to test the GSED SF in the setting for which it was intended for future use (i.e. household surveys) and to administer the study measures (see Annex 2) aimed at capturing the child's everyday home living environment. In Pakistan and the United Republic of Tanzania, the second visit was carried out in a mobile clinic/clinic setting to facilitate anthropometric testing and standardize the setting for concurrent validation (with Bayley-III). In Bangladesh, the second visit was done at home due to transportation and travel times between the sites. This process worked well and was confirmed for the validation. The two-visit model was also confirmed in all countries, including Bangladesh, due to time needed for administration. Fine-tuning of the sequence in which the study measures were administered took into account how to optimize the privacy needed for some of them.

Reliability testing Two reliability testing processes were carried out in a subsample of a minimum of 16 children per country to explore the feasibility of conducting multiple administrations of study measures within fixed time frames. First, for GSED LF, both a video recording approach (Bangladesh N=15 and Pakistan N=12) of the GSED LF assessment was tested (with videos watched and independently scored by other assessors and the study master trainers as well as a simultaneous coding by another assessor (N=8 per country). For GSED SF (and GSED PF) audio recording (Bangladesh and Pakistan, N=16) the administration was carried out and independently scored by other assessors on the study team. In Pemba, United Republic of Tanzania, the reliability testing was completed via live coding of the administration of the test by another assessor (GSED SF N=22, GSED LF N=23]. Several limitations were encountered with the video and audio recordings. As the camera was placed in a fixed location for the video recordings, the motor component of the GSED LF assessment was difficult to capture consistently and be coded. In addition, some sites faced issues with providing sufficient lighting during recording to enable later coding. The audio recording method for GSED SF posed a challenge pertaining to the quality of the voice recordings. The in-person approach with an independent assessor coding the observations simultaneously with another assessor administering the GSED measures was found to be feasible and of higher quality for inter-rater reliability and was therefore implemented in all sites during the validation phase.

Quality control Ten per cent of the scheduled visits of each assessor in each site were randomly selected for a quality control visit by the master trainer or local supervisor. During these visits, the master trainer or supervisor completed a checklist for the administration of the tests that included verifying the child's and mother's ages, date of birth, consent, rapport building and accuracy in administration of the study measures. In this testing phase, at least 10% of the total visits at each site were also video recorded. The direct administration approach for quality control was found to be feasible and reliable as compared to video assessment, due to challenges with the video recording mentioned above.

Qualitative data - exit interviews During the feasibility phase, exit interviews with caregivers were conducted to understand the caregiver experience in the consent process, the acceptability of GSED administration, visit schedules (N=63) and the acceptability of the GSED LF items, materials, instructions and procedures (N=72). The assessors recorded the responses and any narrative comments by caregivers on paper, and these were later translated into English. From these exit interviews, it emerged that most (> 90%) respondents said that the various aspects of the implementation process (e.g. comfort with items asked, where and when they were asked to respond to items) were acceptable. However, 14% of respondents said that the duration of the visit was very long. The duration of the total interview time was expected to decrease with familiarity with the study measures; nonetheless, the training process was reinforced with a focus on supervised practice to ensure that data collectors were sufficiently comfortable with and efficient about administration of the study measures. Additionally, emphasis was placed on further clarifying the time commitment for study participation at the consent stage and deciding to implement data collection in two separate visits at all sites. Some concerns were raised by caregivers and assessors about the sensitivity of some of the questions (see Annex 2 for study measures), which were addressed by ensuring adequate privacy for conversations with the caregiver (e.g. arranged at a health clinic or outside of the home, or during a time when a private setting could be found at home). In the GSED LF exit interviews, some parents said the skills tested (28%) and materials administered (43%) were unfamiliar to children in their community. Based on the preliminary quantitative analyses the items in question (e.g. those including blocks, shape board, peg board) seemed to perform well and were kept without significant change. Moreover, 15% of respondents also specifically offered positive comments about their experiences with the GSED LF, such as learning what their children could do and the need for more education on ECD.

Qualitative data – focus group discussions (FGDs) At the end of the feasibility phase, virtual structured FGDs with the assessors, supervisors and study managers from each country were conducted (N=42). The purpose was to elicit local field team feedback on implementation processes (consent process, ease of administration of study measures, visit schedules, use of the GSED App, training needs, comprehension of the items by assessors and caregivers), and cultural appropriateness of GSED SF and GSED LF items and the GSED LF Kit materials. FGDs were audio-recorded, transcribed and translated by local staff with specialized training in qualitative methods. These data were analysed using Dedoose *(26)*. Codes were created for the items included in the FGDs and applied to the responses from participants. Themes were identified for summary analysis. Overall, the results indicated that the assessors‘ experiences with the administration of the study measures were positive. The main concerns expressed were, consistent with caregivers‘ experiences shared in exit interviews, about ensuring privacy for measures with sensitive items (such as caregiver depression or exposure to violence, see Annex 2). The FGDs also prompted changes in how to introduce the GSED SF (e.g. more information to be provided about the fact that, by design, the validation of the measures implies that some items may seem repetitive to the respondent) and tips for optimizing administration of the GSED LF (e.g. offering materials, such as blocks, to children to familiarize them with the objects prior to administration). The FGDs were also useful for teams to strategize how to manage implementation challenges during GSED LF administration (e.g. keeping children engaged, avoiding distractions, etc.).

Based on the exit interviews and FGDs, along with feedback from the translation process, 13 GSED LF items were identified as difficult to administer or understand. These items were adapted to facilitate administration, or rephrased to improve ease of translation and/or clarity of administration instructions. For example, the GSED LF item to assess the child's understanding of the concepts of "more" and "less" initially asked the child to indicate which of two cups contained more water. This item was reportedly hard for children to understand and complete as some children wanted to play with the water or drink from the cups. This item was adapted by asking the child to indicate which of two piles of blocks had more blocks (in place of the cups with water).

Validation phase

Study population

The study population included children 2 weeks to 41 months of age (inclusive) living in the study areas. Children up to 41 months were included to ensure that measure parameters could be estimated with adequate precision at 36 months. Inclusion criteria were the child's age, the child's primary caregiver (person most familiar with the child and spending most time with him/her) was available to participate in the study, and the family spoke to the child in one of the GSED translated languages. Children were excluded if gestational age or birth weight data were missing.

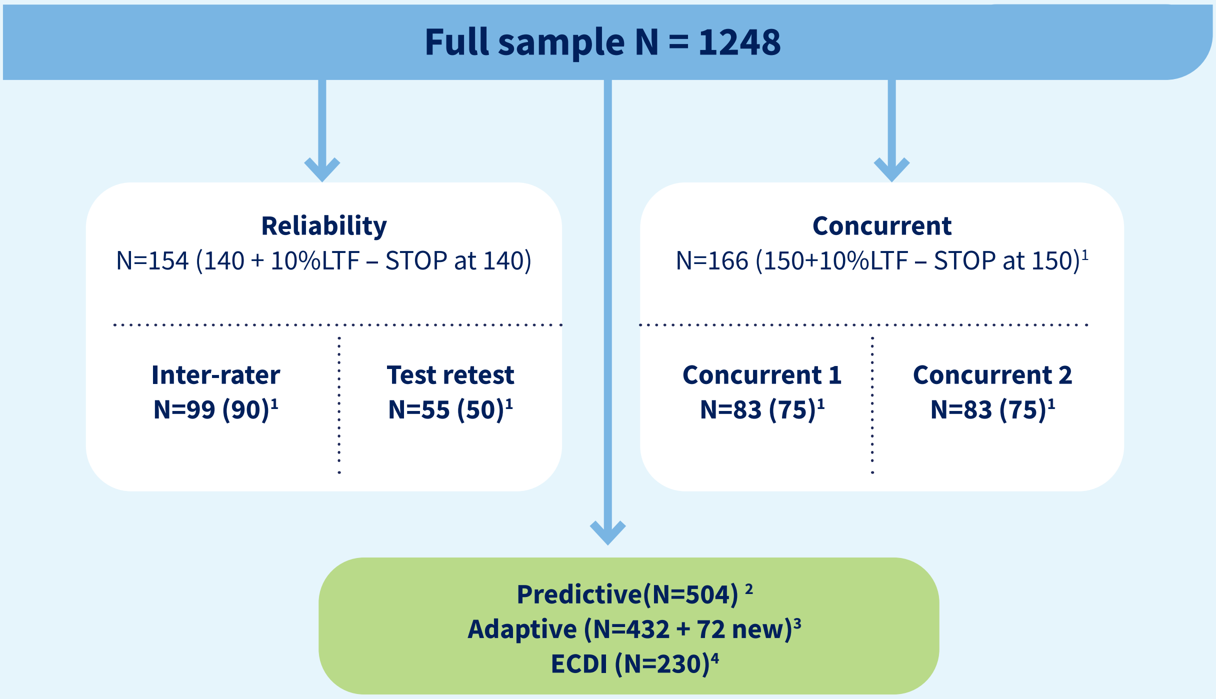
A subsample of children was used to estimate preliminary reference scores and is henceforth referred to as the "reference" sample. Additional exclusion criteria were applied to this subsample to exclude children who were low birth weight (< 2500 g); born preterm or late term (gestational age < 37 or ≥ 42 completed weeks); undernourished (weight-for-age, length-for-age or weight-for-height z-score < –2 SD based on the WHO Child Growth Standards) at the time of developmental assessment; had a known severe congenital birth defect, history of birth asphyxia or neonatal sepsis requiring hospitalization, known neurodevelopmental disorder or disability, or other chronic health problem; or primary caregiver had less than secondary level education.

Sample size and sampling scheme

The target sample size per site was 1248 children (total 3744 in the three countries). In each site, children were sampled from a list of potentially eligible caregiver-child dyads residing in the defined study areas. Only one child per caregiver or multi-family household was selected and the target children's primary caregiver approached for consent and enrolment. Children who were acutely unwell at the time of assessment were rescheduled after seven days. Refusals to participate and drop-outs were registered and replaced. After consent was obtained, children were allocated to sex and age groups using the sampling scheme in Figure 7. Larger quotas were set for the youngest age groups where rates of development are steepest. Out of the full sample of 1248 children per site, 154 were randomly allocated to either inter-rater (N=99) or test-retest (N=55) reliability testing; 166 to concurrent validity testing with the order of GSED and Bayley-III administrations counter-balanced (N=83 with GSED administered first and vice versa); and 504 to re-evaluation six months later for predictive validity.

Within the predictive validity subsample, children were further divided into groups that also received the GSED adaptive measure to determine whether adaptive testing is a feasible and valid option to measure child development within the GSED (see Box 2 for further details), or the UNICEF's Early Childhood Development Index 2030 (ECDI2030) measure to inform harmonization of measurement of child development up to 5 years at population level (see Box 3 for further details).

FIGURE 7. GSED SAMPLING SCHEME IN EACH COUNTRY



[1] The number inside parentheses is the number collected and the number outside is the number randomized to account for loss to follow-up.

[2] Two additional participants have been added to the predictive to have equal numbers in each experimental group.

[3] 72 new children between 2 weeks and 6 months of age have been added to the adaptive sample to ensure coverage at the lower ages.

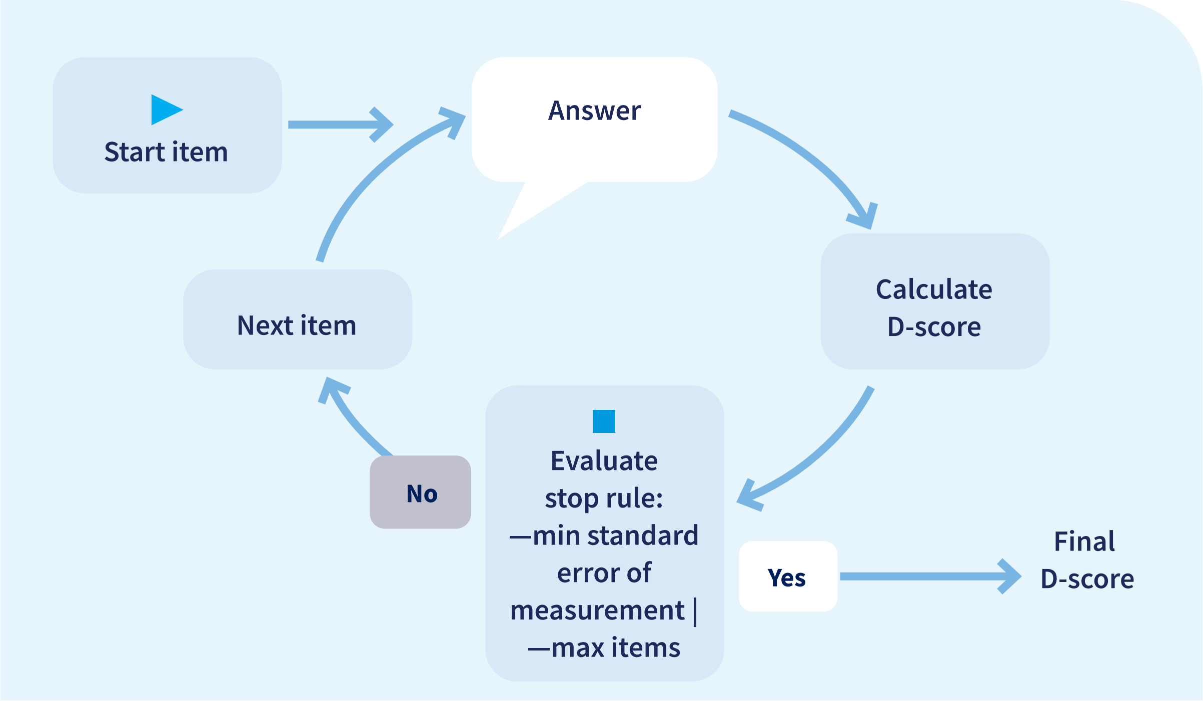
[4] ECDI was only done on N=230 children of 24 months and above at the time of the predictive data collection.

© WHO / NOOR / Sebastian Liste

BOX 2. GSED ADAPTIVE TESTING

Computerized adaptive tests (CAT) are widely used in education. They tailor the order of administration for each participant based on the participant's responses to the prior questions. The GSED adaptive test selects a first item (i.e. starting item) based on an external estimate of the child's proficiency, for example, based on age. After recording the response to that item, the underlying algorithm calculates the D-score and determines whether the result is precise enough to meet the "stop" criterion. If this is not met, the algorithm continues to select the next item, such that it provides maximal information given all previous responses, and the cycle continues. The process ends once the predefined "stop" criterion is satisfied. **Figure 8** visualizes the algorithm. Adaptive testing is a modern, flexible and personalized method to quantify each child's D-score. As yet, there is little published research and experience utilizing adaptive testing for measuring child development, beyond the efficiency of adaptive testing in simulation studies *(27,28)* or clinical settings *(27)*. The adaptive testing approach for GSED SF and GSED LF has been tested within the GSED validation study and the results are under review.

FIGURE 8. CAT ALGORITHM FOR THE D-SCORE


Start item
Answer
Calculate D-score
Next item
No
Evaluate stop rule: —min standard error of measurement | —max items
Yes
Final D-score

BOX 3. LINKING GSED AND ECDI2030

Population-level assessments of ECD outcomes allow countries to track their progress in achieving SDGs while also generating data for advocacy and policy purposes. One of the most relevant for ECD is SDG indicator 4.2.1 (the proportion of children aged 24 - 59 months who are developmentally on track in health, learning and psychosocial well-being). UNICEF's ECDI2030 is recognized as a viable measure for assessing progress towards this target. However, research indicates that the first 1000 days of life provides an important window of opportunity to build a strong foundation for future development. Therefore, an open question remains of how to harmoniously and meaningfully track the development of children starting from birth to 2 years, when inequities in health and education begin. Establishing the feasibility of connecting population-level monitoring in the first 24 months of life with that of older children and improving the evidence-building capacity in this area (e.g. providing metrics that allow for comparison of programme/policy impact across the first 5 years of life) is clearly imperative for advancing the child well-being and equity agenda beyond 2030. Considering the urgency to address the current gap in measurement of ECD in children less than 2 years of age, WHO and UNICEF advanced the dialogue on how to align the GSED and ECDI2030, to enable governments and partners to plan for continuity in measurement of ECD across the 0 – 59 months age range, in an interpretable and psychometrically valid manner.

To explore this objective and given the complementarity, scope and format of the two tools, leveraging the ongoing validation efforts, WHO collected data on 628 children aged 24 – 41 months across three countries where both GSED SF and ECDI2030 were administered in their entirety for the same child. The specific aim of the work was to explore how scores from the two measures link together and relate to one another so that continuity can be established in measuring children's development from birth. The results from this work will be published separately.

Study measures

The GSED SF, GSED LF (and GSED PF) measures, as well as measures of children's growth (anthropometry) and nutrition, health, environmental and contextual information such as family and home environment through Home Observation Measurement of the Environment (HOME) or Family Care Indicator (FCI), Childhood Psychosocial Adversity Scale (CPAS), Patient Health Questionnaire-9 (PHQ9), Brief Resilience Scale (BRS) and Family Support Scale (FSS) (see Annex 3) were administered to all children to examine convergent and discriminant validity and to identify a subsample of children with minimal constraints on their development in order to create preliminary reference scores. The selection of measures was based on known biological and social determinants of development (*29*), the demonstrated validity of the contextual measures in at least one low- and middle-income country, and efficiency for data collection. While there is no universal gold standard that can be recommended for concurrent validity testing, the Bayley-III measures a similar construct to the GSED and had previously been used by the country teams. The Bayley-III was therefore administered in the concurrent validity subsample and the ECDI2030 (see Box 3) to inform harmonization of measurement beyond 3 years in a subgroup of the predictive validity subsample.

Visit schedule and quality control

Data collection was scheduled over one to three visits depending on the study site and subsample. The first administration of the GSED SF was completed at home to test it in the setting intended for future use (e.g. household surveys) and prior to administration of the GSED LF to avoid influencing caregiver responses. The GSED LF and Bayley-III were administered in a controlled environment (e.g. clinic or quiet residential area) to match the required testing protocols of the concurrent validity measure. To ensure high quality, 10% of all study visits were observed in person by study supervisors, covering each child age band and certified assessor. The supervisors had at least five years of experience in community-based research and/or formal education in fields related to ECD (e.g. teaching, nursing, psychology). Supervisors independently completed questionnaires administered by the assessor and completed a fidelity checklist to provide feedback to assessors. Supervisors reviewed quality assurance findings with WHO biweekly. The GSED App for data collection provided built-in data range and consistency checks. Data managers reviewed and resolved issues daily in consultation with the local field and/or WHO team.

Data management

As described above, to optimize ease of administration of the GSED measures and minimize data entry errors, the GSED App was designed which also improved standardization of data collection across study sites. In addition to GSED measures, all other study measures were designed and incorporated in the same app. ODK aggregate with MySQL database was used as an aggregator. Lastly, a separate data management and monitoring module was designed enabling the study team to effectively manage, monitor and generate analysable output files. The module for data management allowed data managers to check for the completion status of the forms, flagged missing data, status of the visit schedule and the visit windows for the participants for scheduling, data collection status for the study age and gender bins and completion status of the participants in the study. The data from the module were reviewed weekly by the country teams and also with WHO for study status monitoring.

The prototype was pre-tested through Google Play with two rounds of feedback by the field teams, SMEs and the statistics teams. The key feedback received focused on the visual interface, colours and fonts, number of questions per screen, and the functioning of the administration rules as intended, as well as the facility of use of the media files (GSED SF) and in-built administration instructions support (GSED LF). The GSED App was then revised and field-tested in the feasibility phase of the validation for ease and accuracy of data collection and transfer. Following the feasibility phase, the revised GSED App version was released and tested for the following features: placement of media files and questions on screen for improved speed in using the App, inclusion of a pop-up screen to inform the assessor the age of the child to be assessed, and inclusion of a pop-up asking if all questions have been completed for the child's age as per the rules prior to saving the forms.

Once collected, the data were stored in local password-protected user authenticated servers. The de-identified data were securely transferred to the WHO central data repository by each site. They were transferred weekly for the first month of data collection and then bi-monthly. The weekly data collected in the first month were reviewed for consistency with the data dictionary, checks for missing data, data formatting and diagnosing any potential problems (missing or non-sensical data). Teams maintained detailed logs related to procedures for rescheduling or incomplete visits. These were reviewed weekly for the resolution of queries with the WHO team. The final data set was transmitted by each country to the WHO repository for analysis.

5. GSED psychometric properties

This section addresses various aspects of the GSED's psychometric performance in the three Round 1 countries (Bangladesh, Pakistan and the United Republic of Tanzania). The analyses in this section, with the exception of the analysis of reliability, are performed on the combined measure, i.e. the GSED SF and GSED LF data together. This is to reflect that primarily the scale from which the measures are drawn is being validated, rather than the individual tools. However, in order to show the limited differences between the psychometric properties of the GSED LF versus the GSED SF versus the CB, concurrent, convergent and short-term predictive validity for all three forms of the measure are presented in Annex 3.

In total, data were collected on 4452 children across the three countries. Some countries collected more data than the specified sample size in order to: i) meet the minimum quota of the reference sample; and ii) ensure every age group stratum was sampled to the specified level. Data were analysed on 4349 children, with randomly selected children contributing to the various reliability and validity subsamples. Table 2 shows the numbers collected for each of the sites by measure. Data from 41 children were removed from the analysis based on notes provided by the countries that the data were invalid for various reasons (e.g. duplicate entries, withdrew from the study, etc.), and 62 participants were removed as they had neither GSED LF nor GSED SF data available at baseline due to incomplete administration of the battery of tools.

TABLE 2. NUMBERS OF CHILDREN BY STUDY MEASURE COLLECTED AND BY COUNTRY

	Bangladesh	Pakistan	United Republic of Tanzania	Total
GSED SF sample	1336	1663	1350	**4349**
GSED LF sample	1332	1642	1344	**4318**
Test-retest reliability GSED SF subsample	48	59	52	**159**
Test-retest reliability GSED LF subsample	48	59	52	**159**
Inter-rater reliability GSED SF subsample	95	100	96	**291**
Inter-rater reliability GSED LF subsample	95	99	95	**289**
Predictive validity subsample	472	455	469	**1396**
Concurrent validity subsample (Bayley-III)	159	158	161	**478**

Table 3 presents the demographic characteristics of the sample. The mean age of the children is higher in Pakistan than in the other countries because more older children were recruited to ensure a sufficient sample for the predictive validity subsample after 6-month follow-up challenges. The large sample size, relative to similar studies, and the standardized implementation of the tools across multiple countries lends strength to the validity inferences and robustness of the results of the study.

TABLE 3. DEMOGRAPHIC CHARACTERISTICS BY COUNTRY

	Bangladesh	Pakistan	United Republic of Tanzania	Total
Male (%)	50	50	50	**50**
Mean age in days (SD)	432 (375)	475 (381)	432 (377)	**448 (378)**
Mean age in months (SD)	14.20 (12.32)	15.61 (12.52)	14.19 (12.38)	**14.72 (12.42)**
GSED DAZ (SD)	0.24 (0.75)	-0.34 (0.85)	0.07 (0.81)	**-0.03 (0.84)**
Gestational age – weeks (SD)	38.67 (1.71)	38.41 (2.00)	38.71 (1.76)	**38.59 (1.85)**
Birthweight - grams (SD)	2921 (422)	3251 (723)	3226 (513)	**3143 (599)**
Anthropometry – HAZ (SD)	-1.30 (1.09)	-1.33 (1.17)	-1.27 (1.12)	**-1.30 (1.13)**
Anthropometry - WAZ (SD)	-1.09 (1.08)	-1.45(1.02)	-0.77 (1.07)	**-1.13 (1.09)**
Maternal education – N (%)				
No schooling	25 (2%)	517 (32%)	163 (12%)	**705**
Primary	299 (22%)	224 (14%)	343 (25%)	**866**
Secondary	735 (55%)	358 (22%)	795 (59%)	**1888**
Post-secondary	277 (21%)	528 (32%)	49 (4%)	**854**
PHQ9 – N (%)*				
Minimal	587 (44%)	1215 (74%)	808 (61%)	**2610**
Mild	437 (33%)	220 (12%)	422 (32%)	**1079**
Moderate	121 (9%)	89 (5%)	52 (4%)	**262**
Moderate-severe	84 (6%)	52 (3%)	27 (2%)	**163**
Severe	103 (8%)	65 (4%)	19 (1%)	**187**
HOME stimulation score (SD)	40.63 (3.77)	38.27 (5.44)	38.70 (3.81)	**39.13 (4.61)**

* For more information on definitions of these categories, see https://www.med.umich.edu/1info/FHP/practiceguides/depress/score.pdf.

Internal reliability

The precision of the CB at any given age is greater than that of the individual forms from which it is constituted because of the larger number of items. Figure 9 gives the standard error of estimation (SEE) *(30)* of the D-scores obtained for the combined GSED SF and GSED LF under the Rasch model. As there are varying numbers of items pertinent to different sections of the scale, the precision of an estimate can vary at different points along the scale. The y-axis gives the SEE, the bottom x-axis gives the D-score scale, and the top x-axis gives the average D-score for various child ages, expressed in months. Note that the average D-scores for any given age are non-linearly related to age, reflecting the decreasing rate of development seen as children become older.

Figure 10 shows the pseudo-reliability [*Reliability*=1−(1/Test information) *(31)*] along the D-score scale. This value, at any point on the scale, can be interpreted in the same way as traditional measures of internal reliability (e.g. Kuder-Richardson 20 or Cronbach's alpha), which indicate the consistency of scores across items. For the GSED LF the internal reliability is above 0.8 for almost all of the scale, for the GSED SF it is above 0.8, and for the GSED CB above 0.9 for almost all of the scale. Lower reliability at a point on the scale indicates a lower density of items at a given point.

FIGURE 9. SEE PLOT BY GSED FORM

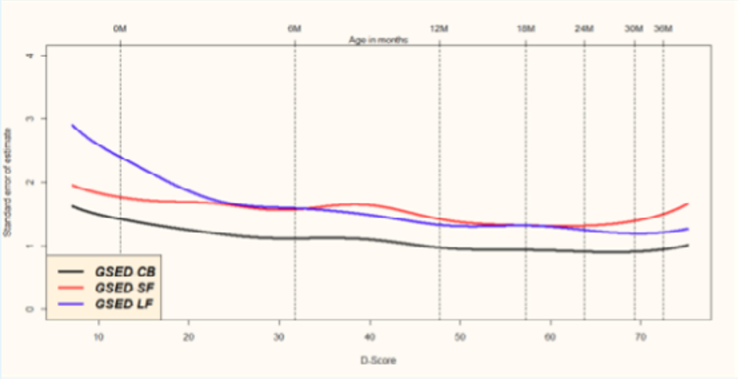
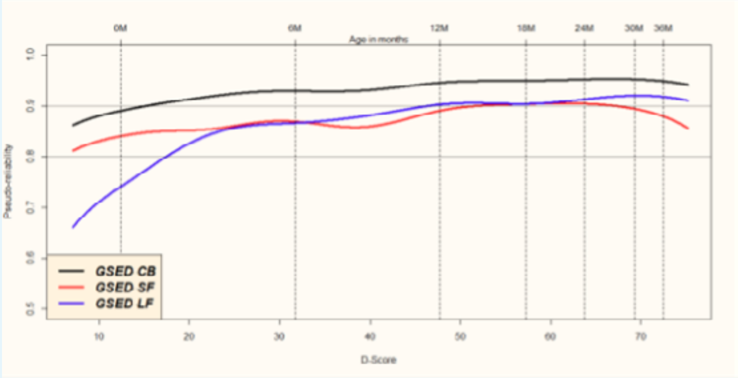
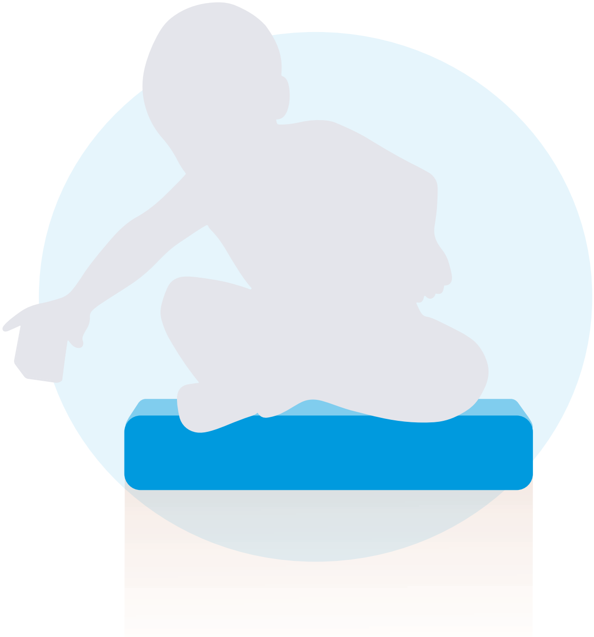


FIGURE 10. PSEUDO-RELIABILITY PLOT BY GSED FORM





External reliability

All the reliability metrics are excellent on the D-score scale. External reliability is the extent to which a measure produces the same score over theoretically identical administrations. Inter-rater reliability is the measurement of agreement between different raters and test-retest reliability is the measurement of agreement for the same rater at two different time points. Reliability is expressed here (see Table 4) as an intraclass correlation coefficient (ICC), where a value of 0 represents no reliability and a value of 1 represents perfect reliability. Common benchmark values suggest that 0 - 0.2 indicates poor reliability; 0.2 - 0.4 indicates fair reliability; 0.4 - 0.6 indicates moderate reliability; and values > 0.8 indicate excellent reliability (*32*).

TABLE 4. INTRACLASS CORRELATIONS FOR RELIABILITY OF GSED D-SCORE (95% CONFIDENCE INTERVALS [CIs])

		D-score scale			
		Bangladesh	**Pakistan**	**United Republic of Tanzania**	**Total**
Inter-rater reliability	**GSED CB**	0.99 (0.99-1.00) N=95	0.98 (0.97-0.99) N=99	0.99 (0.98-0.99) N=95	**0.99 (0.98-0.99)** **N=289**
	GSED SF	0.99 (0.99-0.99) N=95	0.97 (0.96-0.98) N=100	0.99 (0.99-0.99) N=96	**0.99 (0.98-0.99)** **N=291**
	GSED LF	0.99 (0.98-0.99) N=95	0.98 (0.96-0.98) N=99	0.97 (0.95-0.99) N=95	**0.98 (0.97-0.99)** **N=289**
Test-retest	**GSED CB**	1.00 (0.98-1.00) N=48	0.99 (0.98-0.99) N=59	0.99 (0.99-1.00) N=52	**0.99 (0.99-1.00)** **N=159**
	GSED SF	0.99 (0.99-1.00) N=48	0.99 (0.98-1.00) N=59	0.99 (0.99-1.00) N=52	**0.99 (0.99-1.00)** **N=159**
	GSED LF	0.99 (0.97-1.00) N=48	0.98 (0.96-0.99) N=59	0.99 (0.98-1.00) N=52	**0.99 (0.98-0.99)** **N=159**

> “A value of 0 represents no reliability and a value of 1 represents perfect reliability.”

Concurrent validity

Concurrent validity, a type of criterion validity, is the extent to which a measure correlates with another measure of the same construct, possibly a gold standard, given at the same time. Here the criterion measure is the Bayley-III, a widely-used measure that frequently acts as a reference *(33-35)*. To assess the correlation of the Bayley-III raw scores with GSED DAZ on the same scale, a 2PL item response theory (IRT) model was fitted to the Bayley-III item responses and a Generalized Additive Models for Location Scale and Shape (GAMLSS) model *(36)* was used to remove the effect of age, in line with the methodology used to construct the GSED DAZ. Sample-specific norms were constructed to ensure that the age adjustments were pertinent to the specific sample, as far as possible. Age-adjusted z-scores were only generated for the total score, as the individual domains did not contain enough items to calculate the age standardization robustly.

Table 5 gives the Pearson's correlations between the Bayley-III individual domains and total scores, with the GSED D-scores. The GSED D-score correlates > 0.90 with the domains of the Bayley-III in most domains and countries. The correlations are higher for the cognitive and motor items than for the communication items, although in the total score the correlation is very high (0.98).

TABLE 5. CONCURRENT VALIDITY FOR GSED BY BAYLEY-III AND BAYLEY-III DOMAINS FOR EACH COUNTRY AND OVERALL – CORRELATION COEFFICIENT (95% CIs)

		GSED D-score		
		D-score scale		
Bayley-III domain	**Bangladesh** **(N=159)**	**Pakistan** **(N=158)**	**United Republic of Tanzania** **(N=161)**	**Total** **(N=478)**
Cognitive	0.98 (0.97 - 0.98)	0.95 (0.93 - 0.96)	0.97 (0.96 - 0.98)	**0.97 (0.96 - 0.97)**
Receptive communication	0.91 (0.88 - 0.93)	0.88 (0.84 - 0.91)	0.91 (0.88 - 0.94)	**0.90 (0.88 - 0.92)**
Expressive communication	0.94 (0.92 - 0.96)	0.87 (0.83 - 0.90)	0.89 (0.85 - 0.92)	**0.90 (0.88 - 0.91)**
Fine motor	0.97 (0.96 - 0.98)	0.97 (0.95 - 0.97)	0.97 (0.95 - 0.97)	**0.97 (0.96 - 0.97)**
Gross motor	0.98 (0.97 - 0.98)	0.97 (0.97 - 0.98)	0.98 (0.97 - 0.98)	**0.98 (0.97 - 0.98)**
Overall Bayley-III score	0.99 (0.98 - 0.99)	0.97 (0.96 - 0.98)	0.98 (0.98 - 0.99)	**0.98 (0.98 - 0.98)**
		DAZ scale		
Overall Bayley-III age-adjusted z-score*	0.55 (0.44-0.65)	0.26 (0.11-0.41)	0.56 (0.44-0.66)	**0.53 (0.47-0.6)**

* DAZ domain scores were not produced at a domain level as insufficient data existed to do this robustly.

Convergent validity

Convergent validity is the assessment of how closely a measure is correlated with other variables where correlation is expected. Table 6 gives a selection of variables that were, a priori, expected to correlate with the GSED DAZ score based on evidence from the literature. The table contains Pearson's correlation coefficients with 95% CIs, unless otherwise specified. Some variables were ordinal in nature and required the use of Spearman's correlation coefficient.

For several of the multi-item variables which contained large amounts of missing data, a unidimensional 2PL IRT model was fitted to extract summary scores, under the Missing at Random missingness assumption. The 2PL model may also be more appropriate and accurate for the DHS Wealth Index based on generated internal scores rather than relying on the published quintiles, according to weights established some years ago. A total score (i.e. all countries combined) was not generated for the DHS Wealth Index and maternal education because the items and categories were not comparable across countries. For the total scores, all measures are statistically significant from zero in the hypothesized directions at the 5% level of significance. However, some measures do not differ significantly from zero in the expected directions in the country level analyses. Table 6 compares the results of the GSED SF and GSED LF when administered in a CB against the convergent contextual variables.

TABLE 6. CONVERGENT VALIDITY WITH GSED DAZ – PEARSON'S CORRELATION COEFFICIENT (95% CIs)

	Bangladesh	Pakistan	United Republic of Tanzania	Total
SES - DHS Wealth Index***	0.10 (0.05-0.16)	0.14 (0.09-0.19)	0.15 (0.10-0.20)	NA
Anthropometry - HAZ	0.21 (0.16-0.26)	0.18 (0.13-0.23)	0.21 (0.16-0.26)	**0.19 (0.16-0.22)**
Anthropometry - WAZ	0.21 (0.16-0.26)	0.17 (0.12-0.22)	0.17 (0.12-0.23)	**0.23 (0.20-0.26)**
Birth weight	0.16 (0.10-0.21)	0.03 (-0.02-0.08)	0.20 (0.14-0.25)	**0.04 (0.01-0.07)**
Gestational age	0.11 (0.05-0.16)	0.16 (0.11-0.21)	0.21 (0.15-0.26)	**0.17 (0.14-0.20)**
Maternal education*, ***	0.14 (0.08-0.19)	0.21 (0.16-0.26)	0.06 (0.01-0.11)	NA
PHQ9 category*	-0.05 (-0.10-0.01)	-0.04 (-0.08-0.02)	0.02 (-0.03-0.07)	**0.05 (0.02-0.08)**
HOME	0.21 (0.16-0.26)	0.17 (0.12-0.21)	0.21 (0.16-0.26)	**0.23 (0.20-0.26)**
CPAS **	-0.05 (-0.1-0.01)	-0.05 (-0.10-0.01)	-0.01 (-0.06-0.04)	**-0.07 (-0.1--0.04)**
FSS**	0.11 (0.06-0.16)	0.07 (0.02-0.12)	0.03 (-0.02-0.09)	**0.22 (0.19-0.25)**
BRS*	-0.1 (-0.15--0.05)	-0.09 (-0.14--0.04)	-0.01 (-0.06-0.04)	**-0.09 (-0.12--0.06)**

*Spearman's correlation: maternal education [no schooling, primary, secondary, higher], PHQ9 [minimal, mild, moderate, moderate-severe, severe depression].

** Scale created via a unidimensional 2-parameter IRT model.

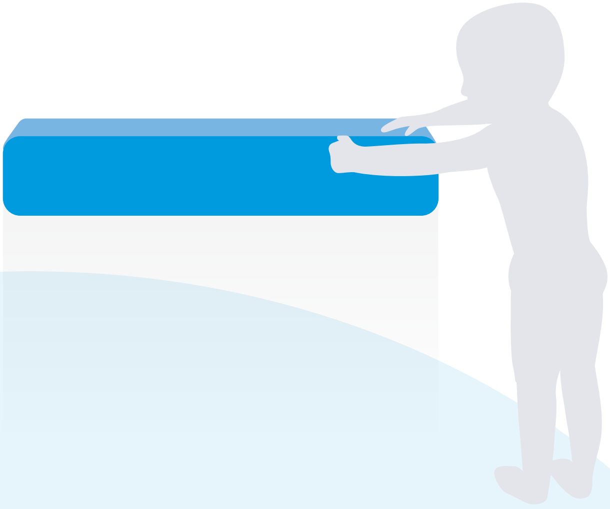
*** For these variables a cross-national scale was not considered appropriate.

Known groups validity

To assess whether the GSED scores were associated with known factors which influence development, a series of logistic regression models, corrected for differential prevalence across country, were fitted to predict the probability of a child having lower than average DAZ given their membership of one of the groups in Table 7. None of the CIs for the odds ratios contain zero, indicating statistical significance at the 5% level. Children who are stunted or underweight (HAZ or WAZ < -2 SD below the mean), children with low birth weight, and children who were premature are all about twice as likely to have lower-than-average DAZ scores. Premature children and those whose mother's used tobacco during pregnancy are 1.77 and 1.34 times more likely to have a lower-than-average DAZ score. Children whose mothers took supplements during pregnancy are 25% less likely to have lower-than-average DAZ scores.

TABLE 7. KNOWN GROUP VALIDITY ODDS RATIOS (95% CIs)

Known group	Odds ratio
Stunted HAZ	2.12 (1.83-2.46)
Underweight WAZ	2.15 (1.81-2.54)
Low birth weight (< 2500 g)	2.24 (1.73-2.90)
Premature (< 38 weeks gestation)	1.77 (1.51-2.08)
Maternal supplement use during pregnancy	0.75 (0.66-0.86)
Maternal tobacco use during pregnancy	1.34(1.05-1.70)



Short-term predictive validity

Short-term predictive validity was assessed via the correlation of the GSED measure at baseline with a GSED measure taken six months (± two weeks) later.

Overall, the correlation between GSED DAZ scores in a six-month interval was 0.59 with similar values across countries (see Table 8).

TABLE 8. PREDICTIVE VALIDITY – CORRELATION COEFFICIENT (95% CIs)

	Bangladesh	Pakistan	United Republic of Tanzania	Total
GSED DAZ at baseline vs GSED DAZ at 6 months	0.55 (0.48 - 0.61)	0.57 (0.51 - 0.63)	0.57 (0.50 - 0.63)	**0.59 (0.56 - 0.63)**

© WHO / Christopher Black

6. GSED package v1.0

This section describes all components for the GSED package v1.0. Package has been tested in three countries and met the target criteria of concurrent, convergent and short-term predictive validity and reliability. Further testing is underway in four additional countries to extend the validity checks (see Section 7).

The package has been created to serve as an open-access population-level measure of ECD for the global community that is comparable across countries. There are no fees nor royalties involved when using it, and it was designed and tested to be linguistically and culturally neutral. It includes: i) GSED SF and GSED LF measures as both a paper version and app; ii) GSED measures *Item Guides*; iii) GSED measures *Administration Manuals*; iv) *Adaptation and Translation Guide*; and v) *Scoring Guide*.

GSED SF and LF measures v1.0

The GSED SF is a caregiver-reported interview, while the GSED LF is comprised of items that are directly administered (e.g. making sounds, fixing gaze and following an object, sitting and standing or identifying objects provided in a picture on a tablet). They both have administration rules ("start" and "stop" rules) based on the age of the child and responses to the items. Table 9 summarizes the key aspects of the GSED SF and GSED LF. Both forms are available in multiple languages, and more translations will become available.

The GSED SF includes media files to be used as prompts to facilitate understanding of the item by the caregiver, such as images and short animations showing actions and skills being asked about (e.g. walking, kicking a ball) or audio files for caregivers to hear the sounds related to the questions (e.g. giggling or cooing).

The GSED LF is organized into three streams which group tasks that are likely observed together in order to streamline and facilitate administration. The GSED LF administration relies partially on low-cost materials organized into a kit that is assembled locally by users following detailed guidance found in the GSED LF User Manual.

The GSED measures should be administered using the **GSED App** on a smart device that includes the in-built administration rules and a user-friendly interface for both caregiver report and direct administration (through a grid-based interface which includes instructions for the assessor) data collection. The GSED App is supported by any mobile device running on the Android operating system, and can be downloaded from Google Play Store. The GSED XLSForm and Xform (xml) version of the GSED SF and GSED LF measures along with the associated media files can be uploaded and configured to any aggregator server (ODK Aggregate/Central, Kobo, etc.) of choice for data collection using the GSED App.

If necessary, a paper version of GSED measures may be used as long as administration rules are followed and assessors have access to the accompanying materials specific to each measure. Alternatively, these forms can be printed. For the GSED LF, the **kit** must be complemented by printing the components that are in-built in the GSED App (i.e. images and booklets). Self-administration and remote administration (e.g. via phone) of the GSED SF are being tested. Current recommendations to maximize the quality of data generated are to use face-to-face administration of the GSED SF and GSED LF with the GSED App.

The GSED HF (Box 4) and GSED PF (Box 5) are not yet fully tested but can be made available on request.

TABLE 9. GSED MEASURES DESCRIPTION

	GSED SF	GSED LF
Primary purpose	Large-scale data collection and monitoring efforts Research and programme evaluation	Research and programme evaluation
Score interpretation	Population-level *NOT for individual-level interpretation*	
Target age	< 36 months	
Format	Caregiver report	Direct administration
Form structure	Ordered questions to caregiver	Items grouped in three streams to facilitate form administration Each stream organized into grid (items do not need to be scored in sequential order)
Total number of Items	139	155
Average number of items administered per child/ caregiver*	30 - 50	45 - 60 (15 - 20 per stream)
Response options	Binary (Yes/No) + "Don't Know" response option (to be used only when absolutely necessary)	Binary (Yes/No) Only observed items qualify to be scored as Yes

	GSED SF	GSED LF
Administration rules	"Start" rule based on age of child and "stop" rule based on reported abilities "Go back" rule to allow measure to be administered to children of all abilities and developmental levels Items should not be skipped to complete the form	"Start" and "stop" rules based on the age of child and performance "Bookends" to allow measure to be administered to children of all abilities and developmental levels All streams must be administered to complete the form
Time range to administer measure*	15 - 25 minutes	30 - 75 minutes
Administration modality	GSED App (recommended) Paper based (also available)	
Materials needed to administer instrument	GSED App: tablet or similar device Administration on paper: printed paper form with a tablet or similar device for audio/visual prompts (may also be printed)	GSED LF Kit GSED App: tablet or similar device Administration on paper: printed paper form and instruction manual

*GSED measures length and administration time are intended to be reduced by revisions planned once data from Round 2 countries are available (see Section 7). The same is true when the adaptive version is available.

Dedicated training courses are required to learn to administer the measures and to train others to administer them. These courses are available in English in person or via Zoom upon request. The suggested length of the training is seven to 10 days (for both measures); however, the training sessions may be tailored according to the users' experience with child development-related tools and depending on whether the GSED SF or GSED LF only is to be used.

Additionally, a series of self-paced online courses are in development. Training courses include resources, such as the GSED Training Manual (available for trainees), which provides guidelines for standardized administration to ensure that the same procedures are used consistently by all assessors. Individuals administering the GSED must familiarize themselves thoroughly with the guidelines and follow them carefully.

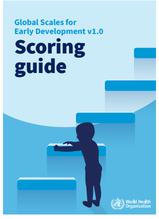
BOX 4. GSED HOUSEHOLD FORM [FOR FURTHER TESTING]

The GSED HF is a population-level, caregiver-reported data collection measure designed to be suitable for integration into multi-topic household surveys to monitor child development globally. It measures the proportion of children up to 24 months of age who are developmentally on track. The GSED HF uses the same metric, the D-score, as the GSED SF and GSED LF. The GSED HF captures the youngest children's achievement of key developmental milestones. It is comprised of 55 items organized by five age bands (0 to < 3 months; 3 to < 6 months; 6 to < 12 months; 12 to < 18 months; 18 to < 24 months). Depending on the age of the child, primary caregivers are asked a set of 20 questions (tailored to the age band, but overlapping across bands) about their children's behaviour, skills and knowledge. The total of 55 questions and their allocation to the specific age bands are the result of a rigorous methodological process to identify the shortest and most informative sets of items to capture child development. The questions were intentionally selected to reflect the increasing complexity of skills children acquire as they become older. Therefore, some questions may seem too easy or too difficult for some children. The GSED HF is accompanied by a package of implementation tools. It is specifically designed to be used in surveys that also collect a wide spectrum of additional data on other family members, and whose focus may not necessarily be ECD. The GSED HF allows inclusion of an ECD component for children < 24 months in multi-topic investigations with minimal additional burden. When used in surveys that are adequately designed and implemented, it allows for the generation of data that are comparable across countries. The GSED HF will be tested in multi-topic household surveys before being released for scale up.

BOX 5. GSED PSYCHOSOCIAL FORM [FOR FURTHER TESTING]

Understanding the emergence of early mental health challenges, including disorders in sleeping, eating and emotion regulation, is an important component of tracking young children's development. Because the GSED was designed to track normative development rather than the emergence of mental health challenges, an additional form intended to be included in the package, called the PF, was created to focus specifically on young children's mental and behavioural well-being. Using items from existing tools for emotion regulation and behaviour problems, an initial set of 49 was selected to index difficulties in eating, sleeping, internalizing and externalizing behaviours, and social competence. In Bangladesh, Pakistan and the United Republic of Tanzania, qualitative data through exit interviews and cognitive testing were collected from a sample of 16 caregiver-child dyads. The qualitative information was used to evaluate the cultural relevance of the GSED measures, caregiver understanding of the items, and feedback on the study implementation processes. In one additional site, the USA, cognitive testing was conducted in both Spanish and English to inform final item selection. In the three Round 1 countries the PF was administered alongside the GSED SF and GSED LF, and in the USA, as part of an ongoing study, it was administered online together with the GSED SF to a sample of approximately 1000 parents *(Marcus Waldman and colleagues, University of Nebraska, unpublished observations, 2023).*

Scoring



Once the GSED SF, GSED LF or both have been administered to one or more children, the next step is to calculate the D-score and the DAZ for each child. This step is known as scoring. The present section provides instructions on how to calculate these scores. Either of two methods may be used:

1. **online calculator.** The online Shiny App (https://tnochildhealthstatistics.shinyapps.io/dcalculator/) is a convenient option for users not familiar with R. The app contains online documentation and instructions;

2. **R package `dscore`** (https://CRAN.R-project.org/package=dscore) is a flexible option with all the tools needed to calculate the D-score. It is an excellent choice for users familiar with R and users who like to incorporate D-score calculations into a workflow.

Revisions to the GSED measures planned for 2024 will not impact the interpretation of the D-scores calculated on previous versions. Scoring for previous versions of the GSED instruments will continue to be available. While procedures for future versions may change, they will continue to produce scores on the same standard D-score scale and thus will remain comparable.

Detailed instructions on how to calculate the D-score and DAZ with the above methods can be found in the GSED Scoring Guide.

Because development naturally occurs with age, it can be difficult to compare D-scores for children of different ages. To help solve this problem, the package also calculates preliminary DAZ scores. These DAZ scores are calculated in reference to same-age children from the Round 1 GSED data from both the GSED SF and GSED LF in Bangladesh, Pakistan and United Republic of Tanzania to estimate the age-conditional distributions of scores. Using this reference group, the D-scores of new data can then be converted into standardized Z scores (with a mean of 0 and SD of 1) at all ages.

While these preliminary norms are useful to adjust scores to remove the age effect, DAZ scores with the current reference population should not be interpreted as representative of any specific population or hold any special normative importance. They are calculated on a non-representative convenience sample. The main utility of these preliminary reference scores is to provide estimates of the stability of GSED and the D-score over time, without artificially inflating correlations due to the strong association between D-scores and age. They can also be used to provide a rough estimate of the association between D-scores and other concurrent and predictive measures.

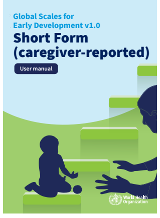
The DAZ is **not** currently an appropriate basis for determining whether children are on or off track developmentally. A Norms and Standards study will be carried out by WHO (see Section 7) which aims to create a better estimate of how D-scores vary by age in a restricted population of children living without major constraints on their development. This updated DAZ will be the focus of ongoing norms and standards work and provide a better justification of cut-off points.

Other components of the GSED package v1.0

Global Scales for Early Development v1.0
Short Form (caregiver-reported)
Item guide

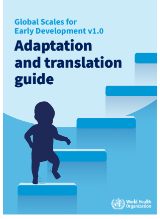
Global Scales for Early Development v1.0
Long Form (directly administered)
Item guide

The GSED package includes the GSED measures as well as accompanying materials to facilitate their implementation and use. A detailed item-by-item description is available through the GSED SF and LF **Item Guides**. They can be used as a resource for both the translation process (to ensure that the translations reflect the original purpose of the questions), adaptation (to ensure instructions are relevant for the context) and training (to ensure that assessors have clear instructions on how to administer and score items). The Item Guides include instructions on how to administer, assess and score each item. In particular, the GSED SF Item Guide further clarifies the purpose of each item, and the GSED LF Item Guide includes indications for methods and props to use, referring to whether the item should be administered by observation, by listening, by demonstrating, etc., and whether any particular GSED LF Kit tool should be used as well.


Global Scales for Early Development v1.0
Short Form (caregiver-reported)
User manual

Global Scales for Early Development v1.0
Long Form (directly administered)
User manual

Each GSED measure is accompanied by a **User Manual** to guide assessors' understanding and use of the measures. Assessors require training and certification to administer the GSED measures, with the manuals and item guides as support. The manuals are organized into four main sections: (a) description of the measure; (b) administration of the measure; (c) what to do and what not to do when administering the measure; and (d) how to address possible challenging situations when administering the measures.


Global Scales for Early Development v1.0
Adaptation and translation guide

To generate high-quality comparable data, the GSED measures should be used in their entirety (no item should be removed or added) without modifications to the item wording and sequence or to the response options. Only the specific adaptation options indicated are acceptable, as well as best practices for translation. If needed, guidance found in the **Adaptation and Translation Guide** should be followed.

© WHO / Kiana Hayeri

7. Next steps

The GSED package aims to provide a feasible and reliable means of collecting population-level data on early development that could be used to monitor progress and policy-level changes, and evaluate programmes and interventions. Data from the GSED will be useful for policy-makers and governments in deciding priorities for funding. Global organizations will be able to use the data for cross-country comparisons and trend analyses. The GSED measures are expected to provide countries with an indication of how the youngest children are developing and become a motivation to invest in and promote healthy development.

This Technical Report provides an overview of the GSED creation and validation methodology with results from three countries. The GSED measures have been shown to be valid and reliable for measurement of child development up to 36 months at the population level. Additional work is ongoing to expand evidence of global validity and reliability (in Round 2 countries and inclusion of GSED in external studies). The analysis related to the work conducted for field-testing of the adaptive testing approach as well as the psychometric property description of the GSED PF across different cultures and contexts, and linkages of GSED SF with ECDI2030 will be finalized and disseminated as soon as available. Similarly, the results of the testing of the GSED measures within the context of programmatic evaluations and use of GSED HF within multi-topic household surveys are expected to be made available soon.

Moreover, the GSED project has been expanded, under GSED 2.0, to answer further research questions. Firstly, additional validation evidence will be generated for: i) predictive validity until 5 years of age by following the cohorts from Round 1 countries; and ii) assessing the association of the GSED measures with biomarkers (including brain imaging). Secondly, a global age-normed distribution of GSED scores through 36 months will be created, based on a rigorous methodology, among children raised with minimal constraints on their development. While the existing D-score package calculates DAZ relative to three possible references (of the Round 1 countries validation data), they are considered an interim reference. Round 2 countries validation data will replace these references using data from all seven countries, but a final revision will be provided when data from the normative sample (under GSED 2.0) are available. These norms will then be used as references to set standards for on- and off-track development, including exploratory adjustment for moderate-to-late preterm babies. Thirdly, both conceptual and field work will address the adaptation of GSED for individual-level identification of children at risk for neurodevelopmental impairment.

Lastly, the D-score approach may be used to harmonize measurements across ages and instruments. Scores from multiple instruments can be translated into D-scores and compared to scores from a different instrument. Future work will evaluate the extension of the D-score to instruments for children beyond 36 months of age. Extending the age limit of the D-score will provide improved guidance to users on tracking children's development over time.

As with other measures of child development, the GSED will continue to evolve as more knowledge is acquired about the capabilities and learning processes that occur in the earliest years of life and about environmental influences on children's early development. It is also expected that technical innovations (e.g. adaptive testing) will facilitate future measurements. Consistent with the SDGs, equity is strived for by developing measurements that provide data to enable government leaders and programme planners to implement strategies that enable all children to reach their developmental potential.

References

1. Clark H, Coll-Seck AM, Banerjee A, Peterson S, Dalglish SL, Ameratunga S et al. A future for the world's children? A WHO–UNICEF–Lancet Commission. Lancet. 2020;395(10224):605-58.
2. Daelmans B, Darmstadt GL, Lombardi J, Black MM, Britto PR, Lye S et al. Early childhood development: the foundation of sustainable development. Lancet. 2017;389(10064):9–11.
3. Lu C, Black MM, Richter LM. Risk of poor development in young children in low-income and middle-income countries: an estimation and analysis at the global, regional, and country level. Lancet Global Health. 2016;4:e916-e22.
4. Black MM, Walker SP, Fernald LCH, Anderson CT, DiGirolamo A, Lu C et al. Early child development coming of age: science through the life-course. Lancet. 2017;389(10064):77-90. doi:10.1016/S0140-6736(16)31389-7.
5. Rubio-Codina M, Grantham-McGregor S. (Predictive validity in middle childhood of short tests of early childhood development used in large scale studies compared to the Bayley-III, the Family Care Indicators, height-for-age, and stunting: a longitudinal study in Bogota, Colombia. PLOS ONE. 2020;15(4): e0231317.
6. Hoddinott J, Behrman JR, Maluccio JA, Melgar P, Quisumbing AR, Ramirez-Zea M et al. Adult consequences of growth failure in early childhood. Am J Clin Nutr. 2013;98(5):1170-8. doi:10.3945/ajcn.113.064584.
7. Fink G, Peet E, Danaei G, Andrews K, McCoy DC, Sudfeld CR et al. Schooling and wage income losses due to early-childhood growth faltering in developing countries: national, regional, and global estimates. Am J Clin Nutr. 2016;104(1):104-12.
8. United Nations. Sustainable Development Goal 4 (Education) [Online] (https://www.sdg4education2030.org/the-goal#:~:text=Sustainable%20Development%20Goal%204%20(SDG%204)%20is%20the%20education%20goal,lifelong%20learning%20opportunities%20for%20all.%E2%80%9D).
9. WHO Multicentre Growth Reference Study Group, de Onis M. WHO Child Growth Standards based on length/height, weight and age. Acta Paediatrica. 2006;95(S450):7685.
10. Ertem IO, Krishnamurthy V, Mulaudzi MC, Sguassero Y, Balta H, Gulumser O 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(3):e279-e91.
11. Fernandes M, Villar J, Stein A, Staines Urias E, Garza C, Victora CG et al. INTERGROWTH-21st Project international INTER-NDA standards for child development at 2 years of age: an international prospective population-based study. BMJ Open. 2020;10(6):e035258. doi:10.1136/bmjopen-2019-035258.
12. Jacobusse G, van Buuren S, Verkerk PH. An interval scale for development of children aged 0–2 years. Stat Med. 2006;25(13):2272–83.
13. Weber AM, Rubio-Codina M, Walker SP, van Buuren S, Eekhout I, Grantham-McGregor SM et al. The D-score: a metric for interpreting the early development of infants and toddlers across global settings. BMJ Glob Health. 2019;4(6):e001724.
14. Ayoub CC, Fischer KW. Developmental pathways and intersections among domains of development. In: McCartney K, Phillips D, editors. Blackwell handbook of early childhood development. Massachusetts: Blackwell Publishing; 2006:62-81.
15. Fischer KW, Bidell TR. Dynamic development of action and thought. In: Lerner RM, Damon W, editors. Handbook of child psychology: theoretical models of human development. New York: John Wiley & Sons Inc.; 2006:313-99.
16. McCray G, McCoy D, Kariger P, Janus M, Black MM, Chang-Lopez S et al. The creation of the Global Scales for Early Development (GSED) for children aged 0-3 years: combining subject matter expert judgements with big data. BMJ Global Health. 2023;8:e009827.
17. McCoy DC, Waldman M, CREDI Field Team, Fink F. Measuring early childhood development at a global scale: evidence from the Caregiver-Reported Early Development Instruments. Early Childhood Research Quarterly. 2018;45:58–68.
18. Gladstone M, Lancaster G, McCray G, Cavallera V, Alves CRL, Maliwichi L et al. Validation of the Infant and Young Child Development (IYCD) indicators in three countries: Brazil, Malawi and Pakistan. Int. J. Environ. Res. Public Health. 2021;18(11):6117.
19. Aylward GP. Brain, environment, and development: a synthesis and a conceptual model. In: Aylward GP, editor. Bayley 4 clinical use and interpretation. London: Academic Press; 2020:1-19.
20. Gesell AL, Halverson HM, Amatruda C. The first five years of life; a guide to the study of the pre-school child. New York: Harper & Brothers; 1940.
21. Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press; 1980.
22. van Buuren S, Eekhout I. Child development with the D-score: turning milestones into measurement [version 1]. Gates Open Res. 2021;5:81 (https://doi.org/10.12688/gatesopenres.13222.1).
23. McCoy DC, Waldman M, CREDI Field Team, Fink G. Measuring early childhood development at a global scale: evidence from the Caregiver-Reported Early Development Instruments. Early Child. Res. Quart. 2018;45:58–68.
24. Chang W, Cheng J, Allaire JJ, Sievert C, Schloerke B, Xie Y et al. shiny: web application framework for R. R package version 1.7.1. 2021; (https://CRAN.R-project.org/package=shiny).

25. Cavallera V, Lancaster G, Gladstone M, Black MM, McCray G, Nizar A et al. Protocol for validation of the Global Scales for Early Development (GSED) for children under 3 years of age in seven countries. BMJ Open 2023;13:e06256227.
26. Salmona M, Lieber E, Kaczynski D. Qualitative and mixed methods data analysis using Dedoose: A practical approach for research across the social sciences. Thousand Oaks: Sage Publications; 2019.
27. Huang CY, Tung LC, Chou YT, Wu HM, Chen KL, Hsieh CL. Development of a computerized adaptive testing of children's gross motor skills. Archives of Physical Medicine & Rehabilitation. 2018;99:512-20.
28. Jacobusse G, van Buuren S. Computerized adaptive testing for measuring development of young children. Statistics in Medicine. 2007;26(13):2629–38 (https://stefvanbuuren.name/publication/2007-01-01_jacobusse2007/).
29. Walker SP, Wachs TD, Gardner JM, Lozoff B, Wasserman GA, Pollitt E et al. Child development: risk factors for adverse outcomes in developing countries. Lancet. 2007;369(9556):145-57.
30. De Ayala RJ. The theory and practice of item response theory. New York: Guilford Publications; 2013.
31. Thissen D. Reliability and measurement precision. In: Wainer H, editor. Computerized adaptive testing: a primer (2nd ed.). New Jersey: Lawrence Erlbaum Associates Publishers; 2000:159-84.
32. Landis R, Koch G. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.
33. Fernald Lia CH, Prado E, Kariger P, Raikes A. A toolkit for measuring early childhood development in low and middle-income countries. Washington, DC: The World Bank; 2017.
34. Lennon EM, Gardner JM, Karmel BZ, Flory MJ. Bayley Scales of Infant Development. In: Haith MM, Benson JB. Encyclopedia of infant and early childhood development. Massachusetts: Academic Press; 2008:145-56.
35. Armstrong KH, Agazzi HC. The Bayley-III cognitive scale. In: Weiss LG, Oakland T, Aylward GP, editors. Practical resources for the mental health professional, Bayley-III clinical use and interpretation. Massachusetts: Academic Press; 2010:29-45.
36. Stasinopoulos DM, Rigby RA. Generalized additive models for location scale and shape (GAMLSS) in R. Journal of Statistical Software. 2008;23:1-46.
37. Sudfeld CR, McCoy DC, Fink G, Muhihi A, Bellinger DC, Masanji H et al. Malnutrition and its determinants are associated with suboptimal cognitive, communication, and motor development in Tanzanian children. J Nutr. 2015;145(12):2705-14. .

Bibliography

Ages and Stages Questionnaire, third edition (ASQ-3)
Squires J, Bricker D. Ages and Stages Questionnaire (ASQ): a parent-completed child monitoring system, third edition. Baltimore (MD): Brooks Publishing Company; 2009.

Bayley Scales of Infant Development (Bayley)
Bayley N. Bayley Scales of Infant Development. San Antonio (TX): The Psychological Corporation; 1969.

Bayley Scales of Infant Development, second edition (Bayley-II)
Bayley N. Bayley Scales of Infant Development, second edition. San Antonio (TX): The Psychological Corporation; 1993.

Caregiver-Reported Early Development Instruments (CREDI)
McCoy DC, Waldman M, CREDI Field Team, Fink G. Measuring early childhood development at a global scale: evidence from the Caregiver-Reported Early Development Instruments. Early Child Res Q. 2018;45(4):58–68. doi.org/10.1016/j.ecresq.2018.05.002.

Denver Developmental Screening Test (DDST)
Frankenburg WK. The Denver Developmental Screening Test. J Pediatr. 1967;71(2):181–91. doi:10.1016/S0022-3476(67)80070-2.

Denver Developmental Screening Test, second edition (DDST II)
Frankenburg WK, Dodds J, Archer P, Shapiro H, Bresnick B. The Denver II: a major revision and restandardization of the Denver Developmental Screening Test. Pediatrics. 1992;89(1):91–7. PMID: 1370185.

Developmental Milestones Checklist (DMC)
Abubakar A, Holding P, van de Vijver FJ, Bomu G, Van Baar A. Developmental monitoring using caregiver reports in a resource-limited setting: the case of Kilifi, Kenya. Acta Paediatr. 2010;99(2):291–7. doi.org/10.1111/j.1651-2227.2009.01561.x.

Developmental Milestones Checklist II (DMC II)
Prado EL, Abubakar AA, Abbeddou S, Jimenez EY, Somé JW, Ouédraogo JB. Extending the Developmental Milestones Checklist for use in a different context in sub-Saharan Africa. Acta Paediatr. 2014;103(4):447–54. doi: 10.1111/apa.12540.

Dutch Development Instrument (DDI)
Laurent de Angulo MS, Brouwers-de JEA, Bijlsma-Schlösser JFM, Bulk-Bunschoten AMW, Pauwels JH, Steinbuch-Linstra I. Ontwikkelingsonderzoek in de Jeugdgezondheidszorg. Het Van Wiechenonderzoek. De Baecke-FassaertMotoriektest. Assen: Van Gorcum; 2008.

Griffiths Mental Development Scales (GMDS)
Huntley M. Griffiths Mental Development Scales from birth to 2 years – manual. Oxford: Association for Research in Infant & Child Development; 1996. doi.org/10.1037/t03301-000.

Griffiths Mental Development Scales – South African Version (GMDS-SA)
Luiz DM, Kotras N, Barnard A, Knoesen N. Technical manual of the Griffiths Mental Development Scales – Extended Revised (GMDS-ER). Amersham: Association for Research in Infant & Child Development; 2004.

Kilifi Developmental Inventory (KDI)
Abubakar A, Holding P, van Baar A, Newton CR, van de Vijver FJR. Monitoring psychomotor development in a resource-limited setting: an evaluation of the Kilifi Developmental Inventory. Ann Trop Paediatr. 2008;28(3):217–26. doi.org/10.1179%2F146532808X335679.

Malawi Developmental Assessment Tool (MDAT)
Gladstone M, Lancaster GA, Umar E, Nyirenda M, Kayira E, Van Den Broek NR et al. The Malawi Developmental Assessment Tool (MDAT): the creation, validation, and reliability of a tool to assess child development in rural African settings. PLoS Med. 2010;7:e1000273. doi:10.1371/journal.pmed.1000273.

Preschool Pediatric Symptoms Checklist (PPSC)
Sheldrick RC, Henson BS, Merchant S, Neger EN, Murphy JM, Perrin EC. The Preschool Pediatric Symptom Checklist (PPSC): development and initial validation of a new social/emotional screening instrument. Acad Pediatr. 2012;12(5):456–67. doi: 10.1016/j.acap.2012.06.008.

Saving Brains Early Childhood Development Scale (SBECD)
McCoy DC, Sudfeld CR, Bellinger DC, Muhihi A, Ashery G, Weary TE et al. Development and validation of an early childhood development scale for use in low-resourced settings. Popul Health Metr. 2017;15(1):3. doi: 10.1186/s12963-017-0122-8.

Stanford-Binet Intelligence Scales, fifth edition (SBIS-5)
Roid GH. Stanford-Binet Intelligence Scales, fifth edition. Itasca (IL): Riverside Publishing; 2003.

Test de Desarrollo Psicomotor [Psychomotor Development Test] (TEPSI)
Haeussler IM, Marchant T. Elaboración y estandarización del Test de Desarrollo Psicomotor 2-5 años TEPSI [Development and standardization of the Psychomotor Development Test 2-5 years]. Rev Educ. 1989.

Vineland Adaptive Behavior Scales (Vineland)
Sparrow SS, Cicchetti DV. The Vineland Adaptive Behavior Scales. In: Newmark CS, editor. Major psychological assessment instruments, Vol. 2. Boston (MD): Allyn & Bacon; 1989:199–231.

Annex 1. Early childhood development dataset for creation of GSED measures

Table A.1.1 lists the cohorts that contributed information to the dataset for creation of GSED with details on number of visits by country, age group and instruments administered.

TABLE A.1.1. COHORTS CONTRIBUTING TO GSED DATASET

Country	Cohort[1]	0 – < 1 year	1 - < 2 years	2 - < 3 years	3+ years	Individual children (N)	Instruments
Bangladesh	CREDI-BGD	49	202	29	0	280	CREDI
Bangladesh	GCDG-BGD-7MO	0	1807	20	0	1827	Bailey-II
Bangladesh	IYCD-BGD-ASQVAL	127	132	88	101	448	Ages and Stages Questionnaire (ASQ)-3, Bailey-III
Brazil	CREDI-BRA-ONLINE	113	287	224	49	673	CREDI
Brazil	CREDI-BRA-SP	472	426	688	65	1651	CREDI
Brazil	GCDG-BRA-1	1875	899	0	0	2774	Denver-II
Brazil	GCDG-BRA-2	3208	4015	551	0	7774	Battelle Developmental Inventory and Screener-2
Brazil	IYCD-BRA-FPS2017	48	26	11	12	97	WHO Indicators of Infant and Young Child Development (IYCD)
Cambodia	CREDI-KHM	126	123	161	83	493	CREDI
Chile	CREDI-CHL	85	88	71	0	244	CREDI
Chile	GCDG-CHL-1	1483	537	0	0	2020	Bailey-I
Chile	GCDG-CHL-2	312	1185	5166	16675	23338	Test de Desarrollo Psicomotor
China	GCDG-CHN	0	982	0	0	982	Bailey-III
Colombia	CREDI-COL	17	121	143	4	285	CREDI
Colombia	GCDG-COL-LT42M	215	417	450	229	1311	Bailey-III
Colombia	GCDG-COL-LT45M	53	632	257	393	1335	Bailey-III, Denver Developmental Screening Test-II, ASQ-3
Costa Rica	IYCD-CRI-PRIDI	0	0	618	1186	1804	Regional Project on Child Development Indicators (PRIDI)
Ecuador	GCDG-ECU	186	259	222	0	667	Barrera
Ethiopia	GCDG-ETH	115	75	440	456	1086	Bailey-III
Ghana	CREDI-GHA	575	541	426	23	1565	CREDI
Guatemala	CREDI-GTM	67	73	57	8	205	CREDI
India	CREDI-IND-ONLINE	85	41	74	0	200	CREDI
India	IYCD-IND-ASQ	1367	1627	17	0	3011	ASQ-3

Country	Cohort[1]	0 – < 1 year	1 - < 2 years	2 - < 3 years	3+ years	Individual children (N)	Instruments
Indonesia	IYCD-IDN-ASQ	757	1006	0	0	1763	ASQ-3
Jamaica	GCDG-JAM-LBW	0	327	116	0	443	Griffiths Mental Development Scales (GMDS)
Jamaica	GCDG-JAM-STUNTED	5	144	151	177	477	GMDS
Jordan	CREDI-JOR	114	98	66	37	315	CREDI
Kenya	IYCD-KEN-DID	79	148	196	0	423	Kilifi Developmental Inventory
Kenya	IYCD-KEN-DMC	188	96	0	0	284	Developmental Milestone Chart
Lao People's Democratic Republic	CREDI-LAO	16	18	9	3	46	CREDI
Lebanon	CREDI-LBN	181	118	84	41	424	CREDI
Madagascar	GCDG-MDG	0	0	18	187	205	Stanford Binet Test
Malawi	IYCD-MWI-FPS2017	39	20	9	9	77	IYCD
Malawi	IYCD-MWI-MDAT	687	276	130	353	1446	Malawi Development Assessment Tool
Nepal	CREDI-NPL	227	136	0	0	363	CREDI
Netherlands	GCDG-NLD-2	0	262	1253	2130	3645	Dutch Development Instrument (DDI)
Netherlands	GCDG-NLD-SMOCC	10 110	5120	1308	0	16 538	DDI
Nicaragua	IYCD-NIC-PRIDI	0	0	583	1251	1834	PRIDI
Pakistan	CREDI-PAK	85	80	76	9	250	CREDI
Pakistan	IYCD-PAK-FPS2017	48	23	12	12	95	IYCD
Paraguay	IYCD-PRY-PRIDI	0	2	456	1044	1502	PRIDI
Peru	IYCD-PER-ASQ	1261	1654	3	0	2918	ASQ-3
Peru	IYCD-PER-PRIDI	0	0	825	1742	2567	PRIDI
Philippines	CREDI-PHL	198	351	170	1	720	CREDI
South Africa	GCDG-ZAF	490	796	1275	1614	4175	Bailey-I, Vineland Adaptive Behavior Scales, GMDS
United Republic of Tanzania	CREDI-TZA-MALARIA	0	56	132	9	197	CREDI
United Republic of Tanzania	CREDI-TZA-NEOVITA	0	938	1467	76	2481	CREDI
USA	CREDI-USA-BOS	61	56	37	2	156	CREDI
USA	CREDI-USA-ONLINE	336	188	221	0	745	CREDI
Zambia	CREDI-ZMB-CHIPATA	223	591	236	0	1050	CREDI
Zambia	CREDI-ZMB-CHOMA	519	378	47	0	944	CREDI

[1] Cohort name is an internal coding representing original group, country and number.

Annex 2. GSED study validation measures

Table A.2.1 lists and describes the study measures in addition to GSED that were collected for validation processes. They capture children's growth and nutrition, health, environmental and contextual information.

TABLE A.2.1. STUDY MEASURES USED FOR VALIDATION PROCESSES

Construct	What the measure captures	Measure	Administration mode	Time to administer (minutes)
Child health and household SES	• Eligibility (exclusion criteria) • Demographic information • Information about acute child health • Delivery and perinatal conditions • Breastfeeding • Child's health history • Household SES* • Caregiver education • Maternal health/chronic illness • COVID-19 exposure	Eligibility and contextual form (specifically developed for the study)	Caregiver report	35
Anthropometry	• Weight at time of assessment • Infant length/child height at time of assessment • Child's mid-upper arm circumference at time of assessment • Child's head circumference at time of assessment	Anthropometry form	Child assessment	15
Family/home environment	• Home environment (HOME only) • Play/stimulation/interactions between the child and other family members in the home (HOME and FCI)	HOME OR FCI	*HOME:* caregiver report & observation *FCI:* caregiver report	*HOME:* 45 *FCI:* 15
	• Child neglect/abuse • Exposure to violence or conflict	CPAS†	Caregiver report	15
	• Family resilience	BRS†	Caregiver report	1
	• Family social support	FSS†	Caregiver report	5
Caregiver health and well-being	• Caregiver depressive symptoms	PHQ9	Caregiver report	5
Child development	• Global child development (0 - 41 months)	Bayley-III OR GMDS‡	Direct child assessment	45 - 60
	• Global child development (24 - 41 months)	ECDI2030§	Caregiver report	10

* SES information on this form comes from the standard DHS multiple assets index; however, some sites have adapted the items to better fit their contexts.

† These measures have been slightly adapted for the purpose of the study.

‡ In a subsample (N=150).

§ In a subsample (all children of 24 - 41 months within the predictive validity subsamples in three countries).

Annex 3. Validity results by GSED measure

The tables in this Annex present the validity and reliability results for each GSED measure individually as well as for the scale as a whole (i.e. the CB).

TABLE A.3.1. CONCURRENT VALIDITY (WITH BAYLEY-III)

D-score scale					
GSED measure	**Bayley-III domain**	**Bangladesh**	**Pakistan**	**United Republic of Tanzania**	**Combined**
CB	Cognitive	0.98 (0.97-0.98)	0.94 (0.92-0.95)	0.97 (0.96-0.98)	0.96 (0.95-0.97)
SF	Cognitive	0.97 (0.96-0.98)	0.93 (0.91-0.95)	0.96 (0.95-0.97)	0.95 (0.94-0.96)
LF	Cognitive	0.98 (0.97-0.98)	0.94 (0.92-0.96)	0.97 (0.96-0.98)	0.96 (0.96-0.97)
CB	Receptive communication	0.91 (0.88-0.93)	0.87 (0.83-0.91)	0.91 (0.88-0.93)	0.90 (0.88-0.92)
SF	Receptive communication	0.90 (0.87-0.93)	0.87 (0.82-0.90)	0.90 (0.86-0.92)	0.89 (0.87-0.91)
LF	Receptive communication	0.91 (0.88-0.93)	0.88 (0.84-0.91)	0.92 (0.89-0.94)	0.90 (0.88-0.92)
CB	Expressive communication	0.94 (0.92-0.96)	0.87 (0.82-0.90)	0.89 (0.85-0.92)	0.90 (0.88-0.91)
SF	Expressive communication	0.93 (0.91-0.95)	0.86 (0.81-0.89)	0.87 (0.82-0.90)	0.88 (0.86-0.90)
LF	Expressive communication	0.94 (0.92-0.96)	0.87 (0.83-0.90)	0.90 (0.87-0.93)	0.90 (0.88-0.92)
CB	Fine motor	0.9 7(0.96-0.98)	0.96 (0.94-0.97)	0.96 (0.95-0.97)	0.96 (0.96-0.97)
SF	Fine motor	0.96 (0.95-0.97)	0.95 (0.94-0.97)	0.95 (0.94-0.97)	0.96 (0.95-0.96)
LF	Fine motor	0.97 (0.96-0.98)	0.96 (0.94-0.97)	0.96 (0.94-0.97)	0.96 (0.95-0.97)
CB	Gross motor	0.98 (0.97-0.98)	0.97 (0.95-0.98)	0.98 (0.97-0.98)	0.97 (0.97-0.98)
SF	Gross motor	0.97 (0.95-0.97)	0.96 (0.95-0.97)	0.97 (0.96-0.98)	0.97 (0.96-0.97)
LF	Gross motor	0.97 (0.97-0.98)	0.96 (0.95-0.97)	0.97 (0.96-0.98)	0.97 (0.96-0.97)
CB	Overall Bayley-III score	0.99 (0.98-0.99)	0.96 (0.95-0.97)	0.98 (0.97-0.99)	0.98 (0.97-0.98)
SF	Overall Bayley-III score	0.97 (0.95-0.97)	0.96 (0.95-0.97)	0.97 (0.96-0.98)	0.97 (0.96-0.97)
LF	Overall Bayley-III score	0.99 (0.98-0.99)	0.96 (0.95-0.97)	0.98 (0.97-0.98)	0.98 (0.97-0.98)
DAZ scale					
CB	Overall Bayley-III score	0.55 (0.44-0.65)	0.26 (0.11-0.41)	0.56 (0.44-0.66)	0.53 (0.47-0.6)
SF	Overall Bayley-III score	0.37 (0.23-0.50)	0.18 (0.03-0.33)	0.40 (0.26-0.52)	0.35 (0.27-0.43)
LF	Overall Bayley-III score	0.59 (0.48-0.68)	0.31 (0.16-0.44)	0.60 (0.49-0.69)	0.58 (0.52-0.64)

TABLE A.3.2. CONVERGENT VALIDITY

		D-score scale			
GSED measure	Bayley-III domain	Bangladesh	Pakistan	United Republic of Tanzania	Combined
CB	SES-DHS Wealth Index**	0.10 (0.05-0.16)	0.14 (0.09-0.19)	0.15 (0.10-0.20)	NA
SF	SES-DHS Wealth Index**	0.07 (0.02–0.12)	0.11 (0.61-0.16)	0.10 (0.05-0.15)	NA
LF	SES-DHS Wealth Index**	0.07 (0.02-0.13)	0.12 (0.08-0.17)	0.14 (0.09-0.19)	NA
CB	Anthro-HAZ	0.21 (0.16-0.26)	0.18 (0.13-0.23)	0.21 (0.16-0.26)	0.19 (0.16-0.22)
SF	Anthro-HAZ	0.16 (0.10-0.21)	0.13 (0.08-0.18)	0.14 (0.08-0.19)	0.14 (0.11-0.17)
LF	Anthro-HAZ	0.2 (0.15-0.25)	0.18 (0.13-0.23)	0.22 (0.17-0.27)	0.19 (0.16-0.22)
CB	Anthro-WAZ	0.21 (0.16-0.26)	0.17 (0.12-0.22)	0.17 (0.12-0.23)	0.23 (0.20-0.26)
SF	Anthro-WAZ	0.16 (0.11-0.21)	0.11 (0.06-0.16)	0.17 (0.11-0.22)	0.16 (0.13-0.19)
LF	Anthro-WAZ	0.22 (0.16-0.27)	0.19 (0.15-0.24)	0.16 (0.11-0.21)	0.24 (0.21-0.27)
CB	Birthweight	0.16 (0.10-0.21)	0.03 (-0.02-0.08)	0.20 (0.14-0.25)	0.04 (0.01-0.07)
SF	Birthweight	0.09 (0.04-0.14)	0.00 (-0.05-0.05)	0.13 (0.07-0.18)	0.03 (0.00-0.06)
LF	Birthweight	0.16 (0.1-0.21)	0.06 (0.01-0.11)	0.19 (0.14-0.24)	0.03 (0.00-0.06)
CB	Gestational age	0.11 (0.05-0.16)	0.16 (0.11-0.21)	0.21 (0.15-0.26)	0.17 (0.14-0.20)
SF	Gestational age	0.06 (0.00-0.11)	0.13 (0.08-0.17)	0.14 (0.09-0.19)	0.12 (0.09-0.15)
LF	Gestational age	0.13 (0.07-0.18)	0.12 (0.07-0.17)	0.18 (0.13-0.23)	0.16 (0.13-0.19)
CB	Maternal education*	0.14 (0.08-0.19)	0.21 (0.16-0.26)	0.06 (0.01-0.11)	NA
SF	Maternal education*	0.12 (0.07-0.18)	0.18 (0.14-0.23)	0.05 (0.00-0.10)	NA
LF	Maternal education*	0.09 (0.41-0.15)	0.15 (0.10-0.20)	0.04 (-0.01-0.10)	NA
CB	PHQ9 category	-0.05 (-0.10-0.01)	-0.04 (-0.08-0.02)	0.02 (-0.03-0.07)	0.05 (0.02-0.08)
SF	PHQ9 category	-0.05 (-0.10-0.01)	-0.01 (-0.06-0.04)	0.01 (-0.05-0.06)	0.01 (-0.02-0.04)
LF	PHQ9 category	-0.02 (-0.08-0.03)	-0.08 (-0.13—0.03)	0.02 (-0.03-0.08)	0.07 (0.04-010)
CB	Home	0.21 (0.16-0.26)	0.17 (0.12-0.21)	0.21 (0.16-0.26)	0.23 (0.20-0.26)
SF	Home	0.18 (0.13-0.23)	0.15 (0.10-0.20)	0.22 (0.17-0.27)	0.19 (0.16-0.22)
LF	Home	0.14 (0.08-0.19)	0.12 (0.07-0.17)	0.09 (0.04-0.15)	0.18 (0.15-0.21)
CB	FSS**	0.11 (0.06-0.16)	0.07 (0.02-0.12)	0.03 (-0.02-0.09)	0.22 (0.19-0.25)
SF	FSS**	0.06 (0-0.11)	0.04 (-0.01-0.09)	0.07 (0.01-0.12)	0.11 (0.08-0.14)
LF	FSS**	0.14 (0.08-0.19)	0.09 (0.05-0.14)	0.01 (-0.05-0.06)	0.28 (0.25-0.30)
CB	BRS**	-0.1 (-0.15--0.05)	-0.09 (-0.14--0.04)	-0.01 (-0.06-0.04)	-0.09 (-0.12--0.06)
SF	BRS**	-0.03 (-0.08-0.02)	-0.05 (-0.1--0.01)	0.00 (-0.06-0.05)	-0.04 (-0.07--0.01)
LF	BRS**	-0.11 (-0.16--0.06)	-0.1 (-0.15--0.06)	0.01 (-0.04-0.06)	-0.10 (-0.13--0.07)
CB	CPAS**	-0.05 (-0.1 - 0.01)	-0.05 (-0.10 - 0.01)	-0.01 (-0.06 - 0.04)	-0.07 (-0.1 - -0.04)
SF	CPAS**	-0.03 (-0.08 - 0.02)	-0.05 (-0.10 - 0.00)	-0.01 (-0.06 - 0.04)	-0.05 (-0.08- -0.02)
LF	CPAS**	-0.05 (-0.11 - 0.00)	-0.03 (-0.07 - 0.02)	0.00 (-0.05 - 0.06)	-0.06 (-0.09- -0.03)

* Spearman's correlation: maternal education (no schooling, primary, secondary, higher), PHQ9 (none, mild, moderate, moderate-severe, severe depression).

** Scale created via a unidimensional 2-parameter IRT model.

*** For these variables a cross-national scale was not considered appropriate.

TABLE A.V.3. SHORT-TERM PREDICTIVE VALIDITY (AT 6 MONTHS)

D-score scale				
	Bangladesh	Pakistan	United Republic of Tanzania	Combined
GSED CB DAZ at baseline vs GSED DAZ at 6 months	0.55 (0.48- 0.61)	0.57 (0.51- 0.63)	0.57 (0.50- 0.63)	0.59 (0.56 - 0.63)
GSED SF DAZ at baseline vs GSED DAZ at 6 months	0.53 (0.46 - 0.59)	0.56 (0.50 - 0.62)	0.58 (0.52 - 0.64)	0.57 (0.53 - 0.6)
GSED LF DAZ at baseline vs GSED DAZ at 6 months	0.38 (0.3 - 0.46)	0.43 (0.35 - 0.50)	0.38 (0.3 - 0.46)	0.48 (0.43 - 0.52)

FOR MORE INFORMATION PLEASE CONTACT:

Brain Health Unit
Department of Mental Health and Substance Use
World Health Organization
Avenue Appia 20
CH-1211 Geneva 27
Switzerland

Email: GSED@who.int

Website: https://www.who.int/teams/mental-health-and-substance-use/data-research/global-scale-for-early-development

World Health Organization