A Toolkit for Measuring Early Childhood Development in Low- and Middle-Income Countries

Lia C. H. Fernald, Elizabeth Prado, Patricia Kariger, Abbie Raikes

Prepared for the Strategic Impact Evaluation Fund, the World Bank

WORLD BANK GROUP

SIEF Strategic Impact Evaluation Fund

1818 H Street NW
Washington DC 20433
Telephone: 202-473-1000
Internet: www.worldbank.org

This work is a product of the staff of The World Bank with external contributions. The findings, interpretations, and conclusions expressed in this work do not necessarily reflect the views of The World Bank, its Board of Executive Directors, or the governments they represent.

The World Bank does not guarantee the accuracy of the data included in this work. The boundaries, colors, denominations, and other information shown on any map in this work do not imply any judgment on the part of The World Bank concerning the legal status of any territory or the endorsement or acceptance of such boundaries.

A Toolkit for Measuring Early Childhood Development in Low- and Middle-Income Countries

Lia C. H. Fernald, Elizabeth Prado, Patricia Kariger and Abbie Raikes

Prepared for the Strategic Impact Evaluation Fund, the World Bank

Table of Contents

List of Figures

List of Tables

List of Sidebars

List of Formulas

Introduction

The primary purpose of this Toolkit is to provide a resource for researchers, evaluators, and program personnel from various disciplines interested in assessing early childhood development (ECD) in low- and middle-income countries—either for planning and evaluating interventions, monitoring development over time, or conducting a situation analysis. The Toolkit is intended to help produce reliable, actionable data on child development.

Such data are essential at this time. Children in low- and middle-income countries are growing up at a disadvantage. According to estimates in the 2017 Early Childhood Development Series series in *The Lancet*, more than 250 million children aged under five years worldwide are living in poverty or are stunted and thus are at risk for not fulfilling their potential for physical growth, cognition, or social-emotional development (Black et al. 2017). The first five years of life lay the groundwork for lifelong development (Shonkoff and Phillips 2000), and skills developed prior to school entry help determine children's academic success. It is important to assess children during this vulnerable period to determine if they are developing appropriately and to design interventions if they are not. The accurate measurement of a young child's abilities—which may reflect future potential and productivity—is essential for understanding the immediate and long-term impacts of such interventions and to inform policy and practice.

The demand for child development measurements is increasing in low- and middle-income countries. The United Nations Sustainable Development Goals (SDGs) have placed early child development on the global policy agenda for the first time. Goal 4.2 aims for access to quality early child development for all, highlighting the importance of early childhood development and the demand for effective interventions in low- and middle-income countries. Measurement can help generate information on progress and challenges in reaching this target and can help evaluate programs and interventions to inform evidence-based policy.

This Toolkit provides a practical, "how-to" guide for selection and adaptation of child development measurements for use in low- and middle-income countries. Users can follow our proposed step-by-step process to select, adapt, implement, and analyze early childhood development data for diverse purposes and projects. We have built on a previous edition of the Toolkit and the work of several excellent academic papers that have recently reviewed the use of child development measurement tools in such countries (Sabanathan, Wills, and Gladstone 2015; Semrud-Clikeman et al. 2016; Fischer, Morris, and Martines 2014). We also use recent collections of common tools put together by organizations such as Saving Brains and the National Institutes of Health (NIH).

The first version of this Toolkit, published in 2009, reviewed 41 assessment tools that had been developed or used for children ages 0–5 years in low- and middle-income countries. Since 2009, many more developmental assessment tools have been created, and the Toolkit now includes 106 new tools for children ages 0–8 years. Fourteen of the 41 tools in the previous version of the Toolkit originated from a low- or middle-income country (e.g., the Malawi Developmental Assessment Tool), or were developed for multiple countries simultaneously, including at least one low- or middle-income country (e.g., the WHO Motor Milestones). A total of 47 (44 percent) of the 106 newly added tools met these criteria. India and Kenya have produced the greatest number of tools (nine and five, respectively), while 17 tools were developed in multiple countries simultaneously. For the current update, we searched for these additional tools by examining the references cited in several reviews (Sabanathan, Wills, and Gladstone 2015; Semrud-Clikeman et al. 2016; Fischer, Morris, and Martines 2014). We also searched in a landscaping analysis[1] commissioned by the Bill & Melinda Gates Foundation, looked through a test inventory for school-age children developed by researchers at New York University (Wuermli et al. 2016), and carried out keyword searches of PubMed, Google Scholar, PsycINFO, and others. This yielded an additional 71 tools that have been developed or used in low- and middle-income countries, or have been suggested by experts as promising tools to assess neurobehavioral development in children ages 0–5 years in low- and middle-income country contexts, reflecting a growing interest in this area in the global community. Our search identified an additional 35 tools for children ages 5–8 years.

The new tests covered nine domains of child development, with cognitive, language, and motor development most commonly assessed. The majority of tests (88 tests, or 60 percent) covered multiple domains, while those that covered a single domain most commonly measured cognitive (12 tests) or social-emotional development (17 tests), academic skills (10 tests), or executive function (10 tests). The majority of tests were individual-level screening or ability tests, with only 11 population-level assessments found.

The most widely used tools, which we found to have been applied in at least 20 different countries, originated from the United States. These include the Achenbach Child Behavior Checklist (CBCL), Bayley Scales of Infant Development (BSID), Wechsler Intelligence Scale for Children (WISC), Wechsler Preschool and Primary Scale of Intelligence (WPPSI), Ages & Stages Questionnaires (ASQ), Strengths and Difficulties Questionnaire (SDQ), and Denver Developmental Screening Test. Additionally, several new tests have been developed in multiple countries and have already gained widespread use, including the Early Grade Reading Assessment (EGRA) (Dubeck and Gove 2015), used in 65 countries; Early Grade Mathematics Assessment (EGMA) (Reubens 2009), used in 22 countries; Save the Children's Literary Boost assessment toolkit, used in 24 countries; and the Multiple Indicator Cluster Surveys (MICS) Early Child Development Index (ECDI), used in 36 countries. Of the tools developed in a low- or middle-income country before 2009, two have been used in a growing number of countries. The Kilifi Developmental Inventory, originally developed in Kenya, has been used in studies in Uganda, Malawi, Ghana, and South Africa. The Guide for Monitoring Child Development, originally developed in Turkey, has been used in Argentina, India, and South Africa.

All 147 measurement tools are listed in the *ECD Measurement Inventory* that accompanies this Toolkit. In the *ECD Measurement Inventory,* we encoded information regarding the domains assessed, age range for which the tool is appropriate, method of administration, purpose of the assessment, origin and locations of use, logistics, and cost.[2]

This newest edition of the Toolkit also includes discussion of evidence for predictive validity of early childhood development measurements to forecast later intelligence quotient (IQ), school achievement, and other adult outcomes; tools appropriate for children up to age eight years, expanded from the previous version for the age range 0–5 years; tools that use rapidly developing technologies, including neuroimaging measures (event-related potential, or ERP; functional near-infrared spectroscopy, or fNIRS), computer-administered cognitive tests, and data collection devices such as accelerometers and eye-trackers; tools to measure the home environment and preschool quality; information related to new adaptations and psychometric evaluations of previously published tests in low-income countries; and a step-by-step guide to adapting and evaluating a test in a new context.

The nine chapters of this Toolkit, along with the accompanying *ECD Measurement Inventory*, provide researchers, evaluators, and program personnel from various disciplines with rationale for measuring particular domains of child development. They also offer guidance on current best practices for using this kind of measurement for population monitoring, program evaluation, and exploratory research.

[1] This is a tool that catalogues and analyzes instruments used in developing countries to assess four domains of child development (cognitive abilities, social and behavioral development, motor skills, and home environment).

[2] Two coders independently looked up the information for the first 18 columns of the "Test" sheet, plus the "Administration time" column to check for accuracy. Sixty-four percent of these fields were double-coded and agreement was 84 percent. We double-checked and corrected any information that did not agree.

Chapter 1 *(Why Measure Early Childhood Development)* explains how assessments of early child development can be used for population monitoring, program evaluation, or exploratory research. The ideal child development assessment has high reliability and validity and cross-cultural appropriateness. It should be easy to administer and should show variance in scores at all ages and ability levels. The selection of any given assessment, however, will likely require trade-offs among these different aspects of ideal tests. The goals of the adaptation process for an assessment are often to develop a locally appropriate tool that is equivalent to the original tool or to assess the same underlying ability or construct in a way that is appropriate in the local setting. The chapter concludes with a proposed framework for selecting tests based on project-specific purposes and priorities.

Chapter 2 *(Early Childhood Development and Its Determinants)* defines exactly what we mean by child development and what we know about the dynamic interplay between biological and environmental factors that determine it. Assessing child development can provide insights into the biological, physiological, behavioral, and psychological growth occurring in the early years of life. Evidence from impact evaluations in both high- and low-income settings suggests that development is malleable and can be improved by early intervention. The breadth and depth of behaviors that can be assessed increase with age, and the advancement in communication and other skills during the preschool and early primary years provides additional modes for testing.

Chapter 3 *(Measurable Skills and Longer-Term Impact)* describes the main domains of child development and school readiness that can be assessed, including cognitive skills, language, executive function and self-regulation, and motor, social and emotional, and pre-academic skills, among others. Across domains, existing assessments in the age range 0–2 years are generally poor predictors of later performance (e.g., during primary school age), but become stronger by ages 3–5 years. Starting at age three years, the predictive validity of assessments within domains is stronger than across domains, and no single domain is the strongest predictor of both academic and behavioral performance at school age.

Chapter 4 *(Assessment Tools and Their Uses)* describes how assessments can be physiological or behavioral and can be conducted via direct assessment, parent or teacher report, naturalistic or structured observation, or direct measures of brain structure and function. The accuracy of conclusions about child development depends on how well the data reflect the underlying construct or ability we intended to measure, so it is crucial that the method accurately reflects the underlying construct and the method is implemented with high quality and fidelity. For program evaluations, brief assessment tools with just five or six items per age category may not show sufficient variance in scores in typically developing children for capturing treatment-related effects.

Chapter 5 *(Measure Selection)* offers guidance for selecting a measurement tool and outlines features that an ideal early childhood assessment would have. The selection of an assessment tool involves a trade-off between the various advantages and disadvantages of any method. There are important ethical considerations for whatever assessment is used. To assist with choice of assessment, the *ECD Measurement Inventory* accompanying this Toolkit is a database of 147 developmental assessment tools. For each tool, the database contains information on the domains assessed, the age range for which the tool is appropriate, the method of administration, the purpose of the assessment, its origin and locations of use, logistics, and costs.

Chapter 6 *(Adaptation and Standardization of Existing Tools)* presents issues that need to be addressed when modifying existing assessments, adapting them to new contexts, and standardizing them. The development of culture-free cognitive tests is currently impossible because no test (even non-verbal) can avoid some bias due to cultural and linguistic differences. Successful adaptation requires several steps, including accurate translation, cultural adaptation, pre-testing, pilot testing, and test modification. Maintaining the reliability and validity of assessments is important, and these properties need to be determined within a given cultural context.

Chapter 7 *(Creating New Assessments)* outlines required steps when developing a new assessment and cautions against embarking on such an exercise given the work and expertise entailed. Creating a new assessment is not easy, requiring a great deal of time, energy, and resources (financial and human), and thus is not generally recommended. A developmental psychologist or someone with equivalent training should lead or be part of the team developing any new tests. Many new assessments have been created, and they have great potential for use in low- and middle-income country contexts.

Chapter 8 *(Children's Home and Early Learning Environments)* reviews methods and tools for assessing children's home and learning environments. The quality of children's early environments has a large influence on development and performance on developmental assessments. When using these measures, it is important to adapt the test to the particular culture and context where the measure is being used.

Chapter 9 *(Summary and Recommendations)* concludes with recommendations on how to carry out effective and accurate measurement of early childhood development.

Acknowledgements

WE ARE THANKFUL FOR THE INSIGHTS AND INPUT of Julieta Trias, Owen Ozier, Emanuela Galasso, Tamara Arnold, Alaka Holla, and Amer Hasan from the World Bank. The book's development was coordinated by Julieta Trias. We also received helpful feedback from Maureen Black, Dana Charles McCoy, Gunther Fink, Charles Nelson, Ann Weber, Anna Pulakka, Virginia Marchman, Marta Rubio-Codina, Aisha Yousafzai, Sally Brinkman, and Fahmida Tofail. Many thanks go to Helen Pitchik for research assistance and literature searches and to Kallista Bley and Heather Knauer, who also contributed. We would also like to acknowledge the late Patrice Engle, who contributed significantly to the preparation of the previous edition of the Toolkit. Kathy Chen and Aliza Marcus copyedited and proofread the Toolkit, and Nicole Hamam was the graphic designer. Any errors or omissions are those of the authors only.

Acronyms and Abbreviations

ASQ	Ages & Stages Questionnaires (and ASQ-3)
ASQ INVENTORY	Ages & Stages Questionnaires Inventory
BDI-2	Battelle Developmental Inventory - Second Edition
BSID	Baley Scales of Infant Development (and BSID-III)
CANTAB	Cambridge Neuropsychological Test Automated Battery
CAPI	Computer-assisted personal interview
CBCL	Child Behavior Checklist
CDI	MacArthur-Bates Communicative Development Inventories
CLASS	Classroom Assessment Scoring System
CREDI	Caregiver-Reported Early Development Index
DDST	Denver Developmental Screening Test
DMC	Developmental Milestones Checklist
DQ	Developmental quotient
EAIS	Escala Argentina de Inteligencia Sensorio-motriz
EAP-ECDS	East Asia-Pacific Early Child Development Scales
ECD	Early childhood development
ECDI	Early childhood development index (from MICS survey)
ECERS-R	Early Childhood Environment Rating Scale-Revised
EDI	Early Development Instrument
EEDP	Escala de Evaluación del Desarrollo Psicomotor
EEG	Electroencephalography
EGMA	Early Grade Mathematics Assessment
EGRA	Early Grade Reading Assessment
EHCI	Early Human Capability Index
ELDS	Early Learning and Development Standards
ERP	Event-related potential
FCI	Family care indicators
fNIRS	Functional near-infrared spectroscopy
GDP	Gross domestic product
GMCD	Guide for Monitoring Child Development
GMDS	Griffiths Mental Development Scale
HSQ	Home Screening Questionnaire
HOME	Home Observation for Measurement of the Environment
HOME-SF	HOME-Short Form
ICMR	Indian Council of Medical Research
IDELA	International Development and Early Learning Assessment
IDS	Infant Development Scale
IEA	International Association for the Evaluation of Educational Achievement
ITC	International Test Commission
ITERS-R	Infant/Toddler Environment Rating Scale-Revised
IQ	Intelligence quotient

KABC	Kaufman Assessment Battery for Children
KDI	Kilifi Developmental Inventory
LENA™	Language Environment Analysis
LMIC	Low- and middle-income countries
MDAT	Malawi Developmental Assessment Tool
MDI	Mental Development Index, component of the Bayley Scales of Infant Development
MELE	Measure of Early Learning Environments
MELQO	Measuring Early Learning Quality and Outcomes
MICS	Multiple Indicator Cluster Surveys (MICS4, MICS5 are MICS surveys conducted at different times)
MRI	Magnetic resonance imaging
NEPSY	Developmental Neuropsychological Assessment
NEGP	(U.S.) National Education Goals Panel
NIH	National Institutes of Health
NLSY	National Longitudinal Survey of Youth
OMCI	Observation of Mother-Child Interactions
PEBL	Psychology Experiment Building Language
PEDS	Parents' Evaluation of Developmental Status
PICCOLO	Parenting Interactions with Children: Checklist of Observations Linked to Outcomes
PRIDI	Regional Project on Child Development Indicators
PROCESS	Pediatric Review and Observation of Children's Environmental Support and Stimulation
RACER	Rapid Assessment of Cognitive and Emotional Regulation
REEL	Receptive-Expressive Emergent Language Test
RS	Reference interviewer
SABER-ECD	Systems Approach for Better Education Results-Early Childhood Development
SDGs	Sustainable Development Goals
SDQ	Strengths and Difficulties Questionnaire
SD	Standard deviation
SES	Socio-economic status
SUMMIT	Supplementation with Multiple Micronutrients Intervention Trial
TEEP	Turkish Early Enrichment Project
TEPSI	Test de Desarrollo Psicomotor
TIPPS	Teacher Instructional Practices and Processes System
TQQ	Ten Questions Questionnaire
UNICEF	United Nations Children's Fund
UNESCO	United Nations Educational, Scientific and Cultural Organization
WASH	Water, sanitation, and hygiene
WISC	Wechsler Intelligence Scale for Children
WHO	World Health Organization

All dollar amounts are in U.S. dollars unless otherwise indicated.

1

Why Measure Early Childhood Development?

KEY MESSAGES:

- Assessments of early child development can be used for population monitoring, program evaluation, or exploratory research. Screening tools can also be used to identify children who may need further testing, diagnosis, and treatment.
- The ideal child development assessment is easy to administer and has high reliability, validity, and cross-cultural appropriateness. It should also show variance in scores at all ages and ability levels.
- The selection of any given assessment will require trade-offs among different aspects of ideal tests.
- When an assessment is appropriately adapted, it becomes a locally appropriate tool that is equivalent to the original tool or that assesses the same underlying ability or construct in a way that is appropriate in the local setting.
- Creating new tests rather than adapting existing tests is a time-consuming, laborious, and expensive process, and it is very difficult to do well.

MEASURING EARLY CHILDHOOD DEVELOPMENT IS CRITICAL if policymakers, development organizations, and others are to have the information needed to develop and implement effective programs and policies. Children in low- and middle-income countries are growing up at a disadvantage. According to estimates in the 2017 Early Childhood Development Series in *The Lancet*, more than 250 million children aged under five years worldwide are living in poverty or are stunted and thus are at risk for not fulfilling their potential for physical growth, cognition, or social-emotional development (Black et al. 2017). The first five years of life lay the groundwork for lifelong development (Shonkoff and Phillips 2000), and skills developed prior to school entry help determine children's academic success. It is important to assess children during this vulnerable period to determine if they are developing appropriately, and to design interventions if they are not. The accurate measurement of a young child's abilities—which may reflect future potential and productivity—is essential for understanding the immediate and long-term impacts of such interventions and for informing policy and practice.

We present three key reasons for assessing child development outcomes in Figure 1.1. Early childhood development assessments can be used for population monitoring, program evaluation, or exploratory research, among other things. A fourth purpose is to screen children to identify those who are at risk for developmental impairment and may require referral to a specialist for further testing, diagnosis, and treatment. In this Toolkit, we have not focused on screening tools to identify neurodevelopmental disorders, because diagnosing and treating individual children ethically requires specialized clinical training and certification.

The reason for assessing child development will drive the decision for what type of assessment to use.

The reason for assessing child development will drive the decision for what type of assessment to use. For example, a program evaluation may want a more detailed assessment of child development in all domains, whereas a population monitoring priority would be to select a test that could be administered on a large scale. For national monitoring, alignment of the ECD assessment with the content of national standards for preschool and primary grades may be an important priority to ensure policy relevance, while a program evaluation assessment could include exploratory outcomes measures. For hypothesis-driven or exploratory research, longer and more sensitive batteries of tests may be required. In most cases, alignment with cultural and national standards increases the odds of detecting relevant program or policy effects.

FIGURE 1.1 Three Primary Reasons for Assessing Child Outcomes

1

GLOBAL OR NATIONAL POPULATION MONITORING

Goal: Detecting broad trends in child development to inform policy

Application: May be intended to be comparable across populations; may not be sufficiently detailed to be sensitive to interventions

Requirements: Alignment with content of national standards for preschool and primary education to ensure policy relevance

2

PROGRAM EVALUATION

Goal: Demonstrating impacts of specific programs or policies

Application: Must be sufficiently detailed to quantify impact on child development

Requirements: Alignment with program or policy goals to detect possible range of impacts; alignment with cultural and national standards to detect program effects relevant to local policy

3

HYPOTHESIS-DRIVEN OR EXPLORATORY RESEARCH

Goal: Exploring a range of impacts on child development in line with theory and existing understanding of neural mechanisms

Application: May be sensitive to wider range of effects, both predicted and not specifically predicted, enabling new discovery; may use new technologies to advance the field

Requirements: Alignment of the method to the local culture and context to ensure valid results

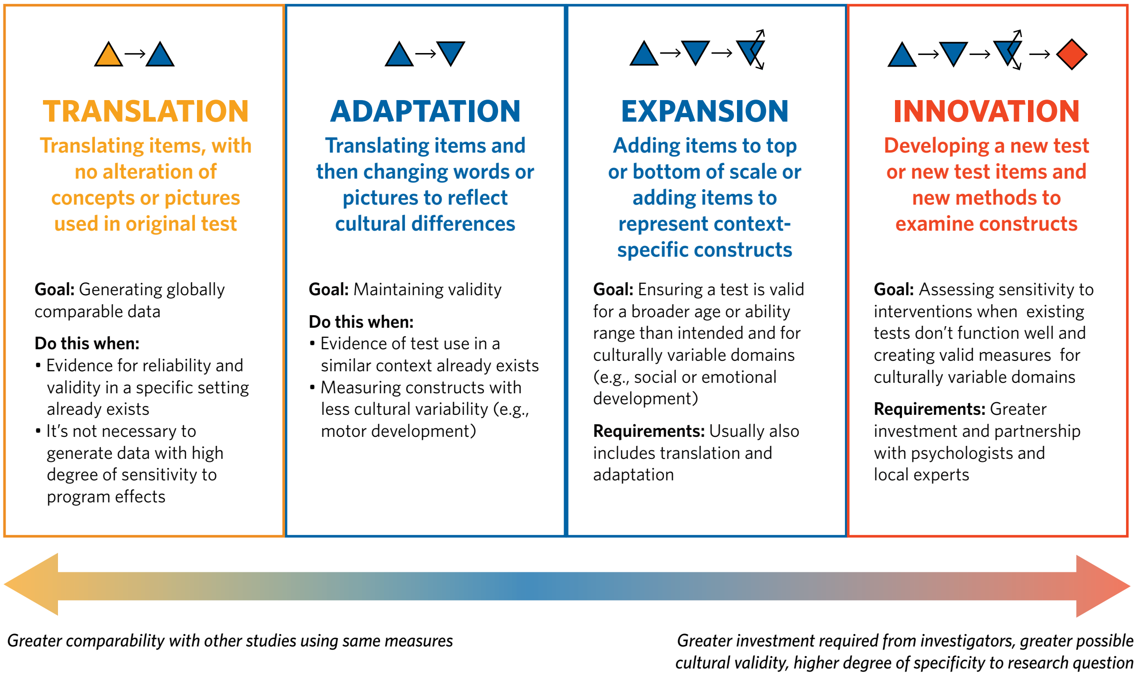
Regardless of the reason for assessing children, the "ideal" ECD assessment is characterized by features such as high reliability and validity, cross-cultural appropriateness, and ease of administration. In reality, no matter what assessment is selected, many of these ideals will be compromised, as demonstrated in Table 1.1.

After selecting the goals for the assessment, the next steps are to choose the instrument or tools, and to translate and adapt the instruments for the local context. Because many early childhood development assessments have been developed and used in high-income countries, many instruments will need modification before use in low- or middle-income countries to maintain validity. In some cases, modifications may be as simple as replacing existing pictures with locally appropriate options, whereas in other cases, more substantial modifications may be needed. Just as selecting what type of assessment to use involves a trade-off among different priorities, choosing a level of adaptation also requires a trade-off between making the modifications necessary for local validity while maintaining enough commonality to compare results across countries and contexts. With the growing interest in using standardized tests throughout low- or middle-income countries, it may be advantageous not to change the original test more than necessary. A framework for determining the degree of adaptation required is shown in Figure 1.2.

TABLE 1.1 **Child Development Assessment: Ideal and Reality**

IDEAL	REALITY
The test score represents the child's true ability in a certain domain.	Every assessment method introduces measurement error. A test that has high reliability and validity minimizes such error.
The test is appropriate, interpretable, and has high reliability and validity in all contexts and cultures, including groups with different ethnic and socio-economic backgrounds within the same country.	Test items and procedures that are appropriate, reliable, and valid in one context or group may not be so in another.
The test shows variance in scores at all ages and ability levels.	Many tests are appropriate only for a limited age range, while children outside that age range score at floor (minimum score) or ceiling (maximum score). Screening tests are not designed to show variance in typically developing children, who normally score at ceiling.
The test is relatively easy to administer.	Many tests require high levels of training and expertise to administer.
The test can be administered quickly and at low cost.	Many tests are time-consuming and expensive to administer.
The test provides information on all developmental domains.	Assessing additional domains adds to the time and resources required for training and administration.
The test score is relevant to a child's practical function in daily life, and therefore relevant to inform policies and programs.	The practical relevance of many tests of low-level cognitive abilities and neural measures has not yet been quantified.
The test is a good indicator of future success.	Child development continues to be malleable throughout childhood, reducing the predictive validity of early assessments. The predictive validity of many tests is not known, especially in low- and middle-income countries.
The specific brain systems and neural mechanisms underlying test performance are well-understood.	For many tests, especially those measuring global cognitive function, the underlying neural systems and mechanisms are not well-understood.
The impact of health, nutrition, and environmental factors on the test score is well-understood.	Due to small numbers of studies and heterogeneity in measurement tools across studies, it is generally not known which specific tests are particularly sensitive to specific exposures common in low- and middle-income countries.

FIGURE 1.2 **Framework for Adapting Assessment Tools**



TRANSLATION
Translating items, with no alteration of concepts or pictures used in original test

Goal: Generating globally comparable data

Do this when:
- Evidence for reliability and validity in a specific setting already exists
- It's not necessary to generate data with high degree of sensitivity to program effects

ADAPTATION
Translating items and then changing words or pictures to reflect cultural differences

Goal: Maintaining validity

Do this when:
- Evidence of test use in a similar context already exists
- Measuring constructs with less cultural variability (e.g., motor development)

EXPANSION
Adding items to top or bottom of scale or adding items to represent context-specific constructs

Goal: Ensuring a test is valid for a broader age or ability range than intended and for culturally variable domains (e.g., social or emotional development)

Requirements: Usually also includes translation and adaptation

INNOVATION
Developing a new test or new test items and new methods to examine constructs

Goal: Assessing sensitivity to interventions when existing tests don't function well and creating valid measures for culturally variable domains

Requirements: Greater investment and partnership with psychologists and local experts

Greater comparability with other studies using same measures

Greater investment required from investigators, greater possible cultural validity, higher degree of specificity to research question

The goal of the adaptation process could be to develop a locally appropriate tool that is equivalent to the original tool, or to assess the same underlying ability or construct in a way that is appropriate in the local setting. Achieving the first goal is more difficult, but allows comparability of scores across studies and contexts. Achieving the second goal allows comparison of groups within the same context—for example, comparing an intervention and control group in an impact evaluation—but test scores may not be comparable to other studies and contexts.

Some research teams may decide to create their own tests rather than adapting existing tests, but this is a time-consuming, laborious, and expensive process, and it is very difficult to do well. Successful generation of new tests involves having an inter-disciplinary research team, an adequate representative sample for testing items and test cohesion, and the concurrent establishment of norms or standards that represent typical child development.

In this Toolkit, we recommend a series of steps to ensure that the measurement is as reliable as possible within each new context, and subsequent chapters elaborate what is entailed in each step.

STEP 1: Produce an accurate translation of the test and the underlying construct(s) to make sure it is culturally and contextually appropriate.

STEP 2: Adapt test content to the local context by working with local stakeholders and researchers to ensure cultural alignment.

STEP 3: Adapt the test administration procedures to the local context, conducting pre-piloting and pilot tests.

STEP 4: Plan and implement a process of iterative adaptation and evaluation of the measurement.

STEP 5: Carefully record all changes made to any assessment.

.

2 Early Childhood Development and Its Determinants

KEY MESSAGES:

- Child development represents a dynamic interplay between biological and environmental factors.
- Evidence from impact evaluations in both high- and low-income settings suggests that development is malleable and can be improved by interventions affecting the child.
- Any assessment of child development should be accompanied by a measure of the quality and quantity of nurturing care that the child experiences in his or her environment to aid the interpretation of developmental scores.
- The breadth and depth of behaviors that can be assessed increase with age, and the advancement in communication and other skills during the preschool and early primary years provides additional modes for testing.

CHILD DEVELOPMENT REFERS TO THE BEHAVIORAL, BIOLOGICAL, PHYSIOLOGICAL, AND psychological changes that occur as a child transitions from a dependent infant to an autonomous teenager. These changes include the development of language (e.g., babbling, learning words, sentence construction), cognitive skills (e.g., symbolic thought, memory, logic), motor skills (e.g., sitting, running, pencil grip), and social-emotional skills (e.g., a sense of self, empathy, ability to interact with others), among other domains.

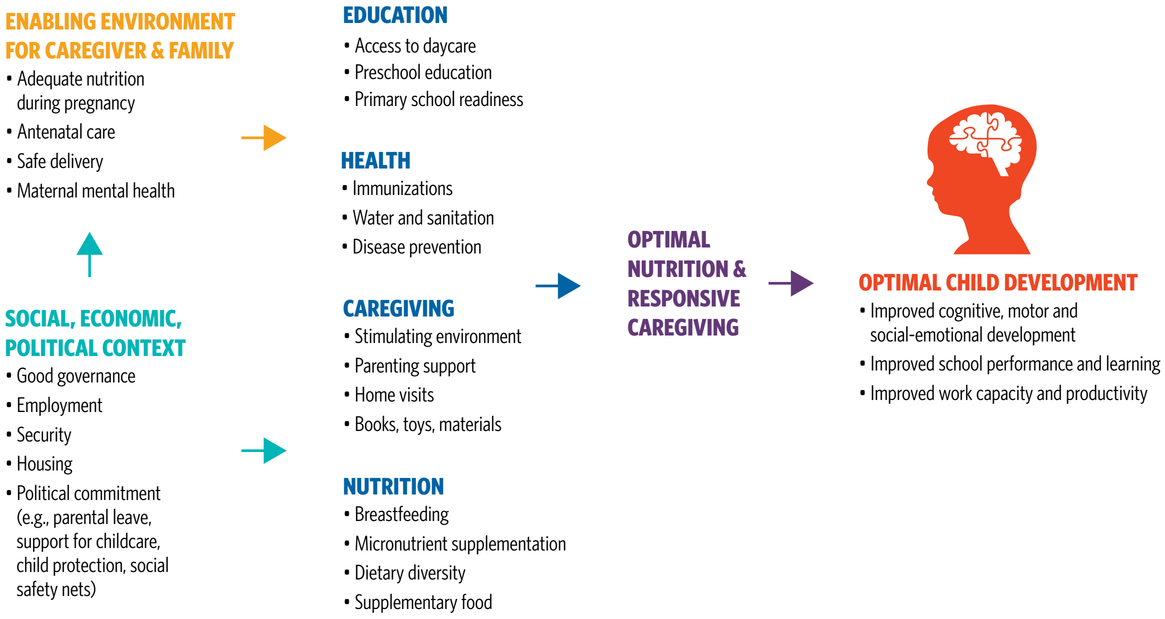
It is now well accepted that development is a process that is not determined independently by nature or nurture alone, but by "nature through nurture" (Shonkoff and Phillips 2000). Changes throughout development result from multidirectional interactions between biological factors (genes, brain growth, neuromuscular maturation) and environmental influences (parent-child relationships, community characteristics, cultural norms) over time (Shonkoff and Phillips 2000; Gottlieb 1991; Pollitt 2001). These interactions lead to the reorganization of various internal systems that allows for new developmental capacities (Thelen 2000). For example, the emergence of locomotive skills results from the co-occurrence and interactions among physiological systems (muscle strength; the ability to balance), social-emotional change (the motivation to move independently), and experience (adequate opportunity to "practice" the emerging skill) (Adolph 2002; Adolph, Vereijken, and Denny 1998; Adolph, Vereijken, and Shrout 2003).

The conceptualization of development as a dynamic interplay between biological and environmental factors suggests that early childhood is a time of great risk and great opportunity. Because young children have developing neuronal systems that are so plastic, they are simultaneously vulnerable to environmental influences and capable of benefiting from interventions. Thus, child development is malleable and can be enhanced by interventions affecting the child, the environment, or both.

Child development depends on the environment and care

Poverty, socio-cultural factors, and psychosocial and biological risk factors all work together to influence child development and long-term adult productivity (Grantham-McGregor et al. 2007; Walker et al. 2007). The acquisition of later skills and learning in middle childhood through adolescence and adulthood builds on foundational capacities established between preconception and early childhood. Life course effects are illustrated below in Figure 2.1, in which exposures during pregnancy affect newborn health and development, which subsequently affect development in early and middle childhood. Preventing exposure to risks or intervening to reduce their effects on development enhances a child's capacity to reach his or her developmental potential.

FIGURE 2.1 The Role of Context, Environment, and Caregiving in Child Development



ENABLING ENVIRONMENT FOR CAREGIVER & FAMILY
- Adequate nutrition during pregnancy
- Antenatal care
- Safe delivery
- Maternal mental health

SOCIAL, ECONOMIC, POLITICAL CONTEXT
- Good governance
- Employment
- Security
- Housing
- Political commitment (e.g., parental leave, support for childcare, child protection, social safety nets)

EDUCATION
- Access to daycare
- Preschool education
- Primary school readiness

HEALTH
- Immunizations
- Water and sanitation
- Disease prevention

CAREGIVING
- Stimulating environment
- Parenting support
- Home visits
- Books, toys, materials

NUTRITION
- Breastfeeding
- Micronutrient supplementation
- Dietary diversity
- Supplementary food

OPTIMAL NUTRITION & RESPONSIVE CAREGIVING

OPTIMAL CHILD DEVELOPMENT
- Improved cognitive, motor and social-emotional development
- Improved school performance and learning
- Improved work capacity and productivity

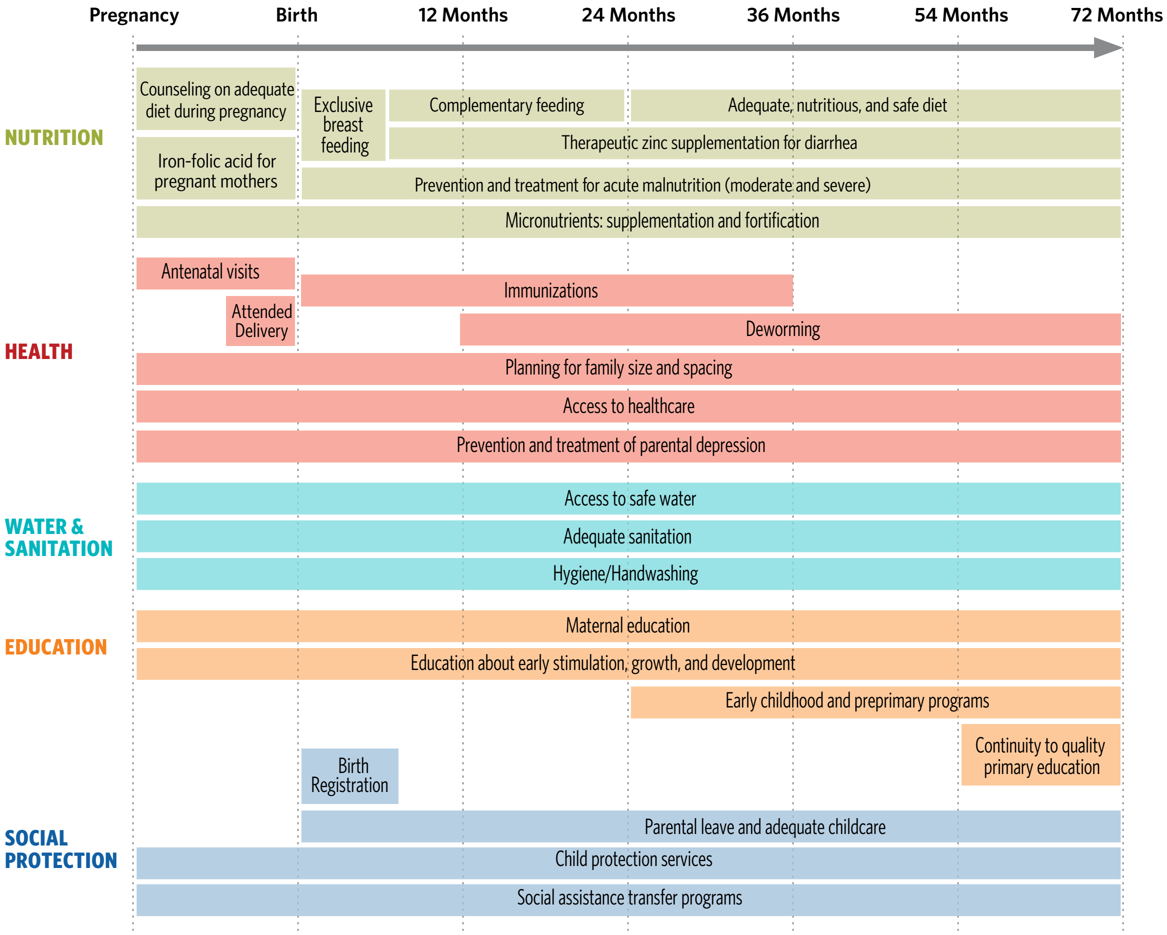
Source: Adapted from Black et al. 2017.

A child's development is determined by the integrity and function of the central nervous system and by positive and negative environmental factors that affect development. Positive environmental factors include many elements of nurturing care, including health (e.g., disease prevention, immunizations, improved water), nutrition (e.g., dietary diversity, macronutrients and micronutrients, breastfeeding), security and safety (e.g., early interventions for vulnerable children, birth registration), responsive caregiving (e.g., home visits, caregiving, support for emotional development), and early learning (e.g., access to quality childcare and preschool, learning materials). These factors are supported by an enabling environment for the caregiver, family, and community, and by social, economic, political, climatic, and cultural contexts. Biological risk factors, such as malnutrition, can influence development by affecting a child's behavior—for example, causing him or her to fuss more or play less—and by directly altering brain development and function (Prado and Dewey 2014). Poverty and socio-cultural factors, such as social marginalization, also increase the likelihood of both physiological and behavioral deficits. When children are not in stimulating and responsive environments, it is unlikely that they will demonstrate the same competencies as children who are in stimulating and otherwise positive environments.

There is substantial diversity in the types of achievements that children demonstrate during the first eight years. Some developmental achievements are more "canalized" than others, meaning that they are on a particular trajectory in which both nature and timing are strongly affected by biological maturation (Bretherton et al. 1979; McCall 1981). Walking and talking are examples of traits that all healthy individuals ultimately demonstrate in the early years, although the timing in which they emerge can vary according to environmental factors. Children in impoverished environments will appear increasingly dissimilar in developmental competencies from their higher socio-economic peers as they grow older (Wagstaff et al. 2004).

In 2014, the World Bank reviewed the existing evidence about the best interventions for young children and their families (Denboba et al. 2014). The evidence shows that interventions in several sectors can improve outcomes for children in lower- and middle-income countries. These sectors include nutrition, health, water and sanitation, education, and social protection, underscoring the importance of both the physical environment that children are exposed to and the level of nurturing care that they receive (Figure 2.2).

FIGURE 2.2 **Intervention Options for Young Children and Families**

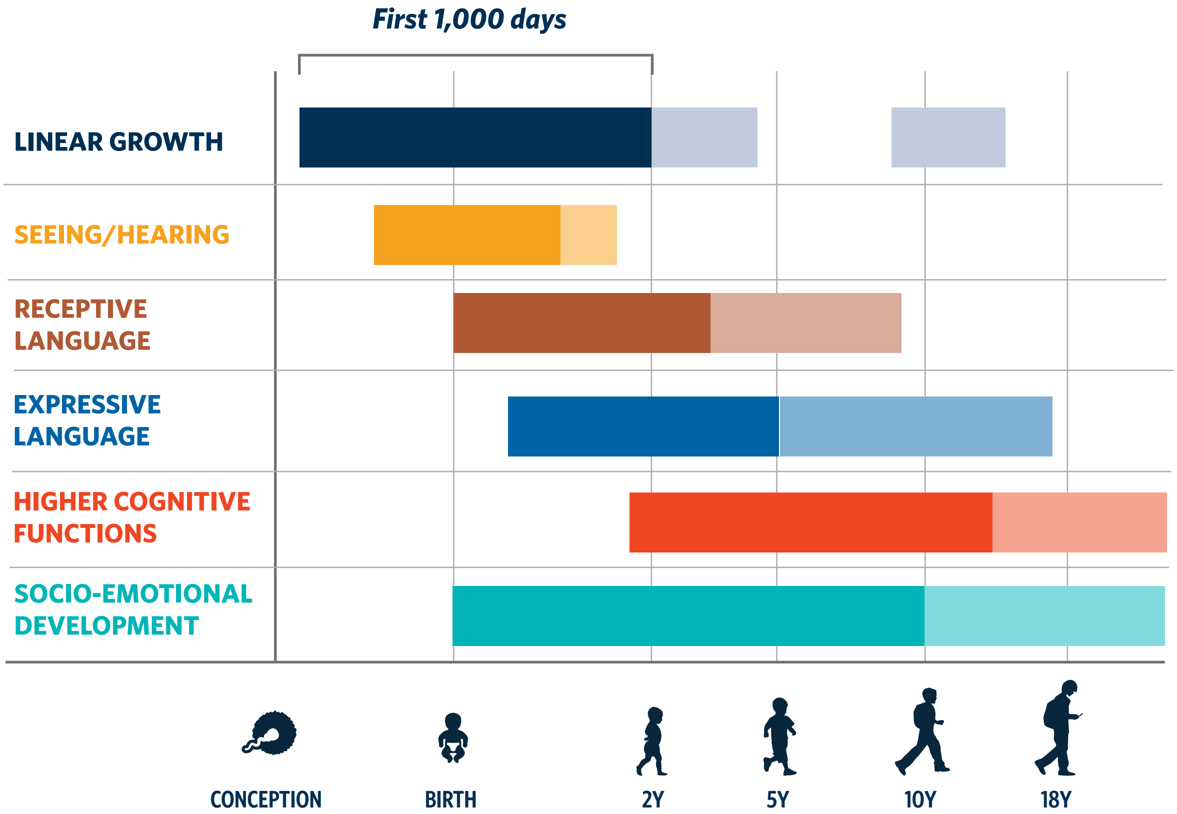


Source: Denboba et al. 2014.

Continuity of development across childhood

Development begins very early in life. Neurodevelopmental processes (neuron proliferation, axon and dendrite growth, and synaptogenesis) begin during gestation and continue throughout infancy, with a different timeline of maturation in different areas of the brain (Figure 2.3). Groups of neurons form pathways, which are refined through the elimination of cells and connections, and this process of neural pathway refinement is affected by a child's experience. This process is a primary mechanism of brain plasticity, allowing the brain to organize itself to adapt to the environment and reorganize itself to recover from injury during development (Couperus and Nelson 2006). Early childhood is characterized by developmental spurts and plateaus (Shonkoff and Marshall 2000). Rapid brain and physical development, social relationships, and environments work together to create phenomenal advances in children's abilities during this time frame.

FIGURE 2.3 **A Timeline for Human Brain Development**



Bars depict periods important for the development of each domain. Darker shading denotes critical periods of development.

Source: Adapted from Grantham-McGregor et al. 2007 and Thompson and Nelson 2001.

New capacities emerge continually and often, in close succession, as developments in one domain are catalysts for development in another. Similarly, children who are slow to develop in one domain (e.g., understanding language) may have limited capacity to display the skills that they possess in other domains (e.g., cognitive tasks that require language skills). There are sensitive periods for the association of adversities with early childhood development (see Sidebar 2.1), and optimal windows for intervention. Figure 2.4 illustrates the importance of interventions across a lifetime—from pregnancy to birth and early childhood through adulthood—that contribute to healthy development.

The speed of a child's development can vary over time, and a child's progression in any particular domain may be unstable rather than advance steadily over time (Pollitt and Triana 1999; Darrah et al. 1998). This variability reflects the fact that development results from interactions among child characteristics, environmental factors, and the demands of the developmental task(s) at hand, and that during periods of rapid change, development tends to occur in one domain at a time. From a clinical perspective, a child with discordant development raises some concerns, and the recommendation is generally to repeat the evaluation or to monitor the child more closely than usual. For example, for a child whose language is delayed, hearing may be a concern that could be evaluated. Motor delay (or abnormality) may also warrant additional attention.

After the first two years of life, development throughout the lifespan becomes more stable and proceeds in trajectories with development at one phase laying the groundwork for development at the next. These trajectories can change over time, but also have a good deal of continuity from one stage to the next. Continuity results from several influences, including the tendency for home and learning environments

SIDEBAR 2.1 The Impact of Adversity on Early Development

Stunting

- Stunting during the prenatal period and the first 24 months after birth is strongly associated with later cognition, executive function, and school attainment. After 24 months, the association is not as strong.
- Some catch-up is possible in height-for-age after 24 months, with uncertain cognitive gains.
- Early supplementation has long-term benefits on wages, but no benefit occurred with supplementation after 36 months.

Poverty

- Poverty is associated with deficits in language and cognition at three years of age. These deficits are larger at five years. Deficits are evident from the first year of life.
- Changes in poverty after age 36 months can affect cognitive development and executive function.

Severe psychosocial deprivation

- Institutionalized children exhibit profound deficits in multiple domains of child development.
- Children in institutions had changes in brain microstructure white matter; foster care caused some improvement, regardless of placement time.
- Foster care can improve IQ, attachment, and brain function. Stress responses, language development, and mental health can improve when children are placed in foster care before 24–26 months.
- Children fostered before age 15 months have been shown to catch up with their environmental peers in language development; children placed after 24 months show less improvement.

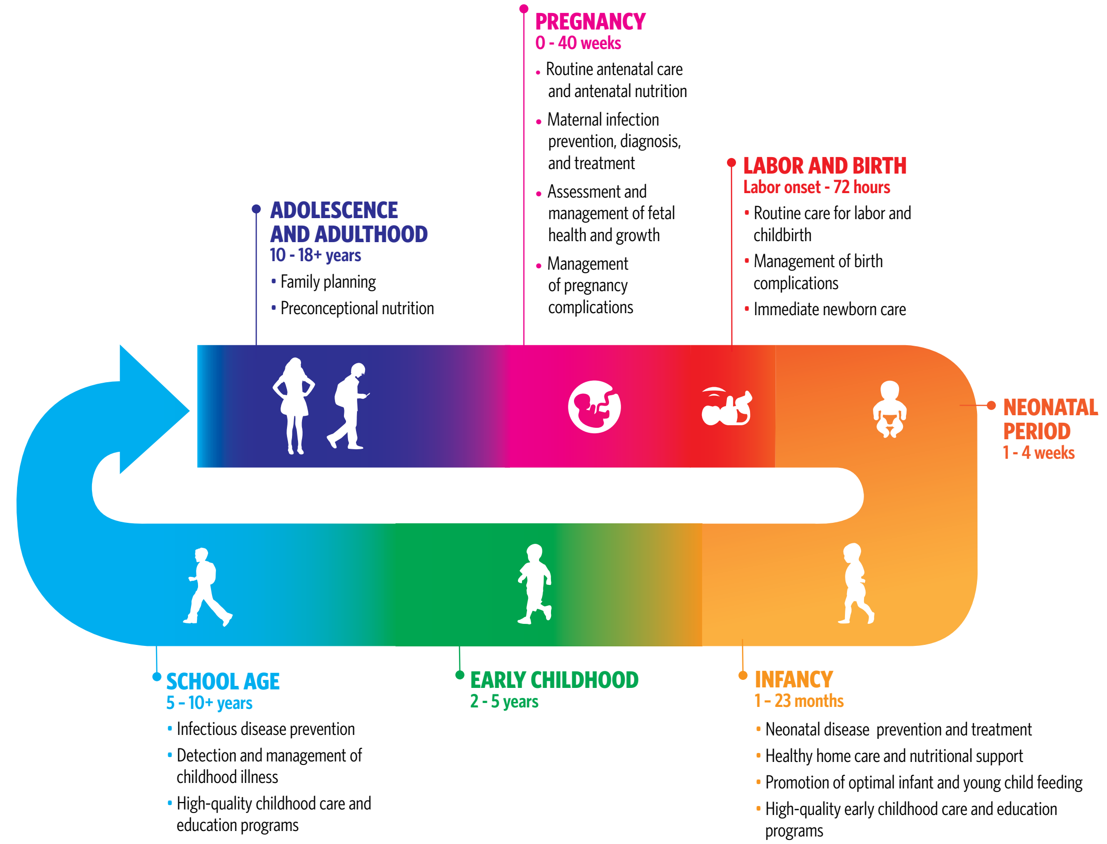
Source: Adapted from Black et al. 2017.

to remain either high- or low-quality over time; the role of biological influences on development; and the temperament of children themselves, as children's experiences at early ages influence how they interpret and learn from later experiences. Children with good language skills at one age, for example, also have more ability to learn through language at the next stage, leading to faster skill acquisition in language and literacy throughout childhood.

The general pattern of continuity in development means that children who have either high or low scores within one developmental domain at one age may also show a similar pattern later in life. Such a development pattern suggests a straightforward process for deciding what to measure. However, some areas of development do not follow linear patterns. Instead, they emerge either in non-linear patterns, with predictors showing a curvilinear relationship with an outcome in question, or as sleeper effects, in which children's abilities at one age do not predict developmental outcomes for many years, only to emerge later with strong predictive power.

The breadth and depth of behaviors that can be assessed increase with age, and the advancement in communication and other skills during the preschool and early primary years provides additional modes for testing (Snow and Van Hemel 2008). Aptitudes important for cognition and school success—e.g., pre-literacy skills, attention and focus, memory, and ability to get along with other children—can be measured at this age level. Children's environments become increasingly differentiated, and individual differences in abilities become more pronounced as children grow older (Shonkoff and Marshall 2000; Rydz et al. 2005).

FIGURE 2.4 Evidence-Based Interventions to Promote Care Across a Lifespan



PREGNANCY
0 - 40 weeks
- Routine antenatal care and antenatal nutrition
- Maternal infection prevention, diagnosis, and treatment
- Assessment and management of fetal health and growth
- Management of pregnancy complications

LABOR AND BIRTH
Labor onset - 72 hours
- Routine care for labor and childbirth
- Management of birth complications
- Immediate newborn care

ADOLESCENCE AND ADULTHOOD
10 - 18+ years
- Family planning
- Preconceptional nutrition

NEONATAL PERIOD
1 - 4 weeks

SCHOOL AGE
5 - 10+ years
- Infectious disease prevention
- Detection and management of childhood illness
- High-quality childhood care and education programs

EARLY CHILDHOOD
2 - 5 years

INFANCY
1 - 23 months
- Neonatal disease prevention and treatment
- Healthy home care and nutritional support
- Promotion of optimal infant and young child feeding
- High-quality early childhood care and education programs

INTERVENTIONS THROUGHOUT THE LIFE COURSE

PARENTING PROGRAMS
- Psychosocial stimulation
- Positive parenting and responsibility
- Maltreatment prevention

MATERNAL MENTAL HEALTH AND WELLBEING
- Assessment and treatment for anxiety, psychosis, and depression

SOCIAL PROTECTION
- Conditional cash transfers

WATER, SANITATION, AND HYGIENE (WASH)
- Ensuring access to clean water
- Creating sanitation infrastructure
- Promoting hygiene behaviors

Source: Adapted from Britto et al. 2017.

At an early age, the rate of emergence of abilities differs considerably among children. When the emergence of a child’s ability is significantly slower than average for age, the child is considered to be “delayed” in terms of that ability; this delay is usually defined by being below a certain cutoff based on nationally representative norms. "Delay" is always determined relative to normative development within a given population. Cutoff scores that define delay in one population cannot be assumed to define delay in another. Delays as well as abilities become evident with age, and problems in specific areas are not apparent until the child reaches an age when those skills are typically learned and can be effectively evaluated (Rydz et al. 2005; Glascoe 2001). A child with no apparent delays in communication or cognitive skills at three years may nevertheless be diagnosed with reading difficulties at six years (Glascoe 2001). Continued testing and tracking through school-age years is important for evaluating the long-term benefits of programs and interventions that begin early in life (Shonkoff and Marshall 2000; Snow and Van Hemel 2008; Rydz et al. 2005; Glascoe 2001).

Poverty increases the risk of delayed development

Compared to children in high-income countries, children in low- and middle-income countries are more likely to be vulnerable to deficiencies in basic health and nutrition. These deficiencies contribute to delayed physical and cognitive development.

Infants and children growing up in poverty are more likely to be exposed to poor sanitation, crowded living conditions, inadequate diets, lack of psychosocial stimulation, and fewer household resources (Walker et al. 2007)—co-occuring risk factors that can negatively impact development and growth (Bradley and Corwyn 2002; Brooks-Gunn et al. 1995; Bolig, Borkowski, and Brandenberger 1999). Not surprisingly, significant associations exist between low height-for-age (stunting) and other deficits, such as delayed cognitive and psychomotor development, poor fine motor skills, and altered behavior (Sudfeld et al. 2015). There is increasing evidence that economic status in the United States is associated with children's brain development and function. Compared to better-off peers, low-income children show restricted development of the hippocampus and frontal and temporal lobes (Hanson et al. 2015; Noble et al. 2015), and these alterations are associated with deficits in language, reading, memory, visuospatial skills, and executive function (Noble, Tottenham, and Casey 2005).

Experiencing poverty during childhood can have permanent effects. Research in the United States suggests that the developmental scores of children in low-income households are in the normal range during infancy and then decline in comparison to normal samples during the preschool years; this pattern is not apparent in middle-income samples (Black, Hess, and Berenson-Howard 2000). Socio-economic disparities in child development scores have also been consistently found in lower- and middle-income countries, with the gap between rich and poor increasing from infancy throughout childhood (Fernald et al. 2011; Lopez Boo 2016; Paxson and Schady 2007; Hamadani et al. 2014; Schady et al. 2015; Rubio-Codina et al. 2015). When they grow up, children living in poverty in these countries are likely to have substantially lower wages than healthier adults (Boissiere, Knight, and Sabot 1985), and are thus less likely to be able to provide increased stimulation and resources for their own children, thereby perpetuating the cycle of poverty (Sen 1999).

Exposure to stress and a child's physiological stress response may mediate or moderate the effects of poverty on development (Hackman et al. 2015). For example, environmental adversity may cause high levels of chronic stress, and therefore chronic exposure to elevated levels of stress hormones, which can affect brain and behavioral development (Lupien et al. 2009).

Some evidence suggests that the effects of early adversity on stress response can be reversed through later nurturing care. Among a group of institutionalized Romanian children, those who were randomized to foster care showed a normal stress response to a psychosocial stressor in later childhood, while those who remained institutionalized showed an abnormal response (McLaughlin et al. 2015). This highlights the plasticity of early child development and the existence of continual windows of opportunity for intervention throughout childhood.

Although children who are exposed to deprivation and adversity are at risk for poor developmental outcomes, some children thrive despite such exposure (Boyce and Ellis 2005). Researchers have identified several traits that may modify children's susceptibility to environmental influences, including temperament (Stright, Gallagher, and Kelley 2008), physiological and emotional reactivity to stressful events (Obradović 2016), and certain genetic variations (Belsky and Pluess 2013). Besides these individual-level factors, environmental-level supports, such as responsive caregiving, can also mitigate the effects of risks.

Norms for development can differ across cultures

Child development occurs within a social context and culture. Culture refers to a set of beliefs, values, goals, attitudes, and activities that guide the manner in which a group of people live (Payne and Taylor 2002). Any particular culture is shaped by a broad spectrum of factors, such as geography, religion, political and economic structures, access to educational and health care systems, and the degree to which modern technology is present. Parenting practices and ideas about child development are largely determined by cultural ideals. Cross-cultural studies of development aim to distinguish which skills and abilities are universal from those that are culture-specific or are unique to an individual (Carter et al. 2005).

Cultures have a wide range of values for the skills and abilities that children should develop and when they should be exhibited (i.e., "norms" or normative ages when skills are typically displayed). Abilities may emerge earlier if they are valued and encouraged in a particular culture. However, this does not mean that unencouraged abilities will not emerge at some point. Ideally, these culturally specific patterns can be considered in assessing the validity of a measurement, and are of particular concern when comparisons are made across population or ethnic groups or across countries. If competencies are valued in one culture but not the other, any disparities that emerge between cultures can be easily misinterpreted. When comparisons are made within a group (e.g., intervened versus control), the concern is limited to being sure that the measurement used is actually measuring the capacities that the intervention was designed to change.

As higher levels of educational attainment become more universal, the necessary skills for life success become more consistent across cultures. These include not only academically related skills, such as language and symbol recognition, but also social skills such as the ability to function in groups, wait for a turn, or inhibit an initial response. These skills are useful not only for school but also for overall productivity and adaptability throughout later life.

Cultures have a wide range of values for the skills and abilities that children should develop and when they should be exhibited (i.e., "norms" or normative ages when skills are typically displayed). But as school becomes more universal, the skills necessary for the future become more consistent across cultures.

There is no simple way to ensure cross-cultural comparability of early cognitive tests. An extreme position suggests that each culture is totally unique and requires special measurement methods. In contrast, a position that holds that all children should be judged by the same measurement tool—even if well adapted—ignores the wide range of values and ways of learning that can change how quickly abilities develop in different cultures (e.g., using rules of social conduct and respect). Certain domains of early childhood development, such as social-emotional development, are likely to be more susceptible to cultural influences than other domains, such as motor development.

Some evaluations have attempted to relate scores on measures that assess skills necessary for children to do well in school and be productive as adults (e.g., literacy and problem-solving skills) with culturally valued attributes deemed important for being successful within a particular society (e.g., responsibility for carrying out tasks necessary for daily living). Among the Yoruba in Nigeria, for example, children 22–26 months of age who were rated as more responsible by parents to purchase items or retrieve particular objects scored higher on a modified (shortened) version of the Bayley Scales of Infant Development (BSID) (Bayley 1969) than children with lower responsibility ratings (Ogunnaike and Houser 2002). This suggests that the two types of measures were related.

In Zambia, adult ratings of a school-age child's capacity to complete specific tasks were highly predictive of school grade completion and adult literacy scores; however, this finding was true only for girls (Serpell and Jere-Folotiya 2008). The authors speculated that because girls participate more in domestic chores than boys, adults may have had greater opportunity to observe and evaluate their abilities. In contrast, scores on other locally developed tests more rooted in Western notions of abilities were strongly predictive of grade attainment and literacy for boys, especially those living in urban areas, but these were not predictive for girls' later literacy scores, in particular for girls in rural areas. In rural Guatemala, test performance was associated with children's behavior—in particular their ability to complete a series of three chores without additional instruction—as well as with adults' ratings of children's "smartness" (Nerlove 1974). These examples illustrate both the links between tests and local conceptions of ability, and the complexities of using local notions of attributes to predict later capacities, and they highlight the need for scrutinizing all types of measurements.

3 Measurable Skills and Longer-Term Impact

KEY MESSAGES:

- Domains of early child development and school readiness include cognitive skills, language, executive function and self-regulation, and motor, social-emotional and pre-academic skills, among others.
- Across domains, existing assessments in the 0–2 year age range are generally poor predictors of later performance (e.g., during primary school age), but become stronger predictors by ages 3–5 years.
- Starting at age three years, the predictive validity of assessments within domains is stronger than across domains, and no single domain is the strongest predictor of both academic and behavioral performance at school age.

THE VARIOUS DOMAINS OF EARLY CHILD DEVELOPMENT AND SCHOOL READINESS include cognitive skills, language, executive function and self-regulation, and motor, social-emotional and pre-academic skills, among others. Table 3.1 lists various ECD domains and their definitions. While we have chosen these definitions for the purposes of this Toolkit, it is important to note that these domains are not always consistently defined in the literature and that varying definitions for these terms may be found. It is also evident from these definitions that many of these domains overlap with one another (e.g., executive function, social-emotional skills, approaches to learning). In addition, while developmental tasks such as walking and learning letters are divided into domains for categorical purposes, the underlying skills in different domains overlap and mutually influence each other. These domains involve overlapping skills, and a test focusing on one domain inevitably taps abilities in other domains as well.

Predictive validity

The predictive validity of a test refers to the association of test scores at one time point (time 1) with other scores or indicators collected months, years, or decades later (times 2, 3, and so forth). A strong association between a score at time 1 and a later time point can be interpreted to mean that scores at time 1 are a meaningful indicator of a child's future ability, while a weak association can be interpreted to mean that scores at time 1 are not relevant for later knowledge, function, or performance. For example, an association between a score on an expressive vocabulary assessment at age 18 months and a score on a reading comprehension assessment at age 18 years implies that children who say more words at an early age are likely to be better readers in high school. The lack of such an association implies that how many words a child says at an early age is not relevant to reading ability in high school.

Many early childhood interventions are implemented because they are intended to affect children's development later in life, reflecting the strong scientific evidence demonstrating that early development lays the groundwork for later development. Clearly, it is important for researchers to consider the predictive validity of a test, especially in the case when they want to select tests that measure something now while also meaningfully telling them something about the future.

TABLE 3.1 **Description of ECD Domains**

DOMAIN	DESCRIPTION
COGNITIVE SKILLS	The processes or faculties by which knowledge is acquired and manipulated, including abilities such as memory, problem solving, and analytical skills
LANGUAGE SKILLS	The ability to understand and express verbal communication
MOTOR SKILLS	The ability to control and coordinate gross movements of the legs and arms (e.g., jumping, throwing) and fine movements of the fingers
EXECUTIVE FUNCTION/SELF-REGULATION/EFFORTFUL CONTROL	Intentional control over behavior and cognition. Executive function includes abilities such as inhibitory control, cognitive flexibility, attention, and working memory
TEMPERAMENT	Biological influences on the experience and expression of emotion, including extraversion/surgency (positive affect, activity level, impulsivity, risk-taking), negative affectivity (fear, anger, sadness, discomfort), and effortful control (attention shifting and focusing, perceptual sensitivity, inhibitory and activational control)
SOCIAL-EMOTIONAL SKILLS	The regulation of emotional responses and social interactions, which is a function of both temperament and self-regulation, including behavior problems, social competency, and emotional competency
PERSONAL-SOCIAL/ADAPTIVE SKILLS	The ability to perform daily-life skills, such as self-feeding, dressing, toilet training, interacting with others, and adjusting to new situations
PRE- AND EARLY-ACADEMIC SKILLS	Skills needed to learn reading and math, such as counting and letters
APPROACHES TO LEARNING	Behaviors related to how children become engaged in learning experiences, such as the ability to stay focused, interested, and engaged in activities

Studies of predictive validity in the United States and in low- and middle-income countries have generally found weak associations between scores for existing developmental tests taken by children aged under two years and their abilities in later childhood (Snow and Van Hemel 2008; Bracken 2007). Current recommendations support the use of comprehensive assessments in children under two for measuring concurrent abilities and identifying severe delay, but caution against using such scores for predicting future development (Snow and Van Hemel 2008; Bradley-Johnson and Johnson 2007). Starting at age three years, the predictive validity of assessments within domains is stronger than across domains, and no single domain is the strongest predictor of both academic and behavioral performance at school age (Duncan et al. 2007; Halle et al. 2012; La Paro and Pianta 2000). Thus, children need to develop a constellation of skills in early childhood to reach success in all areas of achievement and behavior.

Many researchers want to use measures that have already demonstrated predictive validity. However, conducting and interpreting predictive validity studies can be challenging. First, longitudinal studies are costly in terms of both time and financial and material resources, which is why they are not often done. In the best scenarios, longitudinal follow-ups repeatedly test the same children at infancy and then follow them throughout childhood and into adulthood. For many tests, though, predictive validity studies have not been conducted, and for others, the intervening time between time 1 and time 2 is only a few years. Most predictive validity studies have been conducted in high-income countries, and may not be generalizable to low- and middle-income countries. Evidence of predictive validity does not yet exist for tests that have been recently developed and published because not enough time has elapsed for a meaningful assessment.

Second, it can be very difficult to keep track of participants over time, resulting in high attrition at the later time points. Even if the sample at time 1 was representative of the population, attrition can result in a biased sample at time 2, which compromises the generalizability of the results. Although sample attrition can be addressed with post hoc statistical techniques, it makes the analytic process and interpretation more complex.

Third, socio-demographic characteristics are likely to confound the association between scores at time 1 and later time points. For example, children from households with low socio-economic status may have both lower expressive vocabulary at 18 months and lower reading comprehension at 18 years. The raw correlation between scores at time 1 and time 2 would be inflated by this confound. Adjusting for socio-economic status, however, could also adjust out meaningful variance and mask a true correlation (Snow and Van Hemel 2008; Duncan et al. 2007).

A fourth related reason results from the fact that the interpretation of predictive validity studies can be clouded by environmental influences that occur in the intervening period between time 1 and time 2. Often, these are not measured or controlled in such studies (Marks et al. 2008). For example, children with high scores at time 1 who attend a poor-quality school may score lower at time 2, while children with low scores at time 1 who attend a high-quality school may score higher at time 2. This would result in a low correlation between time 1 and time 2, even though a high correlation may have been found if school quality had been taken into account.

In this chapter, we review the evidence for predictive validity of various early childhood development domains. However, for the reasons stated above, we caution against placing a very high priority on this criterion when selecting ECD tests. The predictive validity of ECD tests in low- and middle-income countries is a research question greatly in need of further evidence, and we encourage those with longitudinal datasets to publish such analyses.

Cognitive skills

Cognition is the processes or faculties by which knowledge is acquired and manipulated, including abilities such as memory, problem-solving, and analytical skills (Bjorklund and Causey 2017). For infants and toddlers, early cognitive development involves problem-solving with objects, such as learning to stack or nest objects, and early understanding of math, demonstrated by such behaviors as sorting objects and knowing what it means when someone asks for “one” or “two” of something (Damon, Kuhn, and Siegler 1998). By age three years, many children in high-income countries are capable of solving simple puzzles and matching colors and shapes and also show awareness of concepts such as “more” and “less.” As children approach school age, they are able to process and learn more complex information, while showing individual differences in cognitive skills such as memory, processing speed, and logical reasoning.

Cognition is the processes or faculties by which knowledge is acquired and manipulated, including abilities such as memory, problem-solving, and analytical skills

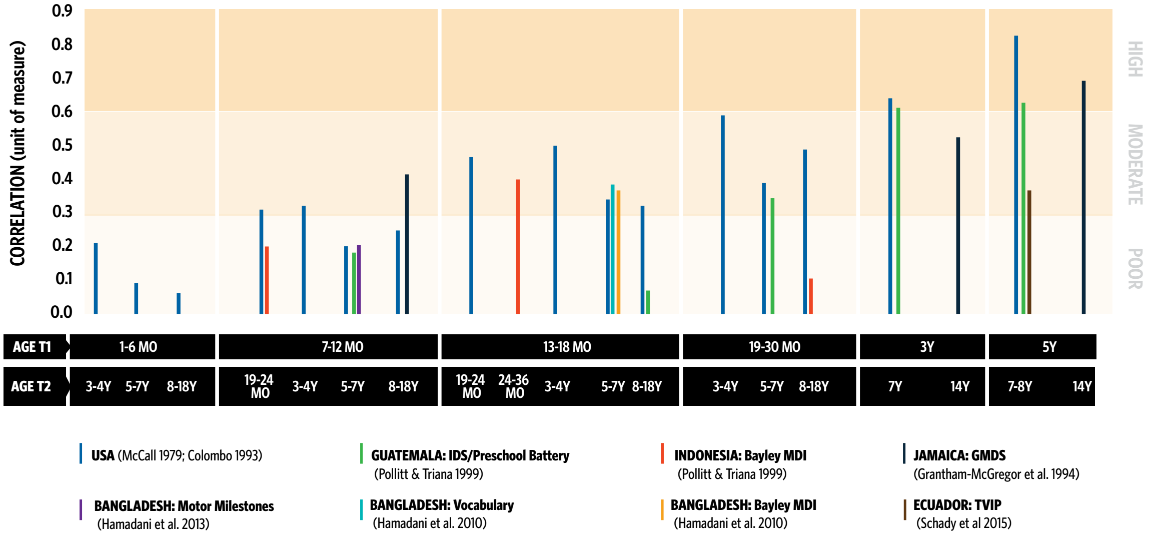
Genetic influences are generally considered to account for approximately half of the variance in cognitive abilities based on studies of identical twins (Kovas et al. 2007). According to one estimate, the heritability of intelligence, which is the variance in intelligence accounted for by genetics, increases linearly, from 20 percent in infancy to 40 percent in adolescence, and to 60 percent in adulthood (Plomin and Deary 2015).

Although genetics play a role in a child’s developing abilities, evidence also shows the importance of genetic and environment interactions in how those genes are expressed, and in the important role that environmental variations play in development and education. Children with responsive caregivers, and those who are in more stimulating environments, are more cognitively advanced at the start of school than children in less stimulating homes; parents who interact frequently with their children promote their cognitive, social, and emotional development (Shonkoff and Phillips 2000). These environmental influences are possibly even more important in conditions of poverty, malnutrition, and ill health.

Predictive validity of early cognitive assessments

As discussed previously, continuity in cognitive development throughout infancy and childhood arises from both genetic influences and the tendency for home and learning environments to remain consistent in terms of quality over time. In both high- and low-income countries, the predictive validity of general mental development assessments, such as the Bayley Scales of Infant Development (BSID) (Bayley 2006), is low to moderate for children under age two years, with correlations in the range 0 to 0.5, and increases for children around age three to five years, to correlations in the range 0.5 to 0.8 (Figure 3.1). Information-processing measures in infancy, such as Fagan's novelty test, have also shown low to moderate correlations with later childhood IQ, in the range 0 to 0.4 (Fagan, Holland, and Wheeler 2007; Rose et al. 2012; Andersson 1996; Tasbihsazan, Nettelbeck, and Kirby 2003).

FIGURE 3.1 **Correlations Between Early Developmental Scores and Later Performance**



METHODS USED IN EACH STUDY

USA
Median correlation across multiple studies of the correlation between infant/preschool scores at time 1 (T1) and later IQ at time 2 (T2)

BANGLADESH
Median correlation across age of acquisition of six WHO motor milestones with IQ at T2

GUATEMALA
Infant Development Scale (IDS) and preschool battery developed for that study predicting a performance on a battery of psychoeducational tests at T2

BANGLADESH
Sixty-word vocabulary checklist based on the MacArthur-Bates Communicative Development Inventories (CDI) predicting IQ at T2

INDONESIA
Bayley Scales of Infant Development Mental Development Index (MDI) used at both T1 and T2

BANGLADESH
Bayley Scales of Infant Development Mental Development Index (MDI) at T1 predicting IQ at T2.

JAMAICA
Griffiths Mental Development Scale predicting IQ at T2; sample included large proportion of children who had been severely malnourished in infancy

ECUADOR
Test de Vocabulario en Imagenes Peabody (TVIP), the Spanish version of the Peabody Picture Vocabulary Test at T1 predicting math and language scores at T2

Cognitive development and general knowledge are important for success in school and together are included in the five dimensions of school readiness specified by the United States' National Education Goals Panel (Kagan, Moore, and Bradekamp 1995). The other four Panel dimensions are physical well-being and motor development, social-emotional development, approaches to learning, and language development. In high-income countries, cognitive and general knowledge measures at school entry predict later school achievement (Halle et al. 2012). An analysis of three longitudinal datasets in low- and middle-income countries also showed that cognitive scores in the age range four to eight years predicted later school achievement and grade attainment (Grantham-McGregor 2007).

Early development clearly matters for later learning, though the size and nature of the relationship between early skills and later development may vary by setting. In South Africa, for example, eight cognitive and behavioral assessments administered from age 0–8 years were not strongly associated with delayed age at school entry or with grade repetition before grade six (Richter, Mabaso, and Hsiao 2016). Although Bayley Scales of Infant Development scores at age one year significantly predicted delayed age at school entry, the diagnostic accuracy was low. The only other assessment with significant predictive validity was the Conners' Teacher Rating Scale at age seven years. The scale is a teacher report of children's conduct in the classroom. However, the diagnostic accuracy of this tool was also low.

In Figure 3.1, we have presented the results of all studies that we found that reported predictive validity correlations in low- and middle-income countries. Studies in these countries were not included if they did not report comparable statistics (for example, they reported odds ratios, ROC curves[3] , or unstandardized regression coefficients). We also included in Figure 3.1 results from an analysis of multiple studies in the United States for comparison with results from low- and middle-income countries.

Self-regulation, effortful control, and executive function

In the first years of life, children develop increasing intentional control over their behavior and cognition. Recent work has pointed to the importance of self-regulation, executive function, and effortful control for children's development (Liew 2011), but there is not yet conceptual coherence or agreement on how best to measure these (Jones et al. 2016). Broadly speaking, self-regulation refers to children's abilities to control their emotional responses and behavior, resulting from the integration of emotion and cognition in early childhood (Blair and Razza 2007). Self-regulation refers to both the cognitive and emotional aspects of regulating behavior, attention, and emotional responses.

Self-regulation is influenced by both temperament—the biological influences on the experience and expression of emotion, which is referred to as "effortful control" in research literature—and cognition, which determines children's abilities to focus and shift attention (referred to as "executive function" in research literature) (e.g., Liew 2011 and Blair and Razza 2007). Drawing on the scientific tradition of temperament theory and research, effortful control is typically operationalized by parent or teacher report on children's behaviors, including constructs such as how easy or difficult it is to soothe children, the degree of emotional reactivity (e.g., how often they throw tantrums), and level of impulsivity (e.g., how often they have trouble sitting still) (Rothbart et al. 2001).

Self-regulation refers to both the cognitive and emotional aspects of regulating behavior, attention, and emotional responses.

Executive function processes include impulse control, ability to initiate action, ability to sustain attention, and persistence. Executive function is represented by children's abilities to inhibit responses, hold information in working memory, and shift attention from one set of cues to another. For example, in *Luria's tapping test*, a child is instructed to tap on a table with his or her finger or a pencil one time when the tester taps twice, and to tap twice when the tester taps once. This requires the child to inhibit the automatic response to imitate the tester. The child must use effortful control to tap once or twice according to the instructions rather than copying the demonstrated action.

An example of a working memory task is the *digit span backwards task*. In this test, the child is required to repeat increasingly longer series of digits in the reverse order as the tester (e.g., when I say 9, 1, 2, you say 2, 1, 9). This test requires the child to both hold and manipulate the information in working memory and is quite difficult as the number of digits increases.

A common set-shifting task is the dimensional change card sort task. In this task, the child is shown pictures that differ in both color and shape, such as red trucks, blue trucks, red boats, and blue boats. The child is first instructed to sort the pictures according to color. Then, the rule is switched and the child is instructed to sort according to shape. This requires the child to shift attention from the color of each picture to the shape of each picture.

The range of terms and definitions describing self-regulation, executive function, and social-emotional skills reflects ongoing scientific debates on the origins and developmental significance of a mix of skills related to social behavior, emotion, and attention (Liew 2011). Executive function includes a subcategory of cognitive skills, in spite of the fact that both cognitive and emotional processes are involved. The more cognitive executive function processes are linked to dorsolateral regions of the prefrontal cortex and have been called "cool" processes–such as remembering arbitrary rules and other non-emotional aspects of the task. "Hot" executive function processes have been linked to the ventral and medial regions of the prefrontal cortex and describe the more emotional aspects of executive function–those involving inhibition and, in some studies, delay of gratification (Hongwanishkul et al. 2005), although the distinction between hot and cool executive function is not yet accepted by all researchers.

The concepts of self-regulation, effortful control, and executive function are relatively new—from the

[3] Receiver operating characteristic curve (ROC curve) is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate against the false positive rate at various threshold settings.

past 20–30 years—and work on executive function specifically results from neuropsychological research on the effects of damage to the frontal lobes (Jurado and Rosselli 2007). While the field is still evolving, and definitions of executive function are variable, researchers generally agree that executive function comprises *fluid* abilities or processes that are engaged when a person is confronted with a novel situation, problem, or stimulus. These fluid abilities are distinct from *crystallized* cognition or knowledge of information (such as vocabulary) (Jurado and Rosselli 2007; Cattell 1963).

Executive function components can be measured separately, but often it is the capacity to integrate or coordinate them to solve a problem or reach a goal that is most significant to assess (Welsh, Friedman, and Spieker 2006). Tasks requiring the engagement of multiple processes are considerably more difficult than using only one process (Carlson 2005), but are more likely to reflect real-life demands.

Self-regulation as a whole, specifically both effortful control and executive function, has powerful impacts on children's development and learning, through a number of mechanisms, including the ability to focus attention, control emotions, and navigate relationships with others. Several inter-related domains comprise children's regulation in school settings, or their ability to follow rules, interact successfully with peers and teachers, focus their attention, and control their emotions. The negative effects of socio-economic status on children's school readiness in the United States may be mediated by attention processes, suggesting that low-quality environments affect cognitive development in part by decreasing children's abilities to attend (NICHD Early Child Care Research Network 2003).

Recent work has emphasized the importance of approaching the study of self-regulation using both temperament-based and cognitive approaches, by integrating measures of effortful control and executive function. There is less evidence on measurement of effortful control in low- and middle-income countries to date than in high-income countries. Yet reflecting the strong theoretical basis for predicting that self-regulation matters for children's development, this Toolkit recommends that researchers begin to include effortful control in assessments to complement emphasis on executive function.

While the roots of children's executive functioning are apparent in infancy, executive function develops considerably in early childhood, as the frontal lobe develops (Anderson 1998). In young children (3+ years), some of the processes most commonly cited as measurable are working memory (e.g., holding information in mind for a short time); inhibition of behavior or responses as demanded by the situation or task (e.g., not opening a box until a bell rings or inhibiting a response that was previously correct but no longer is); and sustaining attention as required or being able to switch attention as necessary (e.g., shifting focus from the color of a picture to the shape of the picture) (Carlson 2005).

Although some tasks assess executive function at ages 1–3 years, it is difficult to measure these skills reliably at this age. The trajectory of the development of executive function, and the areas in the frontal lobe that underlie it, continues into later childhood (Figure 2.3), and these functions are measured more reliably at later ages (4+ years).

Predictive validity of self-regulation, effortful control, and executive function assessments

Concurrent measures of executive function have been shown to predict measures of children's school achievement (e.g., Best, Miller, and Naglieri 2011). When it comes to predictive validity, performance on executive function tasks shows low to moderate continuity (r = 0.2–0.5) from infancy (age 7–11 months) through ages three to seven years (Kochanska, Murray, and Harlan 2000; Posner et al. 2014; Tsetlin et al. 2012). In the United States, higher performance on measures of attention, inhibition, and other executive functions in the preschool years is associated with later academic achievement (Duncan et al. 2007). Early development of executive function is also important for daily life beyond academic success. In New Zealand, for example, measures of self-control from age three to 11 years predicted physical health, substance dependence, personal finances, and criminal offenses at age 32 years (Moffitt et al. 2011).

The ability to delay gratification at preschool age, sometimes measured by the "marshmallow test," as it has become known, has also been found to predict later behavior in adolescence and adulthood. In one variation of this test, the child is given a marshmallow by an adult, who says that he or she is leaving the room for a few minutes. The child is told that if he or she waits until the adult returns to the room before eating the marshmallow, he or she will receive two marshmallows. The duration of time children waited before eating the marshmallow at age four years predicted later social competence, self-control, and the ability to cope with frustration and stress, with low to moderate correlations (r = 0.25–0.5), and SAT scores with higher correlations (r = 0.4–0.7) 10–20 years later (Mischel, Shoda, and Peake 1988; Shoda, Mischel, and Peake 1990). We are not aware of any studies in lower- or middle-income countries reporting associations between early executive function and later life outcomes.

Finally, the construct of "grit" has emerged in relation to achievement among older children; grit is defined as the tendency to sustain interest in, and effort towards, long-term goals (Duckworth et al. 2007). While grit is not intended to refer to development among preschoolers, some questions may arise on how it may relate to executive function and delay of gratification (e.g., Mischel and Brooks 2011). Duckworth et al. propose that both delay of gratification and presence of grit may reflect an underlying construct of "self-control," which in turn can predict children's achievement in adolescence and beyond (Duckworth, Tsukayama, and Kirby 2013). These constructs seem to be measurable during the preschool years through the delay-of-gratification task, but it is important to note that both the development of skills supporting delay of gratification and the manifestation of delay of gratification may be especially prone to cultural expectations for children's behaviors, to the point that the tasks may not be applicable outside of high-income countries (Lewis et al. 2009). Similarly, "grit" is not the same construct as executive function and is not to be considered interchangeable with executive function measurement.

Language skills

Children's language development begins long before the emergence of the first word (Bloom 1998). Early indicators of language development include babbling, pointing, and gesturing in infancy, the emergence of first words and sentences in the first two years, leading to an explosion of words between ages two and three years (Woodward and Markman 1998). As children move into the preschool years, indicators of language development include children's production and understanding of words, their abilities to tell stories and identify letters, and their comfort and familiarity with books.

Under age three years, children's vocabulary is a good indicator of language development. In cultures with a history of literacy (written language), this remains a good indicator at older ages. However, in cultures that do not have a long history of literacy, other criteria can be used. In some African cultures, for example, grammatically correct and creative use of alliteration and metaphor are a more appropriate mark of an older child who is linguistically advanced (Harkness and Super 1977).

Like cognitive and social-emotional development, language development is dependent on stimulating home environments and relationships.

Like cognitive and social-emotional development, language development is dependent on stimulating home environments and relationships. Relatively low-income children in the United States build their vocabularies more slowly than higher-income children and speak many fewer words than their higher-income counterparts by kindergarten (Hart and Risley 1995). This pattern occurs in part because low-income children receive less infant-directed speech and also because the speech that they hear has reduced lexical richness and sentence complexity, both of which contribute to vocabulary growth (Hoff 2003; Hart and Risley 1992). In addition, within low-income homes, adult speech is less responsive to children's signals, less directed to infants, and used less in the course of shared attention and shared communication (Tamis-LeMonda, Bornstein, and Baumwell 2001). Reading to children early in life also supports language development. Because children's language development is heavily dependent on their exposure to words and books in the home, children whose parents are not literate may develop speech and vocabulary more slowly (Fernald et al. 2006).

Predictive validity of language assessments

Children's language skills are also critical for their success in school, especially for learning to read and write. In school, language skills are necessary to understand and participate in discussions with teachers and other students, to develop an interest in books and stories, and to recognize letters and sounds. In high-income countries, language scores at ages 4-5 years are associated with subsequent school achievement at ages 6–15 years (Duncan et al. 2007; Halle et al. 2012). Other studies have shown the language scores at even earlier ages are predictive of later performance. In one study in Germany, for example, scores on the Receptive-Expressive Emergent Language Test (REEL) as early as age 10 months were associated with school achievement when the children were 11 years old (Hohm et al. 2007).

Although no studies have reported associations between early language measures and later school outcomes in lower- and middle-income countries, in a study in Bangladesh, children's vocabulary measured at 18 months was associated with IQ when they were five years old (Hamadani et al. 2013). To measure 18-month vocabulary, the authors developed a 60-word vocabulary checklist in the local language, based on the MacArthur-Bates Communicative Development Inventories (Hamadani et al. 2013). The association of the vocabulary score at 18 months with IQ at five years was of similar magnitude (0.39) as the association of 18-month BSID Mental Development Index (MDI) with five-year IQ in the same sample (0.37), and was also comparable to the association between developmental scores at age 18 months and five-year IQ in the United States (0.34; see Figure 3.1). In another study in Ecuador, scores on the Test de Vocabulario en Imagenes Peabody (TVIP), the Spanish version of the Peabody Picture Vocabulary Test (PPVT), at age five years were associated with math and language scores three years later, with a 1 standard deviation increase in TVIP scores associated with a 0.32 standard deviation increase in math and language scores (Araujo et al. 2015).

Motor skills

Large (or gross) motor development refers to the acquisition of movements that promote an individual's mobility (e.g., scooting, walking). While the age and sequence of motor milestone attainment may vary both within and across samples of children, nearly all healthy children will eventually acquire the capacity to walk, as well as develop more advanced behaviors like running and jumping. Advancement in motor skills was once thought to be determined by brain and neuromuscular maturation alone (Gesell 1946), but more recent research indicates that other factors–such as physical growth, caregiving practices (e.g., swaddling or carrying), and opportunities to practice emerging skills–also contribute to motor progression (Adolph, Vereijken, and Denny 1998; Adolph, Vereijken, and Shrout 2003; Kariger et al. 2005; Kuklina et al. 2004).

For infants and young children, large motor skills include learning to walk and run, and for preschool-aged children, large motor skills include walking on a line, controlling movements in games, and jumping. The timing of most large motor skills is generally not indicative of future cognitive development (see Hamadani et al. [2013] for an exception), although a failure to demonstrate these skills may indicate the presence of a developmental delay. For example, a child who does not walk at age two may have a developmental disorder that needs to be addressed, and tests of gross motor skills are created to identify children whose development is far behind expectations. Fine motor skills, utilized for tasks such as drawing and writing letters, involve hand-eye coordination and muscle control. They include such abilities as picking up objects and holding eating utensils. For preschool-aged children, fine motor skills include the ability to hold a pencil, write, and draw. The acquisition of fine motor skills is significant because through them children gain a new way of exploring the environment, and, thus, fine motor skills contribute to developmental achievements (Bushnell and Boudreau 1993). Difficulties in motor skills can indicate the presence of neurological or perceptual problems.

Predictive validity of early motor assessments

Early gross motor skills support cognitive development by allowing children to explore and interact with their environment, while early fine motor skills (e.g., holding a pencil, using scissors) are necessary for academic activities such as writing, drawing, and other creative projects (Halle et al. 2012). Executive function and visuospatial skills are also necessary for performing many early motor assessments, and these skills, along with motor coordination, may contribute to academic achievement through multiple pathways (Cameron et al. 2016).

Early motor scores are weaker predictors of later cognition and school achievement, as compared to language, cognitive, and pre-academic scores, in both high-income countries (Halle et al. 2012) and low- and middle-income ones (Hamadani et al. 2013; Richter et al. 2016). In the age range 3–4 years, some fine motor assessments include items that also tap cognitive skills, such as visuospatial ability (e.g., copying shapes) and memory (e.g., drawing a person from memory), and pre-academic knowledge and skills, such as writing letters. These types of assessments have been found to predict some aspects of academic achievement in kindergarten (Cameron et al. 2012).

Social and emotional development

Social and emotional development has implications for many domains of children's development (Saarni et al. 1998). In the first two years of life, much of children's social and emotional development centers on relationships with caregivers. During these years, children learn whether they will be responded to by others and how much they can trust those around them. Learning to explore is a fundamental task of infants and toddlers, and they are more confident in their explorations when they are confident that their caregivers will be available when they return from their explorations. In the first two years, children also acquire early strategies for dealing with their negative feelings. Warm, responsive relationships with caregivers are essential for teaching children to trust, and for helping them learn to deal effectively with frustration, fear, and other negative emotions (Thompson and Raikes 2006). Healthy infants and toddlers will show preferential attachments to caregivers, are eager to explore novel objects and spaces, and enjoy initiating and responding to social interactions.

In the preschool years, social and emotional development expands to include children's social competence (how well they get along with others, including teachers and peers), behavior management (how well they follow directions and cooperate with requests), social perception (how well they identify thoughts and feelings in themselves and others), and self-regulatory abilities (whether they demonstrate emotional and behavioral control, especially in stressful situations). Thus, social-emotional skills are related to, and overlap with, self-regulatory aspects of executive function. While the "big five" components of personality development have been used to explain adult social-emotional development, these constructs do not apply in full to young children. Instead, the concepts of temperament and effortful control are more relevant to young children, as these constructs may then pave the way for development of personality characteristics that become more stable in adulthood (Duckworth et al. 2007; Ahadi and Rothbart 1994).

Social and emotional development has recently received attention in high-income countries, such as the United States and Canada (Darling 2016). Children's self-regulatory and social-emotional skills are included in many measures (see Denham, Ji, and Hamre [2010] for a compendium of measures developed for use within the United States, including both parent or teacher report and observational scales). But fewer of these measures have been used in lower- and middle-income countries to date. It is also recommended that several elements of children's social-emotional development be included to ensure accurate measurement (Denham, Bassett, and Zinsser 2012).

Children who are not able to discern the thoughts and feelings of others are more likely to behave aggressively and experience peer rejection (Denham et al. 2003), and children with both "internalizing" behavior problems, characterized by depressed, withdrawn behavior, and "externalizing" behavior problems, shown by aggressive, angry behavior, are more likely to have difficulty in school (Rimm-Kaufman, Pianta, and Cox 2000).

Indices of children's behavior problems have often been used in studies in low- and middle-income countries (e.g., the Strengths and Difficulties Questionnaire). It is both quick and cost-effective to ask parents to respond to questionnaires regarding their children's behavior problems, keeping in mind that reports of behavioral and social-emotional problems are likely to be influenced by cultural norms. The prevalence of both internalizing and externalizing behavior problems is quite low in most contexts. Measures of behavior problems alone, however, generally yield few insights into children's social and emotional well-being, although these measures can be useful in cases of extreme psychological distress results (Atwine, Cantor-Graae, and Bajunirwe 2005; Mulatu 1995). On the other hand, the absence of behavior problems cannot be taken as an indication of social and emotional well-being. When possible, it is important to use measures that index children's social competencies (strengths), as well as their problematic behavior (difficulties).

Predictive validity of early social and emotional assessments

As children develop the ability to regulate their social interactions and emotional reactions, common behaviors are temper tantrums, fights over toys, high levels of activity, and excessive shyness. Some children who show these behaviors will continue to have serious problems throughout childhood, but the majority of children overcome these difficulties (Campbell 2006). Nevertheless, low to moderate continuity in social-emotional function has been shown from infancy throughout childhood, with a meta-analysis of 70 studies finding a mean correlation of 0.3 (La Paro and Pianta 2000; Briggs-Gowan and Carter 2008; Jaffari-Bimmel et al. 2006). Social-emotional skills are critical for children's successful behavior in school (Duncan et al. 2007; Halle et al. 2012).

In high-income countries, social-emotional skills at preschool age are the best predictors of later social-emotional and behavioral function, compared to measures of language, cognition, and pre-academic knowledge and skills (Halle et al. 2012). In the only report of such associations in low- and middle-income countries, social-emotional measures at preschool age did not predict age of school entry or grade repetition in South Africa (Richter et al. 2016). However, these outcomes may be less sensitive than direct measures of social-emotional and behavioral function in later childhood, which were not measured in this study.

It is critical to note that little has been published on social-emotional development and its role in children's development and school achievement in low- and middle-income countries. Few researchers have studied how different cultures conceptualize social and emotional behavior among children in the early school years, but such behavior is likely to be highly sensitive to cultural expectations. Cultural norms and values strongly influence what is considered an appropriate way to express emotions and interact in social relationships. For that reason, special attention should be paid to understanding cultural priorities for children's social and emotional skills, and ensuring proper translation of terms and concepts before using the measures. Attention must also be given to the method used for collecting information about children's behavior. While teacher reports are often valid in high-income countries, they may be less accurate in low- and middle-income countries if class sizes are large and teachers do not have the opportunity to closely observe the behavior of each child.

Pre- and early-academic skills

Five important dimensions of school readiness:

1. Physical well-being and motor development
2. Cognitive development and general knowledge
3. Language development
4. Social-emotional development
5. Approaches to learning

These have been specified as the most important by the U.S. National Education Goals

As children move from early to middle childhood, they become more capable of many skills essential for school success, including sitting still, understanding teachers' instructions, and retaining information from group settings. In the late preschool and early school years, children are increasingly expected to participate in group learning environments and to master early academic skills, such as writing, reading, and numeracy. This shift reflects both the demands of school systems and also the emergence of children's abilities in memory, language, and reasoning, which, in turn, facilitate skills in reading and math. The concept of "school readiness," which has become a commonly used term in describing children's development at the start of formal schooling and its implications for later academic achievement, refers to both early academic skills, such as literacy and numeracy, and children's abilities to regulate their attention and behavior (e.g., Blair 2002; Blair and Raver 2015).

Critical early academic skills include letter knowledge, number knowledge, counting, and listening comprehension. These skills are covered in the Early Grade Reading Assessment (EGRA) and Early Grade Math Assessment (EGMA), which have been used extensively in low- and middle-income countries. Tests of academic achievement for children of this age may also include knowledge of science, such as how plants grow; and general knowledge about the world, such as the name of the country in which they live. As children near third grade, the expectation in many countries is that children can fluently read paragraphs and report on the meaning of the passage. Children are also expected to have well-developed math skills by this time and to demonstrate mastery of double-digit addition and subtraction and early multiplication. Measuring these early academic skills is typically quite straightforward, and many countries have national assessments that can be a good source of items for measuring early academic skills (if it is possible to obtain these assessments). Other potential sources for data on children of this age include the citizen-led assessment organizations (e.g., Uwezo in East Africa and ASER Centre in India), which have conducted fast, large-scale administration of learning tests for children aged five years to 15 years in many low-income countries.

In high-income countries, school-entry academic skills are stronger predictors of later math and reading achievement than any other dimension of school readiness (Duncan et al. 2007; Halle et al. 2012). In general, the strongest predictors are within domains (Duncan et al. 2007; Halle et al. 2012; La Paro and Pianta 2000). For example, entry-level math skills provide the best prediction of subsequent math skills, and entry-level reading skills provide the best prediction of subsequent reading skills. Several studies have examined the relation between types of interventions and children's outcomes. These studies point in the direction of preschool-based settings to promote school achievement (Rao et al. 2017). Fewer studies to date have examined the relation between specific pre-academic skills measured in the preschool years and children's later learning in primary school, which would, in turn, help differentiate the predictive power of pre-literacy, math, executive function, and social-emotional skills.

Implications for impact evaluations

A key question in any evaluation of an early childhood program or intervention is whether the children who participate benefit in the long term. However, most program evaluations and research studies are only able to evaluate short-term developmental outcomes. If short-term benefits are found, does this mean those benefits will be sustained in the future? Conversely, if short-term benefits are not found, does this mean that no benefits will be detected at any time in the future? One might look to predictive validity studies to shed light on these questions. However, the answers are not straightforward.

It is possible that developmental effects of an early childhood intervention might not be detected at an early age, but may be found in a follow-up study. A few nutrition studies have shown this pattern. For example, in a randomized controlled trial, infants who received formula containing certain fatty acids did not differ in vocabulary scores or Bayley Scales of Infant Development scores at age 18 months, compared to infants who received formula without these fatty acids. However, in a follow-up study, they showed higher vocabulary and IQ scores at ages 5–6 years (Colombo et al. 2013). In another study, a group of children who experienced thiamine deficiency in infancy did not show neurological symptoms at the time of deficiency, but showed language impairment at age five to seven years (Fattal, Friedmann, and Fattal-Valevski 2011). As we have seen in the previous sections, developmental scores tend to be more stable after age two years, thus assessments before this age may not be accurate or sensitive enough to detect effects.

It is also possible that observed benefits of an intervention decrease over time. Many early childhood interventions have shown this pattern of diminishing effects in follow-up studies (Hart and Risley 2003; Lazer et al. 1982). For example, in the Infant Health and Development Program conducted on infants in the U.S., the intervention resulted in a 14-point IQ advantage at age three years compared to a control group, but diminished to four points at age five years and subsequently remained stable to age 18 years (McCormick et al. 2006). In other trials, follow-up studies have shown no lasting differences between intervention and control groups, despite immediate positive effects (Hart and Risley 2003; Lazar and Darlington 1982; Hauglann et al. 2015; Lee et al. 1990). Diminishing effects may be due to the influence of environmental factors on children's developmental trajectories after the intervention period. Particularly in a high-risk environment, this may result in diminishing effects of an intervention over time (Figure 3.2).

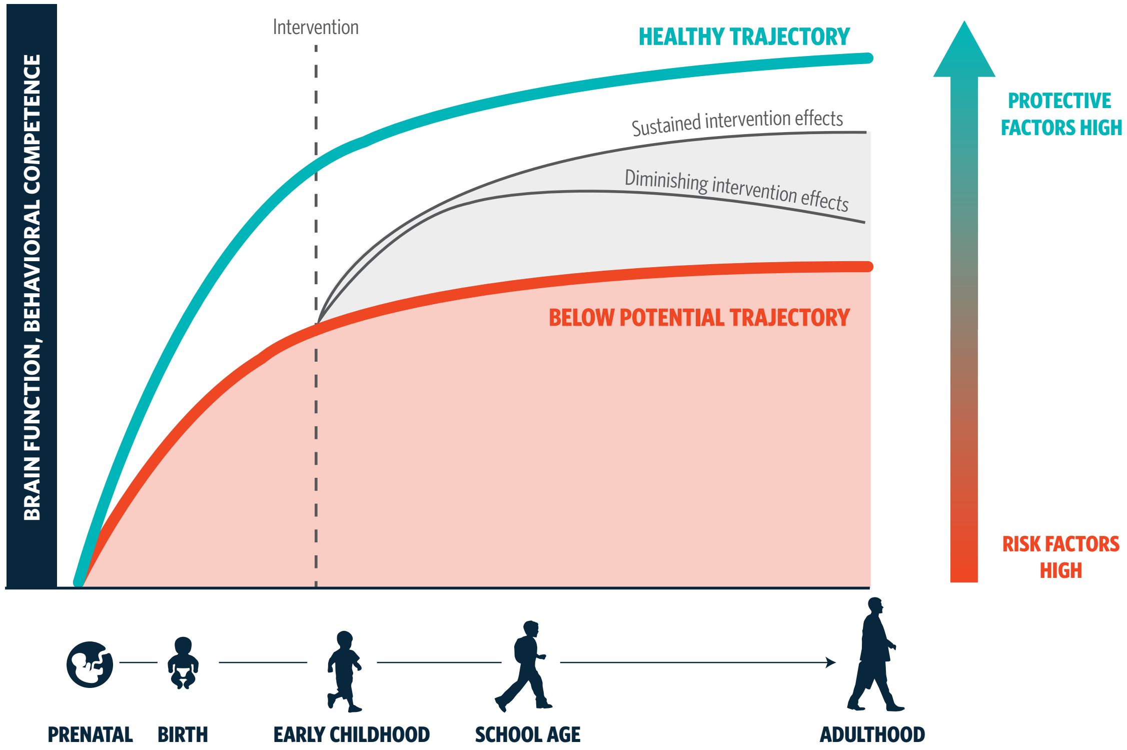
This phenomenon has been called the “sustaining environments” perspective (Bailey et al. 2017). In this perspective, intervention effects are likely to fade out without ongoing, post-program educational supports. Bailey and colleagues (2017) reviewed persistence and fadeout in the impacts of childhood interventions and proposed two additional perspectives, in addition to the sustaining environments perspective: the skill-building perspective and foot-in-the-door perspective. In the skill-building perspective, effective, sustained intervention requires targeting skills that are malleable and fundamental and that would not have developed in the absence of the intervention. In the foot-in-the-door perspective, interventions will have long-lasting effects if they equip a child with the essential skills at the right time to avoid imminent risks (e.g., grade failure, teen pregnancy) or to take advantage of timely opportunities (e.g., exam preparation), which will have later consequences.

Drawing implications from longitudinal follow-up studies in high-income countries for low- and middle-income contexts is challenging, because children in such countries face different types of environmental risk factors. Unfortunately, such data from low- and middle-income countries is scarce. While a number of studies have enrolled children in low- or middle-income countries in early life and tracked them throughout childhood, few assessed early child development. In addition, few randomized trials of early childhood programs have conducted long-term follow-up studies. However, a recent analysis of longitudinal data from the Young Lives studies in Ethiopia and Peru demonstrates both the continuity and plasticity of cognitive trajectories from age five to 15 years (Attanasio et al. 2017). Children from low-income families who were in the 10th percentile of cognitive scores at age five years persisted in having the lowest cognitive scores at all subsequent time points through age 15 years. Conversely, children from high-income families who were in the 90th percentile of cognitive scores at age five years persisted in having the highest cognitive scores through age 15. Children with median cognitive scores at age five years diverged over time depending on their family income level, with the scores of children from low-income families declining and the scores of those from high-income families increasing. However, the scores of these median groups remained in between the two groups with the highest and lowest cognitive scores at age five years. These results underscore the "sustaining environments" perspective and the importance of ongoing environmental support throughout childhood.

Another question related to predictive validity that might be of interest in an impact evaluation is: What measures should be gathered at baseline that would most strongly predict outcome scores? This is an important question because accounting for maximum variance in outcome scores will increase statistical power to detect the effects of the intervention. The best option is to use the same test at baseline and endline. It might not be possible to do this if the age range of the children at baseline is very different from the age range at endline. In this case, use a test at baseline that assesses the same domain(s) as the endline test, since predictive validity within domains is stronger than across domains.

FIGURE 3.2 **Developmental Trajectories and the Impact of Intervention**

The teal line shows a healthy trajectory of brain and behavioral development from the prenatal period throughout adulthood. The red line shows a developmental trajectory that is below a child's potential. The grey lines show potential changes in trajectories following an intervention. Risk and protective factors may influence trajectories both before and after the intervention.


BRAIN FUNCTION, BEHAVIORAL COMPETENCE
Intervention
HEALTHY TRAJECTORY
Sustained intervention effects
Diminishing intervention effects
BELOW POTENTIAL TRAJECTORY
PROTECTIVE FACTORS HIGH
RISK FACTORS HIGH
PRENATAL
BIRTH
EARLY CHILDHOOD
SCHOOL AGE
ADULTHOOD

Source: Adapted from Figure 2 in Walker et al. 2011.

4 Assessment Tools and Their Uses

KEY MESSAGES:

- The first step in selecting measures is to clarify the purpose of the assessment. Three broad purposes of assessments are population monitoring, program evaluation, and hypothesis-driven or exploratory research.
- For program evaluations, brief assessment tools with just five or six items per age category may not show sufficient variance in scores in typically developing children for capturing treatment-related effects.
- The accuracy of conclusions about child development depends on how well the data reflect the underlying construct or ability we intend to measure.
- Assessments can be physiological or behavioral and can be conducted via direct assessment, parent or teacher report, naturalistic or structured observation, or direct measures of brain structure or function.

DECIDING "WHY" TO ASSESS CHILDREN'S DEVELOPMENT, "WHAT" TO ASSESS, AND "HOW" TO assess their development outcomes are crucial steps in any project evaluating early childhood development. The answer to the "why" question, or the purpose of the assessment, will determine which domains of early childhood development—or underlying constructs or abilities—to measure. Then, a method must be selected or developed to assess this underlying ability, which must fit given resource constraints (e.g., training and background of data collectors, financial resources available for test purchase, time available for testing, and testing location). The next step is implementation of the method in the field, resulting in data such as item scores, subscale scores, composite scores, or an indicator of risk for developmental delay (Figure 4.1).

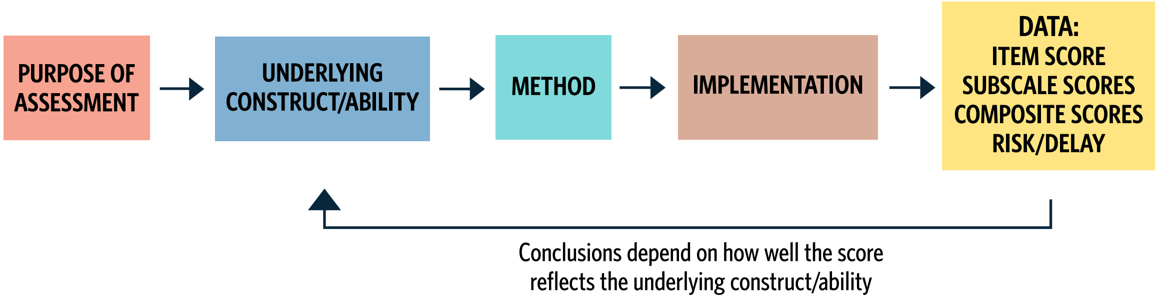
No matter what the purpose for evaluating early childhood development, the accuracy of our conclusions depends on how well the data reflect the underlying construct or ability we intended to measure. Therefore, it is crucial (1) that the method of assessment accurately reflects the underlying construct; and (2) that the method is implemented with high quality and fidelity (Figure 4.1). This chapter will outline the major issues involved with selecting assessment instruments, including the purpose, types, and methods of assessments. Chapter 5 will discuss how to ensure that the method reflects the underlying construct in the local context and that the method is implemented with fidelity.

Purpose of measurement

The first step in selecting measures is to clarify the purpose for the assessment. Assessments of child development can be conducted for various reasons. Our framework, presented in Figure 1.1, includes three broad purposes: (1) population monitoring, (2) program evaluation, and (3) hypothesis-driven or exploratory research. Population monitoring requires less detailed assessment, since the goal is a high-level, broad view of child development at a population level, typically to detect broad trends in differences between groups, with emphasis on description, not explaining or predicting how child development is affected by any one condition or exposure. Program evaluation requires a more detailed assessment battery, measuring a range of skills aligned with the goals of the program to detect changes in child development due to an intervention or program. For hypothesis-driven or exploratory research, even more detailed assessment and innovative methods are appropriate, including physiological measures to explore the mechanisms of behavioral effects to generate new hypotheses or theories of child development. If a tool is to be used for two different purposes, it needs to meet the minimum requirements for both.

A fourth purpose is evaluation of an individual child in a clinical or educational context for referral to diagnostic evaluation or to special programs. We have not focused on this purpose because ethical considerations require that evaluating and treating individual children is conducted by clinicians and educational psychologists with specialized training. However, many of the principles for test adaptation and evaluation presented here are also relevant for clinical diagnostic tools.

FIGURE 4.1 **The Importance of Validity and Quality in Data Collection**



PURPOSE OF ASSESSMENT
UNDERLYING CONSTRUCT/ABILITY
METHOD
IMPLEMENTATION
DATA:
ITEM SCORE
SUBSCALE SCORES
COMPOSITE SCORES
RISK/DELAY
Conclusions depend on how well the score reflects the underlying construct/ability

PURPOSE 1: Population monitoring

The purpose of many child development assessments is to measure how individual children progress following a health, caregiving, or educational intervention, or to relate performance on one test with another. Population-level measures, on the other hand, can be used to draw conclusions about the overall state of children's well-being or used to compare a group of children (such as within a classroom, a school, a region or country) to other groups of children, within and across countries.

Many countries are increasingly relying on population-based measures to inform national and community-level policies on child development, including childcare, preschool access, and health care. If used to generate comparisons between groups either within or across countries, population measures would ideally be invariant across groups, meaning that the constructs and items would be equally applicable regardless of context. Such measures, therefore, can be used to inform system-level decision-making about how best to support young children's development and learning, as well as appropriate planning of interventions. While many population-based tools are commonly administered to representative samples of children, it is important to note that "census-based" approaches have been developed as well to capture the development of every child of certain ages (Brinkman et al. 2012).

Population-based measures differ from measures designed for research or program evaluation because they are designed for use at scale, with an emphasis on feasible, cost-effective measurement, and therefore may be broader in scope than measures used for measuring program impacts or testing specific research hypotheses. One example of such a measure is the Early Development Instrument (EDI), which seeks to assess children's school readiness (Sidebar 4.1). Although some developers maintain that population-based measures can be used for program evaluations, potential users are encouraged to contact test developers to clarify. The population-based monitoring approach encourages a focus on the context of children's abilities and community-level factors and reduces the risk that a test is used to categorize children and in some cases, even stigmatize them, although no test can prevent users from inappropriately using the results. The use of a tool as a measure of "population health" also assumes the importance of community strengths and weak-

SIDEBAR 4.1 EDI: Example of a Population-Based Assessment

The Early Development Instrument (EDI) is a teacher rating of children's readiness for first grade, assessed between ages 3.5–6 years. It differs from many other tests developed to measure children's maturational or experiential readiness for school (Janus and Offord 2007; Janus and Reid-Westoby 2016). In a review of seven of these instruments (Janus and Offord 2007), the authors concluded that although some are reasonably predictive of school success, they must be administered by a professional, they are not cost-effective, nor do they measure all relevant domains (e.g., social-emotional development was missing). To fill this gap, Janus and Offord developed the Early Development Instrument as a teacher-rating scale that can help assess children's school readiness at a much lower cost.

The EDI is a set of questions (initially 103, but there are shorter versions) that a teacher can use for rating an individual child (Janus and Offord 2007; Brinkman et al. 2007). The questions cover five domains: physical health and well-being, emotional maturity, social competence, language and cognitive development, and communication skills and general knowledge. From the instrument, one can quantify the percentage of children who are vulnerable in each of the five dimensions.

The ratings that result from the EDI were found to be associated with other measures of cognitive and social-emotional development (teacher ratings on other measures, direct tests, and parent ratings) and thus had reasonable construct validity in both Canada and Australia (Janus and Offord 2007; Brinkman et al. 2007). These associations were compared using both a continuous scale and a dichotomous measure of vulnerability. Associations with other teacher ratings and with tests were reasonably high, and higher than with parent ratings.

nesses and assesses the value of community-oriented interventions (e.g., providing local libraries) that is not emphasized by individual measurements. For example, each school could be considered a category, and schools could be compared on the percent of children at risk. The intervention, such as financial resources, could be distributed to schools on the basis of the school-level variables (i.e., the percent of children at risk).

It is also possible that not all children are assessed, but that a group could be sampled from the larger population. There is no scientifically accepted way of using population-level data to make decisions about whether an individual child is on or off track using any of these instruments, and the instruments are not standardized in any country.

A review of the benefits of using population-level measures is available (Mustard and Young 2007). Despite the benefits, a number of drawbacks exist. First and foremost, if used to compare between groups of children, it is extremely difficult to develop measures of child development that are invariant across groups, because the nature of early development is such that context and culture influence how and when children develop competencies. Measurement invariance is established by comparing the applicability of constructs and items across a range of different groups within a population, and requires considerable investment and modification of tools in response to results. At present, there is no accepted, valid, and globally comparable indicator of early child development. Table 4.1 summarizes the strengths and weaknesses of population-level assessments.

At present, there is no accepted, valid, and globally comparable indicator of early child development.

Second, population-level data is meant to be collected on representative samples, which may not be possible to obtain in all studies. Indicators, or singular data points that summarize conditions or outcomes across representative populations, can help raise awareness of early childhood development among governments and policymakers of the importance of investing in development during the first five years of life. To provide a useful overview for policymakers and governments, indicators must come from representative data on an entire population, and as a result, household surveys have emerged as a reasonable means for gathering representative data on child development. However, household surveys are limited in their scope and depth, presenting notable challenges on gathering reliable, valid early childhood data.

In response to the emphasis on child development that has arisen through the United Nations Sustainable Development Goals (SDGs), the World Bank (Kim 2016), UNICEF (United Nations News Centre 2016), and the World Health Organization (WHO 2013) have all highlighted the need for early childhood services that promote the development of skills and capacities that will enable children to thrive, and the concomitant need for suitable tools to measure progress towards such goals. Beyond the MICS ECDI (see Sidebar 4.2), several efforts are underway to create and validate universal, population-level measures that can be used to track the developmental status of children 0–72 months of age. These include the WHO's 0-3 Indi-

SIDEBAR 4.2 MICS Surveys: Efforts to Develop a Global Indicator

In an attempt to generate globally comparable data on early childhood, UNICEF developed a 10-item household survey module that asks parents to rate their children's behavior in five domains of development. This new measure has been included in recent versions of the Multiple Indicator Cluster Surveys (MICS4, MICS5, MICS6). Development of this Early Child Development Index (subsequently named the "ECDI," or early childhood development index) began in 2005, with the purpose of identifying a set of simple, practical, holistic, and commonly accepted indicators that could be used to track children's development globally. After reviewing several large-scale surveys of child development conducted both at the national and international level, the MICS team selected a set of items for field-testing, drawing from the EDI, a teacher report instrument developed in Canada (see Sidebar 4.1). The team developed a 48-item survey, reducing it to 18 items for pilot testing, and finally, to 10 items, for use as the MICS ECDI module. The final list of 10 items includes three items on numeracy or literacy; three on social-emotional development; two on approaches to learning; and two on physical development. The MICS ECDI has now been administered to about 99,000 children in more than 50 countries (www.unicef.org/statistics), but evidence of its cross-country validity is limited at present.

Several considerations limit the use of the MICS ECDI as an outcome measure in either program interventions or research studies. Recent analyses examined MICS ECDI results in several countries (McCoy et al. 2016). MICS ECDI items showed predicted associations with family wealth in most, but not all, countries, and some items showed more alignment with national standards than others. Psychometric properties were stronger in some areas (math and literacy) than in others (e.g., social-emotional development).

The reliance on one set of items for children over a relatively large developmental span of two years, coupled with the desire to reduce the total number of items down to 10, creates tension within the MICS scale to fully capture children's development across several domains while remaining feasible for collection at scale. One study concluded that only five of the 10 items were age-appropriate for 3–4 year olds, which was the target age group for the MICS surveys (McCoy et al. 2016). Reliance on a global set of items does not allow countries to align the items with local expectations for children's development, thus potentially decreasing the relevance to policies on preschool curricula or teacher training. Finally, the MICS ECDI is only available for children aged up to four years, 11 months, which is not at the start of formal schooling in all countries.

In sum, the MICS ECDI was designed for population-level monitoring to produce a short, household survey-based instrument that can reveal broad trends in child development. Its utility for this purpose has been debated; its use for program evaluations and research should be even more closely scrutinized as it is likely too short and not culturally responsive enough to generate a meaningful outcome measure of child development.

cators; the Measuring Early Learning Quality and Outcomes (MELQO) modules developed by a consortium led by Brookings Institution, UNESCO, UNICEF, and the World Bank; projects led by the Global Child Development Group; the Caregiver-Reported Early Development Index (CREDI) (CREDI 2016; McCoy et al. 2017) and Save the Children's International Development and Early Learning Assessment (IDELA) (Pisani, Borisova, and Dowd 2015).

In contrast to many copyrighted tools that require payment to the publisher for every child tested, the intended purpose of these efforts is to provide freely available tools that can be used throughout the world to gather data on child development using similar items. The results of these surveys aim to serve dual purposes of generating globally comparable data and yielding measures sensitive enough to detect impacts in program evaluations or research. The proposed approaches include caregiver report and direct measurement. Most of these tools are available now, and may be more suitable for program evaluation and research than the MICS ECDI.

Few studies in low- and middle-income countries have examined the concurrent validity—or the extent to which results of a measurement tool correlate with results of a previously established measurement tool—of relatively short developmental assessments administered in the context of a household survey, as compared to a more comprehensive assessment conducted by psychologists at a clinic, which may be considered a gold standard measure. A recent large study in Bogota evaluated the concurrent validity of five tests in this way, using the Bayley Scales of Infant Development, Third Edition (BSID-III) as the gold standard measure (Rubio-Codina et al. 2016). The five tests administered in a household survey were three multi-domain screening tests (Ages & Stages Questionnaires [ASQ-3], Denver Developmental Screening Test [Denver-II], and Battelle Developmental Inventory - Second Edition screener [BDI-2]) and two single-domain tests

(MacArthur-Bates Communicative Development Inventories short forms and WHO Motor Milestones). The Bogota study found differences in validity, depending on child age and developmental domain. The Denver-II was the most feasible and valid multi-dimensional test, while the ASQ-3 performed poorly for children younger than 31 months. In general, concurrent validity of the multi-dimensional tests' cognitive, language, and fine motor scales with the corresponding BSID-III scale was low for children who were aged under 19 months. However, it increased with age, becoming moderate-to-high for children older than 30 months. By domain, gross motor development had the highest concurrence for those younger than 19 months, and concurrence for language was highest for children who were older than 19 months.

TABLE 4.1 **Pros and Cons of Population Monitoring Tools**

➕ PROS	➖ CONS
Tracking with representative samples. Governments can use population-based measures to track child development and the overall state of children's well-being within representative samples or using a census approach. This can be useful for identifying gaps and successes, and can help organizations and governments make decisions on allocation of funds and programs.	**Lack of sensitivity.** Tools may not be sensitive enough for use with impact evaluations, as population-level measures do not assess children very comprehensively. **Difficulty in getting cross-cultural comparability.** It is highly challenging to obtain equivalent scores across cultures and contexts without careful piloting and adaptation, ideally with the test creator. Systematic bias or error in test scores, which is discussed further in Chapter 6, can occur when applying a test in a culture other than the one in which it was developed. For classroom-based tools, teachers may show systematic biases in their ratings, which can become an issue when measures are compared across cultures or across children varying in age level. Cultural categories such as income, race, or gender may influence ratings, even when respondents are less sensitive to bias.
Nimble data collection. Data collection may take less time and cost less money than surveys done for other assessment purposes, because population estimates sometimes rely on less in-depth questionnaires.	
Potential for research. Some tools may be in-depth enough for researchers and evaluators to use as measures of child development outcomes.	**Sampling challenges.** Sampling may be difficult, but it depends on the goal of the study. If the population of interest includes all children, and not all children are in the group measured (e.g., children attending preschool), it is more difficult to obtain a truly random sample because non-attenders are not included.

PURPOSE 2: Program evaluation

Program evaluations seek to assess the impact of select interventions on child development by measuring a range of skills aligned with the goals of those interventions. Sidebar 4.3 lists questions that can help clarify the purpose of the measurement or assessment. Thorough research relevant to these parameters will narrow the range of tests most suitable for use. After clarifying the purpose for testing, the next step would be to determine the type of assessment to use. A range of tools has been developed to conduct such evaluations in both low- and middle-income countries (see Sidebar 4.4) and high-income ones (see Sidebar 4.5). In program evaluations, the purpose for testing should clearly link to objectives or goals that, in turn, will help guide which domains to measure, the types of tests and testing modes to use, and approaches for interpreting and using the test information (Snow and Van Hemel 2008; Behrman, Glewwe, and Miguel 2007). For example, consider a project concerned with examining the impact of an early parent-stimulation program on child development. The general goal of the project would be to determine whether children 6–24 months of age receiving the intervention perform better on developmental tests than children in a control group.

PURPOSE 3: Hypothesis-driven or exploratory research

The purpose of hypothesis-driven or exploratory research is to develop and test new models and theories of child development and the mechanisms that drive developmental change throughout childhood. For

SIDEBAR 4.3 Questions to Clarify Purpose of Measurement

In the context of an impact evaluation, it would be essential to answer the following clarifying questions to select instruments that will best serve the purpose of the assessment.

- **What dimensions of a child's development are expected to be affected by the intervention?** For example, in the case of an early parent-stimulation program, researchers may hypothesize that the major impact of the intervention would be on changes in the interactions between caregivers and children (e.g., increasing adult-child engagement in learning activities), which would subsequently benefit child performance on language, social-emotional, and problem-solving tasks. It is important to consider measuring aspects of development that link to these kinds of immediate outcomes, as well as to longer-term outcomes (e.g., grade completion or achievement scores, and literacy).

- **What dimensions of a child's development are expected to be affected at the target age?** As explained in Chapter 2, various domains of early childhood development progress according to different trajectories, with motor and language skills developing rapidly at earlier ages and executive function developing at later ages. A domain that is developing rapidly at the target age is likely to show more variance in scores and therefore to be more sensitive to intervention effects.

- **What are the mechanisms at work?** Through which (biological or environmental) mechanism(s) is the intervention expected to operate? What is already known about the functional mechanism linking stimulation, for example, with child performance on various aspects of development that could guide the choice of outcomes? Which processes are most influenced by the intervention, and which biological or environmental risk factors present in the population under study need to be considered in planning and evaluating the intervention? How do these factors change with age (e.g., stimulation programs are generally more effective if started when children are very young)?

- **What are key elements of the context to consider in selecting a test?** These may include: urban or rural setting; level of poverty; parent education and literacy; language spoken in the home; risk factors to which children are likely exposed; and access to, and familiarity with, the media required for the assessment (e.g., pencil and paper).

- **At what level will the effect be measured?** Are the evaluators most interested in demonstrating impact or examining patterns at the individual, household, community, or population level?

- **How will the sample be selected?** Given the study design, is it necessary to test all children or will it suffice (and also be more feasible) to measure a sub-sample of the population? What sample size will be needed to provide sufficient statistical power to detect the anticipated effect, or to detect the minimum meaningful effect?

- **What is the plan for analysis?** Are the assessments occurring in a context where norms (i.e., age-related references for the development of skills or abilities) are available? If so, are the norms relevant and appropriate for the population being tested? For example, even if there are "norms" for the Spanish version of the Peabody Language Development assessment, the Test de Vocabulario en Imagenes Peabody (TVIP), they may not be relevant for all Spanish-speaking countries because they were developed using a limited population sample. Will a cutoff score be used to demonstrate "delay"? If so, how will this cutoff point be determined in the population under study?

- **What are the goals of the analysis?** Is there an interest in showing relative improvement in one group over another or in individual improvement in developmental scores or in domains? Will the scores be used to examine developmental differences? Sometimes evaluations consider relative changes in groups—treatment group versus control—and in such cases, it is helpful if the assessment is extensive enough to demonstrate group differences. Brief assessment tools with just five or six items per age category are often used in large-scale surveys and impact evaluations. However, such tools may not show sufficient variance in scores in typically developing children to capture treatment-related effects. Which measures have been shown in the literature to be most sensitive to detecting treatment effects in similar samples of children? Do these change with age?

SIDEBAR 4.4 Tools Developed in Low- and Middle-Income Countries for Program Evaluation

- **Kilifi Developmental Inventory (KDI)** (Abubakar et al. 2007; Abubakar et al. 2008; Abubakar et al. 2008). This tool was developed to assess psychomotor development in a resource-limited setting, drawing motor items from several standard tests, including the Griffiths Mental Development Scale and the Merrill-Palmer Scales. The KDI showed adequate test-retest reliability (> 0.7). Using the 10th percentile as a cutoff, the KDI showed 89 percent sensitivity and 91 percent specificity to detect children with neurodevelopmental impairment in Kenya. The test has been used in studies in Uganda (Nampijja et al. 2012), Malawi (Prado et al. 2016), Ghana (Prado et al. 2016), and South Africa.

- **Malawi Developmental Assessment Tool (MDAT)**. This tool was developed in Malawi by combining items from the Denver Developmental Screening Test (Frankenburg et al. 1992; Frankenburg 1985), the Griffiths test (Griffiths 1984), and some new items drawn from culturally sanctioned behaviors (Gladstone et al. 2010; Gladstone et al. 2008). The tests assess fine and gross motor, language, and personal-social development. After adaptation and pilot testing, more than 94 percent of items showed high reliability (kappa > 0.4 for inter-observer immediate, delayed, and intra-observer reliability) (Gladstone et al. 2010). Using the screening criterion defined as whether the child failed two items or more in any one domain at the chronological age at which 90 percent of the normal reference population would be expected to pass, the MDAT demonstrated high sensitivity (97 percent) and specificity (82 percent) to detect children with neurodevelopmental impairment in Malawi (Gladstone et al. 2010).

- **Developmental Milestones Checklist (DMC)**. This tool was assembled in Kenya by adapting items selected mainly from the Griffiths Mental Development Scale and Vineland Adaptive Behavior Scale (Abubakar et al. 2010). The first version of the checklist was further adapted and extended in Burkina Faso, creating the DMC-II (Prado et al. 2014), which assesses motor, language, and personal-social development. The DMC-II scores demonstrated internal reliability (Cronbach's alpha), inter-rater reliability, and test-retest reliability (ICC) of greater than 0.75, and showed expected correlations with age, stunting, wasting, and underweight in Burkina Faso (Prado et al. 2014). The tool has also been used in Bangladesh (Matias et al. 2017) and India (Larson et al. 2017).

SIDEBAR 4.5 Tools Developed in High-Income Countries for Program Evaluation

- **Bayley Scales of Infant Development (BSID)**. This is a widely used tool to assess children ages one month to 42 months. The BSID has three versions (I, II, and III). The BSID-III was developed in 2006 to replace the BSID-II, which was developed in 1995 to replace the 1969 BSID-I. All versions of the assessment tool provide scores for both the Mental Development Index and the Motor Development Index. The newest version includes language, cognitive, social-emotional, motor, and adaptive behavior (caregiver report) subscales that can be scored separately, so that domain-specific assessments can be made. Although this tool is well-validated in the United States and has been used in many different countries, several studies have found bias when applying the tool in different contexts and cultures (Ogunnaike and Houser 2002; Vierhaus et al. 2011; Hagie, Gallipo, and Svien 2003).

- **MacArthur-Bates Communicative Development Inventories (CDI)** (Fenson et al. 2006). These tools are parent report forms for assessing language and communication skills in infants and young children. The tools have been adapted to many languages (see https://mb-cdi.stanford.edu/adaptations.html). In Bangladesh, scores on a 60-word vocabulary checklist at age 18 months, developed based on the CDI, showed correlations of 0.30-0.37 with verbal IQ, performance IQ and full scale IQ at age five years (Hamadani et al. 2010). The order in which children learn specific words and grammatical structures varies by language, therefore these tools cannot be directly translated but must be adapted to different languages and contexts.

- **Kaufman Assessment Battery for Children (KABC)**. This tool is an intelligence test of problem-solving ability which is normed for children's performance on three subscales: achievement, simultaneous processing (ability to solve problems by integrating diverse pieces of information simultaneously), and sequential processing (ability to solve problems by ordering items or placing them in sequence). The KABC has been used in a handful of studies evaluating the effects of intervention programs and has shown sensitivity to changes in nutritional status, including iron and iodine. The assessment has also shown sensitivity to exposure to malaria. It has been used in several different languages, including French (spoken in Benin), Laotian, Wolof (spoken in Senegal), and Kikongo (spoken in Democratic Republic of Congo).

- **Wechsler Preschool and Primary Scale of Intelligence (WPPSI)**. This tool is an extension of the Wechsler Intelligence Scale for Children. Both are designed to be measures of intelligence, not achievement. The scale addresses two broad factors: performance and verbal skills. Performance items do not require the child to talk to the experimenter and so may be less sensitive to cultural biases and easier to use across diverse linguistic contexts. The WPPSI has been used widely around the world, including in Brazil, China, Iran, Mexico, Pakistan, and Venezuela.

this type of study, extensive and sensitive batteries of developmental measures are appropriate, possibly including neuroimaging and other advanced methodologies. An advantage of these methods is that they allow us to look "under the hood" or into the "black box" to examine the biological mechanisms of individual and group differences in behavior. Such research has the potential to shed light on critical questions in developmental science regarding the effects of early experience on brain development and to clarify which underlying neurobiological processes are disrupted by, or resilient to, early exposures. For example, some questions that might be answered are: Which risk factors (e.g., illness, deficiencies in specific nutrients, lack of responsive care) are more or less influential at various time periods during development? Which neurobiological processes or brain systems are vulnerable to which risk factors, and which are generally resilient? Which neurobiological processes or brain systems can recover after an early insult, and which show lasting effects of early damage despite later treatment?

Advancing scientific knowledge in these areas is critical to inform early childhood development policy and practice in low- and middle-income countries. Additional research clarifying developmental mechanisms is needed among children in these countries. Although a large proportion of children in the world live in low- or middle-income country contexts, most research in developmental science has been conducted in samples from high-income country populations. In some ways, children's environments and developmental trajectories are consistent across contexts; in other ways, they are vastly different. More research is needed in low- and middle-income countries, where children generally face a higher burden of risk factors for poor early childhood development.

Clarifying underlying mechanisms can inform the design of more effective and efficient interventions. For example, a randomized trial of a family-based intervention that combined parent training and attention training exercises among children in the United States attending Head Start, a national program to improve school readiness among low-income children under the age of five, showed positive effects on both child cognitive skills and changes in event-related potentials (ERPs) during an attention task. These findings suggest that the improvement that children exhibited in cognitive scores after the family-based training was at least partly due to changes in function of the neural structures underlying selective attention (Neville et al. 2013). This leads to the hypothesis that a targeted intervention focusing on selective attention, which may be more efficient and cost-effective than a broad intervention strategy, may have global cognitive benefits. Such hypotheses that arise out of these types of studies can then be tested at a larger scale with a more general population for effectiveness.

Assessments of lower-level cognitive abilities tied to specific brain systems, or measures of neural activity, may be more sensitive to effects of interventions than global behavioral tests. Performance on global behavioral tests often depends on lower-level cognitive abilities and brain systems. For example, performance on an IQ test probably depends in part on the ability to focus and sustain attention; working memory capacity; speed of information processing; reasoning ability; and executive function. The demonstration of any effects of an intervention on an IQ score would not indicate which lower-level ability or combination of abilities might have been specifically affected. Conversely, a lack of an effect on an IQ score does not necessarily mean that all cognitive components are intact, particularly since children may be able to compensate for deficits in one area of ability while carrying out a more global task.

PURPOSE 4: Screening children for further evaluation and diagnosis

In this Toolkit, we have focused on the three purposes described above because diagnosing and treating individual children ethically requires specialized clinical training and certification. However, many of the principles for adapting and evaluating ECD assessments presented here are also relevant to clinicians in lower- and middle-income countries who would like to adapt a standard assessment to a new language and context. In addition, it is a common practice for community health workers to screen children using a screening tool, then refer the children who screen positive for risk of developmental delay to a clinician for further evaluation and diagnosis. Sidebar 4.6 lists several screening tools developed in low- and middle-income countries, and Sidebar 4.7 lists tools developed in high-income settings.

When using screening tests, it is extremely important to be sensitive to the possibility of stigmatization of children who screen positive, given that a positive result on a screening test does not mean that the child is actually delayed. Community workers, clinicians, and others involved in such programs should take steps to ensure that children who either screen positive or who are diagnosed with a developmental problem do not receive negative treatment. For the purpose of screening, diagnosis, and referral, the test's usefulness will be determined foremost by how sensitive and specific it is.

SIDEBAR 4.6 Screening Tools Developed in Low- and Middle-Income Countries

- **The Guide for Monitoring Child Development** (Ertem et al. 2008) is a parent report assessment originally developed in Turkey, which has also been used in Argentina, India, and South Africa. It provides a method for developmental monitoring and early detection of developmental difficulties in children in low- and middle-income countries. The questions are designed to be simple and clear and they pertain to the child's social, emotional, and cognitive development.

- **The Rapid Neurodevelopmental Assessment** was developed in Bangladesh to provide a comprehensive profile of functions in children aged 0–5 years, for use by a range of professionals who work in the health and rehabilitation sectors (Khan et al. 2010; Khan et al. 2013). The tool screens for impairment in reflexes, motor skills, vision, hearing, speech, cognition, and behavior, as well as for seizures. Reliability was found to be good to excellent in Bangladesh and scores correlated strongly with the BSID-III.

- **The Ten Questions Questionnaire (TQQ)** was designed to be a screening tool for neurological difficulties in children aged 2–9 years (Durkin et al. 1994). The tool screens for risk of epilepsy and for cognitive, motor, vision, and hearing impairments. Sensitivity and specificity were evaluated in Kenya, Bangladesh, and Pakistan and ranged from 65 percent to 100 percent, except sensitivity for hearing impairment (54 percent) and vision impairment (34 percent), which were low in Pakistan. Negative predictive value was very high (>95 percent), indicating that a high percentage of children who screened negative were diagnosed as typically developing (no delay). However, positive predictive value was very low (9–32 percent), indicating that a low percentage of children who screened positive were diagnosed with a developmental delay (Mung'ala-Odera et al. 2004).

- **The Intergrowth-21st Neurodevelopment Assessment (INTER-NDA)** was developed as a population-based screening instrument for early childhood disability. This assessment includes measures of auditory-evoked potentials, cognition, language skills, behavior, motor skills, attention, and sleeping patterns. This tool has been used in Brazil, India, Italy, Kenya, and the United Kingdom in children aged less than 14 weeks of gestation to two years. The measure, however, is appropriate for a narrow age range from 22 to 26 months and was designed for children from middle- and upper-income families; it may require adaptation for children from low-income backgrounds (Fernandes et al. 2014).

SIDEBAR 4.7 Screening Tools Developed in High-Income Countries

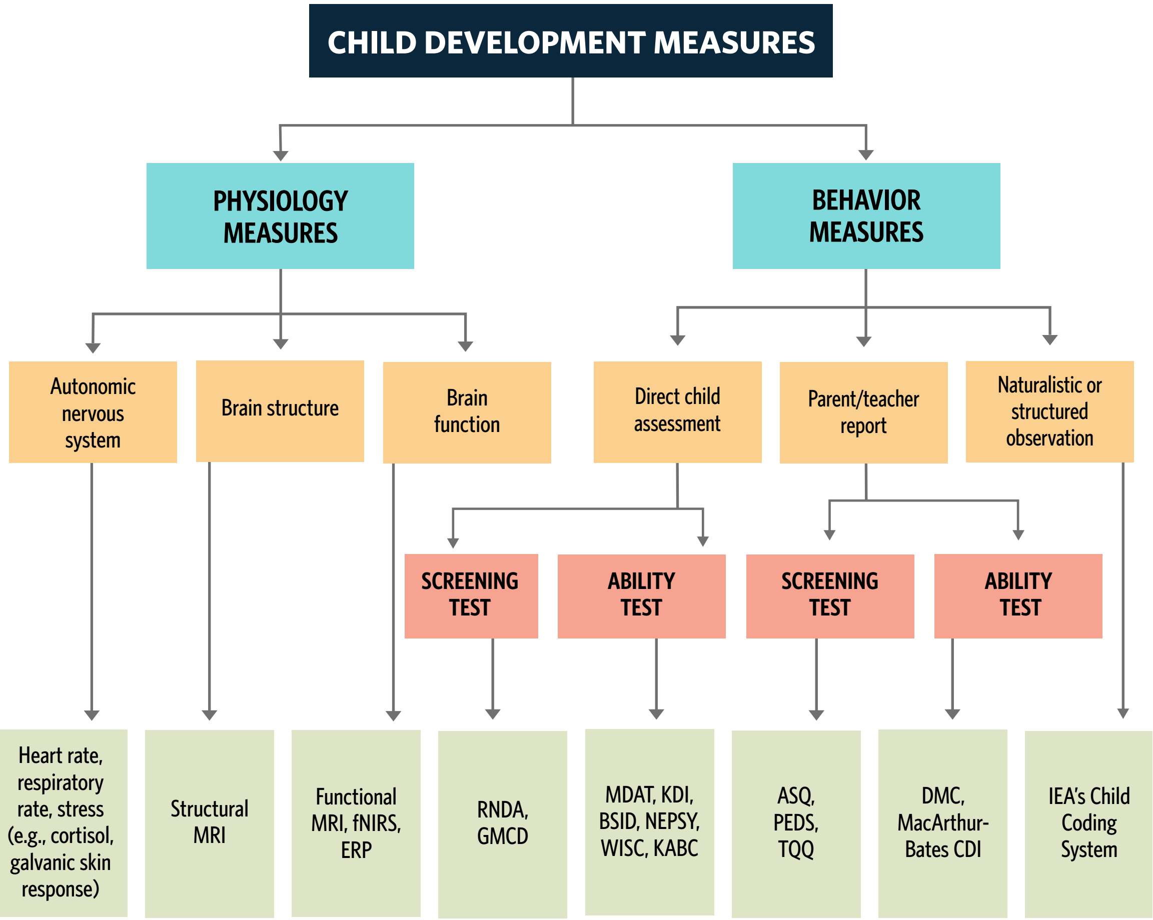
- **The Ages & Stages Questionnaire (ASQ)**, often used in the United States by home visitors to screen for developmental delay or to recommend intervention, is a low-cost and easily administered, comprehensive checklist of developmental milestones. The assessment is a parent report and can be completed by parents alone or administered by a trained assessor. The subscales measure skills in the categories of communication, gross motor skills, fine motor skills, personal-social skills, and problem-solving (or cognitive) domains. The questionnaire is designed for use with children aged 4–60 months, with questions focused on different stages of growth set at two- to three-month intervals. Scores are normed to indicate whether children are developing age-appropriately, but the test does not provide standardized scores, as are available for the Bayley Scales (BSID). The questionnaire is both less detailed and less validated than the BSID, but it may offer an opportunity to systematically obtain information about when children are reaching developmental milestones in diverse contexts. It has been used in a wide variety of countries and contexts.

- **The Strengths and Difficulties Questionnaire (SDQ)** screens for behavioral difficulties in children aged three years and over (Goodman 2001). It is freely available and has been translated into many languages and used in many countries and contexts (Woerner et al. 2004). Subscales provide scores for prosocial behavior and social-emotional difficulties, which can be further divided into emotional symptoms, conduct problems, hyperactivity or inattention, and peer relationship problems.

Types of child development measurements

Early childhood development assessments can be divided into physiological measures and behavioral measures (Figure 4.2). Physiological measures include measures of autonomic nervous system function, brain structure, and brain function. Behavioral measures can be obtained through three methods: (1) direct tests of the child; (2) ratings or reports of the child's behaviors or skills by informants, such as parents, usual caregivers, or teachers; and (3) observation of the child in daily or structured activities (Snow and Van Hemel 2008; Grigorenko and Sternberg 1999). Many tests combine two or more modes of assessment. These methods of individual assessment can be aggregated across groups to create a population-based measure.

FIGURE 4.2 **Taxonomy of Child Development Measures**



Notes: MRI, magnetic resonance imaging; fNIRS, functional near-infrared spectroscopy; ERP, event-related potential; RNDA, Rapid Neurodevelopmental Assessment; GMCD, Guide for Monitoring Child Development; MDAT, Malawi Developmental Assessment Tool; KDI, Kilifi Developmental Inventory; BSID, Bayley Scales of Infant Development; NEPSY, Developmental Neuropsychological Assessment; WISC, Wechsler Intelligence Scale for Children; KABC, Kaufman Assessment Battery for Children; ASQ, Ages & Stages Questionnaires; PEDS, Parents' Evaluation of Developmental Status; TQQ, Ten Questions Questionnaire; DMC, Developmental Milestones Checklist; CDI, Communicative Development Inventories; IEA, International Association for the Evaluation of Educational Achievement

Direct tests

Direct tests assess infants by presenting stimuli, such as objects or sound, to evoke responses, or by asking young children to complete tasks or activities, such as stacking blocks, searching for a hidden item, naming objects, or climbing stairs. Assessors are usually required to complete training on how to administer and score the test and are often professionals who regularly interact with children in some capacity (e.g., pediatricians, psychologists, or teachers). However, other personnel with relevant backgrounds (e.g., community health workers or social workers) can also be trained to conduct these tests. A professional level of training is not necessary for the administration of the tests in an evaluation setting, but a licensed professional would be required to interpret or make a diagnosis for clinical purposes. Examples of direct assessments are the Bayley Scales of Infant Development and the Wechsler Intelligence Scale for Children (WISC). The pros and cons of this approach are outlined in Table 4.2.

TABLE 4.2 Direct Assessment Pros and Cons

⊕ PROS	⊖ CONS
Less bias. Data are gathered first-hand. Information gathered via direct assessment (i.e., requiring responses to fairly structured requests by an adult who may or may not be known to the child) is considered to be an ideal measure of development because there is no concern about recall bias.	**Difficulty of testing young children.** Young children may be shy or refuse to participate in the assessment around strangers, particularly if the child is brought to an unfamiliar location or a location with a negative association, such as clinic where the child has previously been given a shot. The circumstances of direct testing are likely to be unfamiliar to young children—particularly those living in impoverished conditions—and may affect their engagement with the test items. Many tests include a behavior rating scale completed by the tester to indicate the child's mood and interaction with the assessor during the test activities, which can be used as a covariate during analysis.
High quality. Data gathered directly from children can be very high quality with a highly trained interviewer, as well as less biased than a parental report.	**Inaccuracy.** Performance on standardized tests may not be indicative of some children's true abilities (Bracken 2007). Optimal assessment may be challenged by children's internal states (hunger, sleepiness) or other behaviors, such as high activity level, distractibility, shyness with adults, low thresholds for frustration, and fatigue, fussiness, or defiance. Tests that include tasks or activities that are new to the child, use unfamiliar words or language structure, require verbal (rather than demonstrative) responses, or require children to choose between qualitative ("best" or "worst") or quantitative ("more like this" or "less like this") responses will likely reduce the accuracy of the assessment (Snow and Van Hemel 2008).
Benchmarking potential. Standardized norms are sometimes available within country and can allow for comparison with other children. For research purposes, however, changes over time or comparison with a control group do not require the use of standardized norms.	**Sampling challenges.** Sampling may be difficult, but it depends on the goal of the study. If the population of interest includes all children, and not all children are in the group measured (e.g., children attending preschool), it is more difficult to obtain a truly random sample because non-attenders are not included.
Cons that are easily overcome. Many of the "cons" listed can be overcome with good planning—by adapting tests and administrative procedures (see Chapter 6), by scheduling the test during a time when the child is alert, by being familiar with the test so that it moves along seamlessly (without having to fumble around for the proper materials, for example), and by altering the pace of the test in response to the child's behavior (Bradley-Johnson and Johnson 2007).	**Need for highly trained and standardized testers.** Accurate assessment of infants is largely dependent upon testers being able to control the infant's state of arousal, which may be challenged by new stimuli, environments, or unfamiliar persons. As a result, assessments may be more indicative of abilities capable of being demonstrated under novel (and perhaps exciting, or upsetting) situations rather than of true mastery in any domain (Snow and Van Hemel 2008). Testers also differ in their ability to exert this control, which introduces variance in scores due to tester skill and manner of interaction with the child.

Ratings and reports

Ratings and reports are scales or checklists completed by informants who know the child well, such as parents, caregivers, or teachers. The informant answers questions about the child's abilities based on what he or she knows about the child, but does not directly assess the child. Ratings and reports can offer information about how children behave in other (i.e., not standardized testing) settings (Snow and Van Hemel 2008; Squires et al. 1998). The rater may simply report about whether a behavior has occurred and how frequently, as in the parent-reported Ages & Stages Questionnaires (Bricker and Squires 1999). The rater may also be asked to compare the child with other similar children of the same age. Table 4.3 outlines the pros and cons of this approach.

Gathering information from both parents and teachers can provide a more complete picture of child behavior. Children's behavior may be different at home and at school, therefore differences between parent and teacher ratings may reflect true differences in child behavior. However, it is difficult to distinguish whether such differences are due to true differences or due to reporting bias by one of the respondents.

TABLE 4.3 **Ratings and Reports Pros and Cons**

PROS	CONS
Ease of administration. Instruments are easier to administer than direct tests or observations. Ratings are usually easy for respondents to understand, requiring minimal instruction or training; are efficient to use in terms of time and money; tend to be quick and easy to complete; and do not require much time or expertise for scoring and interpretation (Johnson and Marlow 2006). Parent reports may also be used to estimate stages of development where direct tests cannot be used.	**Potential for score inflation.** Teachers may inflate scores if they are used for accountability (Snow and Van Hemel 2008). Respondent bias is particularly problematic in a parenting intervention that teaches parents about child development and encourages developmentally stimulating activities. In this case, parents may feel pressure to provide the "correct" response even if it is not the reality. Across cultures, there may be different tendencies to inflate or deflate scores.
Accuracy. Parents observe their children's behavior over time in a wide range of circumstances. Therefore, parents can have more accurate knowledge of their children's abilities than a stranger is able to observe in the limited time of an assessment session. With older children (aged 3–8 years), teachers may be valuable informants of child development as they have multiple, repeated occasions to observe what children can do, how they behave in a variety of situations, and how this compares with peers of the same age (Snow and Van Hemel 2008).	**Inaccuracy.** Parents may not accurately report abilities. A mother with less education may not be willing or able to report accurately on her child's abilities. If an item is unclear to the respondent, there may be a tendency to simply agree. Teachers might also pay more attention to, or favor, certain children over others, which could bias their ratings. When class sizes are very large, teachers may not have the opportunity to observe each child closely, and may not have accurate perceptions or memory of the behavior of all their students.
Validity. Parent ratings correlate well with direct measurements. Such ratings are used widely within the United States (Bricker and Squires 1999; Doig et al. 1999; Scarborough et al. 2007) and in some low- and middle-income countries (Handal et al. 2007; Heo, Squires, and Yovanoff 2008). Evidence suggests that parents across socio-economic levels can provide accurate measurements of children's development as validated by direct child measurements.	**Variance in interpretation of items.** Parents and teachers may have systematically different interpretation of items in different cultures. For example, cultural norms about how children are supposed to behave at home or in the classroom (e.g., obedience, not speaking to adults beyond greetings) may affect how children are rated, and the intended meanings of the items may be lost.
	Teacher fatigue. Teachers may become confused or fatigued if asked to rate a large number of children. Completing a large number of reports could be a heavy burden, and teachers might not be motivated to put equal thought and care into each report.

While research on the use of teacher reports for the purpose of evaluating programs is scarce, evidence shows that early child educators in the United States can be trained to reliably use an observation-based rating measure (Bagnato et al. 2002). The Early Development Instrument (Sidebar 4.1) is a simple teacher rating measure that requires minimal training and appears to be reliable (Janus and Offord 2007). Moreover, an analysis of its use with some 40,000 children suggests that teachers in certain settings can make unbiased ratings across groups of different children (Guhn, Gadermann, and Zumbo 2007).

Parent ratings can be adapted for better reliability and validity. The following variables of interest (for which parents are reporting about their children) could be considered: (1) current and age-appropriate behaviors, and (2) behaviors likely to occur frequently. Responses that rely upon recognition rather than recall can be more accurate. Ensuring that items and response choices are spoken or written in language that is suitable for populations with low literacy rates is essential.

Observational measures

Observational measures rely upon a trained observer to document the behaviors of a child. Observational ratings may be completed at home or in an institutional setting (e.g., school or daycare facility), but in all cases, observers must be trained. Observational ratings could be recorded in real time or ratings could be made later by viewing videos. Table 4.4 outlines the pros and cons of this approach.

There are three kinds of observational measures that are generally used:

- ***Naturalistic observations.*** Naturalistic observations require the observer to follow the child and observe and record behavior in the normal course of the day. These observations are useful to identify characteristic environments, to detect the meaning of behaviors and skills and capacities, and to find out the cognitive requirements in a child's life. They are often a valuable complement to a standardized assessment.
- ***Sampled observations.*** With sampled observations, specific behaviors can be defined (e.g., caregiver questions a child) and the frequency of these behaviors is observed over a period of time. If the behavior is short and of relatively frequent occurrence (e.g., waving "bye bye"), a time-sampling method can be used. If the behavior can vary in length (e.g., a child's crying), then one can assess an event and its duration. For an example, see the International Association for the Evaluation of Educational Achievement (IEA) observation system available for download on its website.
- ***Structured situations.*** Structured situations are created, and then children are observed in that situation, with a common coding method, to see how they behave. For example, the Strange Situation has been used in many parts of the world to measure a child's attachment to her mother (Ainsworth 1993). The protocol has the mother leaving and reuniting with her child, and the child's response to the returning mother is coded. Other well-known measures are the HOME scale, in which the interviewer observes the caregiver's behavior (Bradley and Corwyn 2005); a book-reading task in which the mother is asked to read a book with her child (Aboud 2007; Rasheed and Yousafzai 2015); observation of play with specific toys in a controlled situation (Wachs et al. 1992; Wachs 1993; Wachs and Desai 1993); and measures of infant emotions and infants' inhibition to respond when they are presented with something novel (Leerkes and Crockenberg 2003; Rubin et al. 2006; Rubin et al. 2006). All of these have been used in several cultures.

TABLE 4.4 Observational Measures Pros and Cons

PROS	CONS
High validity. Because observational measures are based on actual behavior, they are likely to be valid or "true" indicators of typical behavior.	**Requirement of high effort and training.** Collecting this information well requires careful development and examination of the behavioral codes to be used and extensive training of observers to achieve reliability of coding. These assessments are also more time-intensive per child than a test. Some tools involve conducting observations while simultaneously interviewing the parent, which requires adequate practice to develop this skill.
Reflection of true context. These measures allow the observer to determine how the child will behave in an identified context (i.e., home or preschool). They may also help the investigator to develop other, more appropriate measures.	**Requirement of cultural adaptation.** Cultural appropriateness must be determined. Some situational measurements may not be appropriate to all contexts, or may be interpreted very differently depending on the culture.
Provision of confirmatory information. These measures provide additional or confirmatory information about measurement in any domain and thus may be useful to complement other tests and measures.	**Observation bias.** The presence of the observer in the household can change the behavior of children and caregivers. One strategy to reduce this problem is to inform the family members that you are interested in observing a certain type of behavior, for example the child's activity level, when you are actually more interested in something else, such as the child's speech and language environment.
	Difficulty in controlling timing and duration. It may be difficult to control the timing and duration of observation, which is necessary for reliable and comparable information, particularly during home visits. For example, if one child is observed during feeding time and another during sleep time, maternal-child interaction would vary. Naturalistic observation must be of sufficient duration to pick up typical daily practices. Another option is structured observation, in which case observation is conducted during a standardized task or activity, such as when the mother and child are looking at a book together.
	Time-consuming data entry, coding, and analysis. This could be mitigated if observational codes and definitions are clearly defined before data are collected.

Screening tests

Screening tests (e.g., the Denver Developmental Screening Test or the Ages & Stages Questionnaires) are brief measurements used to identify—with some degree of certainty—children who are at risk of having developmental problems in one or more domains (Glascoe 2005). Screens usually include motor, cognitive, and language domains but often do not measure social-emotional development. They are often used in lieu of ability tests because they are less expensive, quicker and relatively easier to administer, and require less time for training. Screening tests may rely upon direct child testing, parent report, or both.

Because screening tests only contain a sample of items per domain (i.e., they do not assess the full range of ability) they do not yield continuous scores, but are used to classify children into categories, such as "delayed," "at risk for delay," or "within normal limits" for age. These categories have been established for specific populations (typically a high-income country where the tests were developed) and do not apply to other populations (e.g., a low-income country). We recommend reporting screening tests by percent of children who are meeting standards as determined by national policy, or by comparing among groups of children ideally using different representative samples. If scores on a screening test show sufficient variance, groups can be compared on the basis of raw scores. However, screening tests usually contain few items per age group, and children who are developing on a normal trajectory usually score at ceiling. That means sufficient variance for comparison of raw scores is unlikely.

Screening tests are not diagnostic. These tests can be used, however, in samples where cutoffs have previously been determined to recommend further evaluation by a clinician or educational psychologist, to refer for intervention, or to monitor development. Screening tests are not appropriate in samples and situations where cutoffs have not been determined.

Ability tests

Ability tests include those designed to assess the maximum skill level for a child at any given age. In contrast to screening tests, they usually produce a normal distribution of scores in a typically developing population. These tests are often direct child measurements (e.g., the Bayley Scales of Infant Development) but can also be parent or other informant report by way of milestones or language checklists (Stoltzfus et al. 2001; Lansdown et al. 1996).

Ability assessments provide detailed, comprehensive information on children's developmental levels within domains and as a summary across domains. Scores are frequently standardized and can be used to compute developmental quotients (developmental age/chronological age x 100), or DQs. The main advantage of ability tests that produce continuous scores is that scores can be used to compare children's developmental levels with more precision. Scores may also be more sensitive to treatment effects, as compared to screening tests, because they measure differences at the upper end of the distribution as well as the lower. Scores of younger children (under three years of age) are typically labeled as developmental quotients, as they may still change, whereas the scores for older children are called intelligence quotients (IQs) as they become more predictive of future development. Some tests are diagnostic, assessing specific skills such as communication and can be used to recommend and design types of remedial assistance.

While ability tests can be time-consuming and require a high degree of training to conduct, they provide flexibility in how scores can be used (that is, as raw scores, DQs or IQs, or with cutoffs for determining delay as specified within a population). Standardization can be parametric or non-parametric (Rubio-Codina et al. 2016).

Tools using rapidly developing technologies

Rapidly advancing technology is creating possibilities for ECD assessment in low- and middle-income countries using methods that were not previously possible in these contexts. These methods include neuroimaging, eye-tracking, and other devices that measure physical activity, the autonomic nervous system, and the language environment. Research using these methods has the potential to shed light on critical questions in developmental science regarding the effects of early experience on brain development and to clarify which underlying neurobiological processes are disrupted by, or resilient to, early exposures. However, whether these techniques are useful as biomarkers to detect the effects of interventions remains to be established.

These methods may be relatively objective and less susceptible to cultural bias as compared to commonly used behavioral assessments. However, as for all developmental assessments, the extent to which these types of measures are likely to be influenced by culture depends on the measure and paradigm. Examples of measures that may be less susceptible to cultural bias are measures of blood pressure, galvanic skin response, and an expected electrophysiological response in newborn infants. Examples of measures that may be more susceptible to cultural bias are tasks in which the stimuli and expected responses are shaped by Western ideas and assumptions, which could be true of some neuroimaging or eye-tracking paradigms.

A high level of technical expertise would be necessary to adapt new technologies into impact evaluations, and in some cases, their high costs could prevent widespread deployment at this time. Still, further advances are underway and these types of techniques may offer promising assessments for the future. For a more complete review of neuroimaging methods, including magnetic resonance imaging (MRI), magnetoencephalography, and positron emission tomography (PET scan), see Sizonenko et al. (2013). Below we discuss a number of assessment tools that harness advanced technology; see Table 4.5 for a comparison of these devices.

- **Functional Near-Infrared Spectroscopy (fNIRS)**

  Measures of a child's neural activity in response to language, objects, and social stimuli also help contribute to our understanding of child development. Near-infrared spectroscopy (NIRS) measures neural activity in the surface layers of the cortex indirectly by measuring changes in blood oxygenation. Active neurons require oxygen at a greater rate than inactive neurons. Release of oxygen to neurons causes changes in the relative levels of oxygenated and deoxygenated hemoglobin and changes in the color of the blood. NIRS measures those changes in blood oxygenation non-invasively, through light sources and detectors placed on the scalp. The light sources shine infrared light up to 2–3 cm from the scalp into the cortex and the detectors measure the color (chromophore concentration) of the light that is reflected back (Ferrari and Quaresima 2012). As a rule, only the cortical surface can be visualized; regions that lie in a sulcus, which is a groove in the folds of the cerebral cortex, may be difficult or impossible to visualize.

Functional NIRS (fNIRS) measures these changes in blood oxygenation time-matched in response to a visual or auditory stimulus. For example, in one experimental paradigm, children view a video that periodically shows social stimuli (a person's face) and non-social stimuli (an inanimate object). For each type of stimulus (social or non-social), the changes in blood oxygenation in response to that type of stimulus are averaged across trials and participants, then subtracted to identify the areas of cortex that show more activation in one condition than the other. It has been shown that children at risk for autism show a different pattern of activation to social versus non-social stimuli than children who are not at high risk (Lloyd-Fox et al. 2013).

In high-income countries, fNIRS has been used to investigate the early cortical localization of perceptual abilities and language processing, as well as developmental changes throughout infancy in the processing of language, objects, and social stimuli. For reviews, see Gervain et al. (2011) and Vanderwert and Nelson (2014). fNIRS has been shown to be feasible to set up and use in a rural clinic in the Gambia (Lloyd-Fox et al. 2014) and more recently in Bangladesh (Nelson 2016). In the Gambia, the equipment was transported to the clinic in a 4x4 vehicle on unpaved roads and the first participant began testing within 2.5 hours of the arrival of the equipment.

■ Event-Related Potential (ERP)

Event-related potentials (ERPs) refer to measured brain responses that are the direct result of a specific sensory, cognitive, or motor event. Similar to near-infrared spectroscopy, they are also measured non-invasively using headgear placed on the scalp, and they provide a measure of the extent to which children are responding to certain stimuli. The headgear contains electrode sensors that measure changes in the electrical activity of groups of cortical neurons in response to a stimulus. These voltage wave-forms that respond to a stimulus are called ERPs. Characteristic peaks in the ERP wave-form, which are generally found by averaging multiple responses to the same type of stimulus, are called ERP components (Nelson and McCleery 2008). Specific ERP components, each of which is presumed to reflect a different neural and cognitive operation, are defined by their latency (how long after the stimulus they occur), amplitude (of the voltage), scalp distribution (the location of the electrodes that detect the signal), and polarity (whether their voltage is positive or negative). ERPs provide highly accurate temporal information regarding the timing of the electrical activity, but it is difficult to identify spatially the source of the activity within the brain (Ullman 2014).

In high-income countries, abnormal ERP patterns (e.g., different from the typical pattern in latency or amplitude) have been identified to indicate delayed brain maturation in children with many types of atypical development, including learning disabilities, dyslexia, autism, and attention deficit disorders. For reviews, see Taylor and Baldeweg (2002) and Nelson and McCleery (2008).

Differences in ERP patterns have also been found between children from high and low socio-economic status in the United States (Pavlakis et al. 2015). In one study, children from high socio-economic status (SES) households showed a typical ERP pattern, while children from low SES households showed a pattern similar to that observed in patients with neural damage in the prefrontal cortex (Kishiyama et al. 2009).

Unlike fNIRS, ERPs have been used to evaluate the effectiveness of interventions. A randomized trial of a family-based intervention that provided training to both parent and children among families with children attending the Head Start program in the United States showed effects on both child cognitive skills and changes in ERPs during an attention task. Specifically, it showed that the general cognitive effects were at least partly due to changes in function of the neural structures that underlie selective attention, which is the capacity to respond to certain stimuli selectively when presented with multiple stimuli simultaneously (Neville et al. 2013). In a baseline assessment of a randomized trial in Romania, institutionalized children (averaging age 22 months) showed reduced ERP amplitudes in response to happy, fearful, angry, and sad faces, compared to children raised by their biological families. After 20 months, children who had been randomly assigned to foster care showed ERP amplitudes in the same task that were midway between those displayed by the continually institutionalized and never-institutionalized children (Nelson and McCleery 2008). Among children in low- and middle-income countries, ERP patterns that deviate from the typical pattern have been found following severe malaria in Kenya (Kihara et al. 2010) and iron deficiency in Chile (Roncagliolo et al. 1998), as well as among a group getting unfortified formula in a study that investigated the effects of fatty acid fortified infant formula in Turkey (Unay et al. 2004).

■ Eye-Tracking

Infants' eye gazes provide meaningful information about their cognitive processing. For example, an infant's novelty preference, demonstrated by looking longer at a new picture compared to a picture that he or she has seen before, shows that the infant remembers the previously seen picture. For decades, researchers have been coding infant looking time by observation of video recordings, but advances in eye-tracking technology have made it much more efficient to evaluate children's fixations and saccades. A fixation is looking at a specific point for a period of time, and a saccade is a rapid eye movement aimed at bringing an object into focus. Eye-trackers typically track looking behavior using an infrared light source and one or a set of cameras to capture the infrared light reflected from the cornea. In cognitive tasks, eye gaze is tracked while children are looking at visual stimuli such as pictures and videos presented on a screen. Looking behavior is quantified by temporal information (e.g., duration of gaze in an area of interest, time to first fixation), spatial information (e.g., fixation position, fixation sequence), and counts (e.g., fixation count, saccade count) (Lai et al. 2013).

Many studies in high-income countries have delineated typical developmental trajectories of performance on saccade tasks, as well as deviations in children with developmental disorders (Karatekin 2007; Feng 2011). For example, children with attention deficit hyperactivity disorder perform poorly on tasks in which they are instructed to look at the space on the opposite side of a screen from a visually presented stimulus, reflecting difficulties with inhibition (Karatekin 2007). In older children and adults, eye-tracking methods have been used to investigate information processing during a variety of cognitive tasks, such as reading, scene perception, visual searching, music reading, and typing. Several published reviews and books present overviews on these methods (Radach, Kennedy, and Rayner 2004; Rayner 2009; Duchowski 2007; Holmqvist et al. 2011).

Tests measuring information processing using infant looking times may be more sensitive to detect intervention effects than global infant development tests, such as the Bayley Scales. For example, in a randomized trial of docosahexaenoic acid (DHA) supplementation in infancy, no differences between intervention groups were found on Bayley scores at age 18 months. However, group differences were found in sustained attention using a visual habituation task at four, six, and nine months, indicating enhanced attention in infants who received higher doses of DHA (Colombo et al. 2011). In visual habituation tasks, an infant's looking time is recorded as he or she views repeated presentations of the same picture. Declining looking time with repeated presentations is interpreted to indicate learning of the picture. Interestingly, a follow-up study of the same DHA trial found differences between intervention groups in several cognitive tasks at age five years. This suggests that the visual habituation task, but not the Bayley, was sensitive to early cognitive effects of DHA that were also detected in later childhood (Colombo et al. 2013).

Eye-tracking has been used extensively in South Africa and has also been used in other countries in Africa and Asia. A recent study tested the feasibility of the technique for assessing visual-orienting and sequence-learning abilities in nine-month-old infants in rural Malawi. A high percentage of parents (92 percent) found the method acceptable, 90 percent of infants were able to complete the entire test, and a moderately high percentage of the test trials were valid (68–73 percent). Test completion rates were slightly higher for eye-tracking (90 percent) than for standard behavioral developmental assessment (87 percent) (Forssman et al. 2016).

■ Accelerometers

An accelerometer is a small device that can be worn on a child's wrist, ankle, or hip, over a period of a few days or a week, and provides a continuous objective measure of physical activity. The device measures the frequency and magnitude of the body's accelerations in one or more planes of movement. Activity is typically recorded several times per second. The raw data can then be used to generate activity counts over a specific time interval or epoch (e.g., five or 30 seconds) (Loprinzi and Cardinal 2011). Validation studies have established different activity count thresholds for different devices, wear locations (where the device is placed on the body), and age groups to make it possible to quantify the proportion of time an individual spent in different activity classes (e.g., sedentary activity, moderate physical activity, and vigorous physical activity) (Pulakka et al. 2013; Rothney et al. 2008; Wong et al. 2011; Kim, Beets, and Welk 2012). Popular accelerometers used in research are ActiGraph (ActiGraph, LLC), Actical (Philips Respironics), and ActivPAL (PAL Technologies, Ltd). Hybrid devices are also available, which combine an accelerometer with other autonomic nervous system measures, such as heart rate and galvanic skin response (e.g., from Intel Corp. or Empatica Inc.). These allow the continuous measurement of stress along with physical activity (Sun et al. 2012).

The feasibility of using accelerometers in large-scale surveys has been shown in the United States (Lee and Shiroma 2014) and in low- and middle-income countries (Katzmarzyk et al. 2015; Dugas et al. 2014). In young children, accelerometers have been used to evaluate the consequences of severe acute malnutrition in Ethiopia (Faurholt-Jepsen et al. 2014) and the effects of a nutrition intervention on physical activity in Malawi (Pulakka et al. 2015).

■ Language Environment Analysis (LENA™)

Language Environment Analysis (LENA™) is an automatic speech recognition system, which can be used to study children's language environment, as well as their own vocalizations. A child wears clothing with a specialized pocket to hold a recording device that continuously records the child's speech and what is spoken to the child within a 1.2–1.8 meter radius. The LENA™ analysis software processes the audio recordings and incorporates speech recognition algorithms to differentiate speech-related sounds from environmental background noise, such as television. The software yields automated measures of adult word counts and child vocalizations, as well as interaction frequencies (i.e., conversational turns between the target child and an adult) and durations of talk between speakers (Greenwood et al. 2011). The LENA™ systems cannot distinguish between child-directed and adult-directed speech that is produced "near and clear" to the child, cannot differentiate among different adult speakers, and does not transcribe the words that are spoken. The LENA™ recorders are particularly susceptible to error in noisy conditions where speech to the child is overlapping with other background noise.

LENA™ was developed for English and has been shown to produce accurate results in French (Canault et al. 2016), Spanish (Weisleder and Fernald 2013), and Chinese (Gilkerson et al. 2015). It has been used to study language development and the linguistic environment of children with hearing loss, autism, and preterm birth (VanDam et al. 2015; Warlaumont et al. 2014; Caskey et al. 2011). LENA™ was used to evaluate the impact of an intervention to increase mother-child interaction among low socio-economic status families in the United States. This study showed increased adult word tokens, conversational turn counts, and child vocalization counts in the intervention group (Suskind et al. 2016). LENA™ was also used to evaluate a parenting intervention in Senegal (Weber, Fernald, and Diop 2017).

TABLE 4.5 **Comparison of Devices for Measuring Brain and Behavioral Development**

TECHNIQUE	BRIEF DESCRIPTION	AGE	COST	ADVANTAGES	DISADVANTAGES	WHY MEASURE THIS?
FNIRS	Charts neural activity by measuring changes in blood oxygenation triggered by light stimulus and captured by detectors placed on the scalp	Ideal for young infants; can be used from the newborn stage through childhood	Headgear: $5,000–$45,000 fNIRS system: $100,000–$400,000 To rent rather than purchase an fNIRS system can cost about $3,000 per month	• Non-invasive • Silent, unlike MRI, enabling easy presentation of auditory stimuli • High spatial resolution of neural activity • Less sensitive to motion artifacts than ERP • Relatively portable	• Expensive • High level of expertise required • Can only measure surface layers of cortex • Temporal resolution lower than ERP • Cross-cultural adaptation of stimuli required • Lack of established paradigm for evaluating interventions	• To examine effects on neural function that might not be detectable by behavioral measures • To examine the underlying neural mechanisms of individual and group differences in behavior • To examine hypotheses regarding intervention effects on specific neural systems, e.g., effects of specific nutrients based on their biological role in brain development
ERP	Measures electrical activity of groups of cortical neurons in response to a stimulus channeled through electrode sensors placed on the scalp	Can be used from the newborn stage through adulthood	< $30,000–$100,000	• Non-invasive • High temporal resolution	• Expensive • High level of expertise required • Highly sensitive to disrupted signal from movement or eye blinking • Low spatial resolution	Same reasons as for fNIRS
EYE-TRACKING	Measures looking behavior, including the location, duration, and shifting of gaze of individuals as they view pictures or videos	Can be used from the newborn stage through adulthood	< $5,000–$25,000	• Non-invasive • Able to measure a wide variety of cognitive processes	• Expensive • Technical expertise required	May be especially useful in very young infants for whom traditional behavioral assessments are not very sensitive
ACCELEROMETERS	Continuously measures physical activity through small device worn on hip, wrist, or ankle; hybrid devices also measure heart rate and galvanic skin response	Can be used from infancy through adulthood, though interpretation is more difficult at ages when children are often carried	$250 per device plus $800–$1,700 per software license; hybrid devices with multiple sensors cost up to $1,700 per device	• Easy to use • Objective measure of continuous physical activity or stress • Sensitive; able to detect short bursts of activity • Useful for collecting data during sleep • Useful for assessing role of physical activity or stress as mediators	• Expensive • Time-consuming data collection (up to seven days) • Some technical expertise required • Data lacking contextual information	To get an objective measure of continuous physical activity or stress
LENA™	Continuously measures child speech and language exposure through small device worn in specialized pocket sewn onto clothing	Can be used from infancy through childhood	$399 per device plus $8,400 for the software	• Easy to use • Minimal effort for data collection • Rich, objective quantitative information from the child's natural language environment	• Expensive • Unable to automatically transcribe speech content or distinguish among adult speakers • Legal and ethical considerations • Time-consuming manual coding • Extended recording time needed • Substantial space required for data storage	To obtain rich, objective quantitative information from the child's natural language environment

Note: Costs are highly variable and very likely to change over time. These estimates are ranges and are only accurate as of December 2017.

Electronic data collection and testing

Data collection by laptop, tablet, and smartphone can be more efficient than using paper-and-pencil forms, which require subsequent data entry into a computer. Many applications have been developed for computer-assisted personal interviewing (CAPI), which can be used in developmental assessments (see Sidebar 4.8). During administration of a traditional developmental assessment, data collectors can directly enter each item score into a CAPI application rather than on paper. Various CAPI applications and their advantages and disadvantages were comprehensively reviewed by Shaw et al. (2011).

Laptops, tablets, and smartphones can also be used to conduct computerized cognitive tests, in which the child interacts directly with the device. The child's score is typically calculated based on the accuracy and speed of his or her responses. These types of tests can generally be administered to children as young as age four years, and even those as young as age two to three years, if the tests involve simple tasks. In a recent study in the United States, a simple touch-screen word recognition task was administered to children aged one to four years. While one-year-olds completed an average of only 44 percent of trials, children aged two to four years completed 86 percent to 100 percent of trials (Frank et al. 2016). See Frank et al. (2016) for a useful overview on developing tablet-based tasks for young children.

Computerized tests, in which the child interacts directly with a computer, tablet, or smart phone, have several advantages over traditional tests, including minimal verbal instructions, relatively quick administration time, and precise information on response times. These precise response time scores may be more sensitive to an intervention or to picking up group differences than accuracy scores such as the percent of total questions that are correct. However, children who are unfamiliar with these devices may require extensive practice to grasp the task, and these tests are generally not appropriate for children under age three to four years. Table 4.6 summarizes the benefits and drawbacks of computerized testing.

TABLE 4.6 **Pros and Cons of Using Computerized Testing**

⊕ PROS	⊖ CONS
Minimal tester effects. Scores will be minimally affected by differences among data collectors. Traditional developmental assessments tend to be strongly influenced by the way the tester interacts with the child, despite efforts to standardize administration procedures. In many computerized tests, verbal instructions are minimal. Rather, the task is demonstrated on the screen and practice items are repeated until high performance shows that the child has understood the instructions, before continuing to the test items.	**Lack of familiarity with devices.** Children may not be familiar with computers or tablets, which could affect their performance. This can be mitigated by providing extensive practice before administering the test. However, this would add to the time required for the assessment. Similarly, the tests may be intimidating to family members.
Minimal verbal instruction. Minimal verbal instruction is an advantage when transferring a test from one language to another.	**Required technical expertise.** Administration of these tests requires a certain level of technical expertise and may be cumbersome for assessments that require a lot of manipulatives (e.g., Bayley Scales of Infant Development).
Quick administration. These tests are relatively quick to administer (1–7 minutes per test).	**Restriction to older children.** The method is probably not appropriate for children under 3–4 years of age, except for tests featuring the simplest tasks.
Low likelihood of missing data. Tablets can be programmed to force a response before an interviewer or respondent moves on to the next question.	**Distraction potential.** The tests may be distracting for children, especially the youngest children. The tests may also increase the likelihood of interference from other adults and children who are curious to see what the child is doing (thus limiting the standardization of the test setting).
Precision. The tests can yield accurate, fine-grained information such as response time in milliseconds. These precise measurements may be more useful for detecting the effects of interventions.	

SIDEBAR 4.8 Examples of Computerized Tests

- **The Cambridge Neuropsychological Test Automated Battery (CANTAB)** is comprised of 12 tests designed to measure frontal and temporal lobe function, with four tests in each of three domains: visual memory; visual attention; and working memory and planning. CANTAB has been used extensively in research in high-income countries. It can be administered to children as young as four years old. CANTAB has been used in a number of studies in low- and middle-income countries, including Brazil (Roque et al. 2011; Teixeira et al. 2011), Costa Rica (Lukowski et al. 2010), and Malawi (Nkhoma et al. 2013) in children as young as six years old. An advantage is the widespread use of the tool, which should help for interpretation of the results and making comparisons of the results with results of other studies. There is also evidence on which specific brain systems are relevant for the tasks, which should help with connecting CANTAB scores and other observed behaviors that might also be guided by those systems. However, a disadvantage is the high cost (about $10,000 for the full battery).

- **Cogstate** is another commercially available battery of computerized tests that has been widely used in high-income countries, with hundreds of peer-reviewed papers reporting use of the tool. It consists of 11 tests of cognitive skills, such as attention, executive function, memory, and learning. In Ugandan children ages 5–13 years, Cogstate subtests showed low to moderate test-retest reliability (correlation ranging from 0.3 to 0.6) and concurrent validity, as compared to subtests of the Kaufman Assessment Battery for Children (KABC-II) (r ranging from 0.2 to 0.5) (Bangirana et al. 2015).

- **The NIH Toolbox for the assessment of neurological and behavioral function** is a set of computerized tests for ages 3–85 years in four domains: cognition (e.g., attention, executive function, memory), emotion (e.g., psychological well-being, social relationships), motor skills (e.g., balance, dexterity, endurance), and sensation (e.g., vision, hearing, taste). The tests were designed by leading academic experts in each respective domain to be open source tools. The goal is for them to become a "common currency" across research in diverse study designs and settings, and the Toolbox has begun to be widely used since its launch in 2012. It costs $5,000 per year for ongoing use of the web-based platform. Both methods of administration—the web-based platform and iPad app—require Internet access, which is not feasible in many low- and middle-income settings. However, offline versions may be available through contacting the researchers who developed the tests. An offline version of the Dimensional Change Card Sort task was implemented among children aged 9–12 years in Indonesia to evaluate the long-term effects of maternal, multiple-micronutrient supplementation (Prado et al. 2017). The NIH EXAMINER battery is a separate set of computerized tests of executive function, which can be used for subjects from age three years to adulthood. It can be administered offline, and is available for free (Kramer et al. 2014).

- **The Psychology Experiment Building Language (PEBL)** test battery is an open-source set of computerized tests. Software for the creation of new tasks is freely available, along with a set of preprogrammed tasks, including the Iowa Gambling Task, Wisconsin Card Sorting Test, and Conners' Continuous Performance Test. We are not aware of use of the PEBL battery in low- or middle-income countries.

- **The Rapid Assessment of Cognitive and Emotional Regulation (RACER)** is an open-source tool that has been used in low- and middle-income countries. RACER is a set of four short cognitive tests (1–4 minutes long), which assess long-term memory, inhibition, working memory, and implicit learning. They can be administered offline using a tablet PC. RACER has been used in the Young Lives study among 4,000 children ages six to 12 years in Peru (Hamoudi and Sheridan, under review) and is currently being used with children as young as four years old in Ghana and Bangladesh. While the tests are available for free, interpretation of the data and use of the programs require fairly extensive consultation with the developers, Margaret Sheridan and Amar Hamoudi.

5 Measure Selection

KEY MESSAGES:

- There are many aspects of the ideal early childhood assessment, and the selection of any assessment tool will involve a trade-off between various advantages and disadvantages.
- Important ethical issues must be considered for whatever assessment is used.
- The *ECD Measurement Inventory*, an accompanying database of 147 developmental assessment tools, can assist in the selection process. For each tool, the database contains information regarding the domains assessed, age range for which the tool is appropriate, method of administration, purpose of the assessment, origin and locations of use, logistics, and cost.

THERE ARE MANY FACTORS THAT CONTRIBUTE TO MAKING AN ECD ASSESSMENT desirable. In Chapter 1, Table 1.1 discusses the characteristics of an ideal assessment, while also presenting the reality: An ideal assessment that meets all 10 criteria does not yet exist. The selection of any assessment tool will involve a trade-off between various advantages and disadvantages. For any project, the first step is to prioritize these criteria according to the project-specific purpose and constraints. The *ECD Measurement Inventory* provided with this Toolkit can then be used to select a tool that meets these project-specific criteria.

In the Measurement Inventory, the "Tests" worksheet contains a database of 147 developmental assessment tools. It contains information for each tool regarding the domains assessed, age range for which the tool is appropriate, method of administration, purpose of the assessment, origin and locations of use, logistics, and cost. The "Definitions" worksheet contains the definition of each column in the "Tests" tab.

To view tools that are appropriate for a given project, filter the database in the "Tests" tab by the project-specific criteria (e.g., domains to be assessed, minimum and maximum age of children to be tested). This will produce a list of tools that meet those criteria. Additional filters can specify the method of administration (child assessment, caregiver or teacher report, computer-administered test) and the purpose of the assessment (screening test, ability test, or population-level assessment).

Table 5.1 summarizes the 10 ideal characteristics of an assessment, followed by detailed recommendations on how to prioritize and achieve them.

Ideal characteristics of an assessment

TABLE 5.1 Ideal Traits of an ECD Assessment

Ideal 1	The test score represents the child's true ability.
Ideal 2	The test is appropriate, interpretable, and has high reliability and validity in all contexts and cultures.
Ideal 3	The test shows variance in scores at all ages and ability levels.
Ideal 4	The test is easy to administer.
Ideal 5	The test can be administered quickly and at low cost.
Ideal 6	The test provides information on all developmental domains.
Ideal 7	The test score is relevant to a child's practical function in daily life and therefore relevant to policy and program design.
Ideal 8	The test is a good indicator of future success.
Ideal 9	The brain systems and neural mechanisms underlying test performance are well-understood.
Ideal 10	The impact of health, nutrition, and environmental factors on the test score is well-understood.

IDEAL 1 *The test score represents the child's true ability.*

Priority for: All purposes.

What this means: As shown in Figure 4.1 in Chapter 4 a test score accurately represents the underlying ability it is intended to measure when the assessment tool has high reliability and validity and is implemented with high fidelity.

How to do this:
Chapter 6 discusses how to establish reliability and validity and high-quality implementation.

- ***Whenever possible, use multiple tests and methodologies to measure both within and across domains.*** Using two or three measures (e.g., parent report, direct child tests, or observation) to assess any domain provides a richer developmental profile than any single test could. The findings can then be analyzed in combination, and the combined results are more likely to indicate a more accurate, thorough, and "true" assessment of that domain (Grigorenko and Sternberg 1999). For example, the International Association for the Evaluation of Educational Achievement (IEA) Preprimary Project, conducted in 15 countries (Montie, Xiang, and Schweinhart 2006), used both observational measures as well as child-administered cognitive and language assessments at age four to examine the impact of preschool activities on development at age seven. Both types of data were useful in understanding later cognitive and language outcomes. The Turkish Early Enrichment Project (TEEP) (Kagitcibasi, Sunar, and Bekman 2001) also used multiple assessments (direct child tests, such as the Stanford-Binet, and parent reports of behavior) to evaluate the effects of a stimulation program on child development outcomes.
- ***Whenever possible, use multiple measures of the same construct and extract a latent factor.*** Latent constructs are complex underlying abilities that cannot be observed or measured directly. To capture a latent construct, researchers measure indicators that represent the underlying construct. In a factor analysis, the latent factor scores are calculated from the observed indicators. The latent factor theoretically represents the child's true ability, free from measurement error (Attanasio et al. 2017).

IDEAL 2 *The test is appropriate, interpretable, and has high reliability and validity in all contexts and cultures.*

Priority for: All purposes, if the goal is comparability across cultures and contexts.

A project designed to inform local policy in a single country or context would not necessarily prioritize this criteria, but instead would want to ensure high reliability and validity in that specific context.

How to do this: Chapter 6 will discuss how to adapt tests across contexts and how to establish reliability and validity.

IDEAL 3 *The test shows variance in scores at all ages and ability levels.*

Priority for: All purposes.

What this means: Test scores should show variance at all ages and for all ability levels that are targeted in the study. Many tests are appropriate only for a limited age range, while children outside that age range score at floor (minimum score) or ceiling (maximum score). When transferring tests from one context to another, the age range for which the test is appropriate may be different.

How to do this: Pilot testing should always be conducted to ensure that the test scores show variance in children at the target age in the local context. Screening tests are not designed to show variance in typically developing children, who normally score at ceiling. Therefore, these are not ideal for program evaluations aiming to analyze differences between groups. A screening test would be appropriate for a purpose such as an exclusion criterion, if a study wanted to exclude children who were developmentally delayed.

IDEAL 4 *The test is easy to administer.*

Priority for: Population monitoring. For program evaluation and hypothesis-driven research, the importance of this ideal depends on project-specific constraints.

What this means:

- ***Training*** (capacity for administration): Some tests require considerable time—one or two months—to adequately train and standardize testers.
- ***Test setting*** (home or school visit versus clinical or lab testing): Children may be uncomfortable being tested in locations that are unfamiliar, and they may perform poorly as a result, so home or school testing can be preferable. The drawback of this is that testing environments will vary according to characteristics that could, in themselves, affect performance (e.g., lighting, noise, seating).
- ***Capacity of respondent:*** In the case of using rating assessments, the ability of respondents (e.g., parents, teachers, doctors) to report accurately on children or rate children is critical to the success of the rating assessments. This issue is particularly important when respondents are illiterate, or have not experienced a similar situation.

How to do this: During training, standard good practice is to ensure that testers are reliable (see Chapter 6 for more details on how to do this). It is also critical to make the testing environments as homogenous as possible to minimize distractions and maximize consistency. For example, the research team could carry a folding table and two chairs and only test during daylight hours so that the testing environment itself is identical even if the location is not. Similarly, the research team could include someone whose job is to maintain a quiet atmosphere, and to keep away observers or other distracting onlookers. If researchers use an unfamiliar testing environment, they should try to make the place cozy and comfortable for the test takers.

IDEAL 5 *The test can be administered quickly and at low cost.*

Priority for: Population monitoring. For program evaluation and hypothesis-driven research, the importance of this ideal depends on project-specific constraints.

What this means:

- ***Time allocated for testing:*** Direct child tests will likely take 20–60 minutes; screening or parent report tests may take 30 minutes or less. Direct testing for infants and toddlers can take longer than expected if children are tired or become hungry during the course of testing. For instance, the Bayley Scales of Infant Development can take 30–90 minutes to administer, and interviewers are usually paid for training time in addition to administration time.
- ***Budget:*** Many standardized tests are prohibitively expensive for use in large-scale studies. For example, the Bayley Scales cost about $1,000 per test kit per interviewer conducting assessments, and could cost between $1 and $3 per child assessed, depending on how many subscales are administered. In addition, the test usually needs extensive amounts of pilot testing and adaptation of materials, which can add to the expense. The Ages & Stages Questionnaires, meanwhile cost about $200 per interviewer with no additional cost per child. Tests vary in terms of how much interviewer time they need. Fortunately, many instruments are non-proprietary and do not charge fees for their use, including the IDELA, MELQO, CREDI, MICS, EHCI, and the Regional Project on Child Development Indicators (PRIDI), to name a few.
- ***Copyright issues:*** Most of the tests developed and licensed in the developed world (e.g., Bayley Scales, Denver Developmental Screening Test, Woodcock-Johnson) are strictly protected by copyrights. In many cases, a licensed psychologist is the only person who can purchase the tests from the publishing companies. Copyright laws prohibit any use of the tests (including photocopying) without explicit permission or purchase. Furthermore, translation is not allowed without approval from the publishing companies' legal department. In the accompanying *ECD Measurement Inventory*, a column in the "Tests" tab called "Accessibility" provides information on this.

IDEAL 6 *The test provides information on all developmental domains.*

Priority for: Program evaluation and hypothesis-driven research.

What this means: Measuring multiple domains (e.g., language, cognition, social-emotional development) provides a more comprehensive assessment of child functioning and can also indicate which domains are or are not affected by an intervention.

IDEAL 7 *The test score is relevant to a child's practical function in daily life and therefore relevant to policy and program design.*

Priority for: Population monitoring and program evaluation.

How to do this: Alignment of ECD assessment with the content of national standards for preschool and primary grades may be important to ensure policy relevance.

IDEAL 8 *The test is a good indicator of future success.*

Priority for: Population monitoring and program evaluation.

What this means: Even in high-income countries, the predictive validity of many ECD assessments has not been evaluated, and even less evidence of predictive validity exists in low- and middle-income countries. The studies that have been conducted show that existing assessments for ages 0–2 years are generally poor predictors of later performance at school age, but become stronger by ages 3–5 years. Predictive validity within domains is stronger than across domains. For example, early language skills predict later language skills more strongly than they predict later social-emotional skills. Academic performance at school age is predicted by preschool measures of language, general knowledge, and executive function, while social-emotional and behavioral function at school age is predicted by early social-emotional skills and self-regulation.

IDEAL 9 *The brain systems and neural mechanisms underlying test performance are well-understood.*

Priority for: Hypothesis-driven research.

What this means: As discussed in Chapter 4, assessments that measure neural activity can provide new insights into the biological mechanisms through which interventions affect brain and behavioral development and can inform the design of targeted interventions to increase their effectiveness.

IDEAL 10 *The impact of health, nutrition, and environmental factors on the test score is well-understood.*

Priority for: Program evaluation and hypothesis-driven research.

How to do this: Programs addressing specific risk factors should select assessment tools that are most likely to show intervention effects. Studies that have examined the effects of the exposure on various aspects of early childhood development should be reviewed to inform test selection. Using tests that have been found to be sensitive to the exposure in previous studies can enhance the likelihood of finding a significant effect and replicating previous results. However, this criterion must also be balanced with innovation and should not preclude using methods that have not been previously examined in order to generate new findings.

Ethical risks and responsibilities in assessing young children

Beyond deciding which instruments to use, measurement teams must also be cognizant of the risks and responsibilities associated with the assessment of young children. All measurement protocols must be reviewed and approved by an ethical review board. Many universities and non-governmental organizations have Institutional Review Boards (IRBs) that fulfill this role. If investigators in the United States or another developed country are working with researchers in a low- or middle-income country, it is generally not sufficient to have approval just from the home institution of the developed-country investigator. Wherever possible, it is also essential to have protocols and permission forms reviewed by a review board in the country where the study is taking place. In the case where the person administering a child's assessment is not affiliated with an organization that has an ethical review board, an external institutional review can be sought. For example, the Western Institutional Review Board is an organization fully accredited by the Association for the Accreditation of Human Research Protection Programs, Inc., which will review and approve study protocols involving human subjects. The Office for Human Research Protections of the U.S. Department of Health and Human Services mandates that all research funded by the U.S.-based National Institutes of Health must be approved by an ethical board before receiving any federal funding.

Accuracy and validity are extremely important, especially where measurements are used to identify children with delays (within a population where such cutoffs have been determined). Non-professionals administering tests must be well trained and understand the objectives of testing when using screening and ability tests, as the failure to identify children who are delayed by local standards (false negatives) may result in children not receiving needed services or interventions. On the other hand, wrongly classifying children as delayed (false positives) within the population can cause needless distress and worry for parents (Tluczek et al. 1992; Fyro and Bodegard 1987). Moreover, being labeled as delayed according to local norms of development—even if later repudiated—can follow a child, possibly affecting self-perceptions as well as the way the child is perceived and treated by peers, teachers, and the broader community. Using a screening test out of context with inappropriate cutoffs for a given population is not ethically justified.

It is important to work closely with local clinicians to whom children can be referred. Many ethical boards require that children who screen positive for risk of developmental delay must be referred to a clinician for further evaluation, diagnosis, and treatment, if needed.

It is also important to specify a protocol with action steps addressing what field workers should do when they encounter specific situations. Field worker training should include instructions on what to do if they encounter various cases, such as critical health issues, severe malnutrition, family violence, and child abuse.

6

Adaptation and Standardization of Existing Tools

KEY MESSAGES:

- All developmental capabilities are affected by the opportunities children have to develop their skills, the attitudes and beliefs of their caregivers, and their caregivers' expectations for healthy development.
- The development of culture-free cognitive tests is impossible because all tests (even nonverbal) are inherently biased.
- Steps to successful adaptation include accurate translation, cultural adaptation, pre-testing, pilot testing, and test modification.
- Maintaining reliability and validity of assessments is crucially important.
- Reliability and validity need to be determined within a given cultural context before any assessment can be conducted.

CHILDREN ARE EMBEDDED IN CULTURAL SYSTEMS FROM BIRTH. THEREFORE, almost all developmental capabilities are in some way affected by the opportunities children have to develop their skills, the attitudes and beliefs of their caregivers, and their caregivers' expectations for healthy development. Some cultural practices may have more substantial implications for development than others. Even the emergence of canalized abilities—skills that all normally developing humans eventually acquire—are affected by culturally dictated child-rearing practices, though the timing of when children acquire these skills may ultimately be of little consequence. For example, children who are carried on their mothers' backs tend to walk at different ages than children who spend more time moving independently, but when they ultimately learn to walk appears to have little bearing on their future development. Cultural practices around literacy (such as a belief that boys are more capable of learning to read than girls), however, may strongly affect development through avenues that are not readily apparent to evaluators, and in turn may affect the impact of an intervention on children's outcomes even when the intervention is working properly. Therefore, when selecting measures to use in each country, researchers should carefully document prevailing cultural beliefs and practices to aid in the interpretation of the data and conclusions on the impact of any intervention on children's development.

Test fairness and bias across cultures

Strategies for Cross-Cultural Assessment

Adoption: direct translation of a test to a new context without modification
Adaptation: modification of items, materials, and procedures to fit the local context
Assembly: creation of a new test by assembling items and methods from various existing sources

Source: van de Vijver and Poortinga 2005.

Test fairness relates to the degree to which a measure is equally valid for individuals with different characteristics. Test bias refers to the degree to which a measure may be biased depending on a person's characteristics.

The term "measurement invariance" also expresses the same idea as test fairness, referring to a statistical property of a measurement tool that indicates that the same construct is being measured across groups, such as gender or ethnic groups. Issues to consider include familiarity with the type of materials (writing, numbers, pictures), with the cultural relevance of items (e.g., horses are unfamiliar in Africa), and with the testing situation (e.g., talking to an adult); as well as the importance of test takers' responding quickly. For example, Zambian children have extensive experience making objects from wire, but little experience with drawing. School-aged children asked to reproduce a wire model of an object (the Panga Munthu Test, based on the Goodenough-Harris Draw-a-Person test [Harris 1963]) did so more effectively than when asked to draw a pictorial figure using paper and pencil, illustrating that the use of a familiar medium was an important factor in the assessment of this skill (Ezeilo 1978; Kathuria and Serpell 1998). More recently, a Zambian study reported higher scores on a pattern completion task when it was presented with three-dimensional (rather than two-dimensional) test stimuli (i.e., using objects to complete patterns rather than selecting pictures of objects to complete patterns) (Zuilkowski et al. 2016). Both of these examples illustrate test bias because the original (unadapted) test procedures would underestimate the true ability of Zambian children. Their low scores would reflect their unfamiliarity with the testing materials, rather than poor visuospatial ability, which is the construct the tests were designed to measure.

While there are methodologies for adapting test items, materials, and administrative procedures to make them as fair as possible, cross-cultural researchers acknowledge that the development of culture-free cognitive tests is impossible, as all tests (even non-verbal) are inherently biased, and they recommend that all tests be adapted for use in a different culture (Cole 1999; Greenfield 1997; Rosselli and Ardila 2003). Adaptations of measurements can at best produce a reduction in cultural differences in performance on any test (Anastasi and Urbina 1997). Within these constraints, we recommend considering assessments that have been shown to be reliable or valid among groups of children in various cultural contexts, and to always bear in mind the necessity for careful selection and adaptation or development of assessments to evaluate young children.

No test is "culture-free"; however, many assessment teams choose to use existing tests rather than develop new ones. Three strategies for early childhood development assessment in new contexts have been classified as *adoption, adaptation,* and *assembly* (van de Vijver and Poortinga 2005). These are not delineated categories but represent a spectrum of adaptation procedures. At the adoption end of the spectrum, a test is directly translated to a new language and context without modification. However, test items, materials, and procedures are often inappropriate for children in a new context and must be adapted (Greenfield 1997). Increasing modifications or merging items from multiple sources leads to the assembly of a new test.

Adaptation refers to processes (including translation and item modification) that researchers undertake to reduce systematic bias or error in test scores that can occur when applying a test in a culture other than the one in which it was developed. There are three main types of bias:

- ***Construct bias*** occurs when the instrument does not measure the same underlying construct (e.g., intelligence, social-emotional development) in both cultures. This may be due to differences in the definition of the construct, variability in the measurable behaviors and skills that represent the construct, or inadequate coverage (too few items or domains) to sufficiently assess the construct. Traditional child development scales, such as the Bayley Scales of Infant Development, illustrate this type of bias. Such scales are based on the attainment of behavioral items in a normative sample of children in the country where the test originates. For example, the typical American child learns to squat after learning to crawl and stand. However, the order of attainment of these milestones may differ in other cultures. In Bali, crawling is explicitly discouraged because it is considered animal-like. Balinese children learn to squat as they progress from flexible movement on all-fours to sitting then squatting and standing (Super 1981). While the failure of an American child of a certain age on a crawling item would indicate delayed motor development, the failure of a Balinese child at the same age might not give the same indication.

- ***Method bias*** occurs when the administration or procedures of the test—the use of unfamiliar stimuli (e.g., blocks, puzzles) or unfamiliar response formats (e.g., scales, multiple choice)—differentially affect the scores of groups being tested. The findings that led to the development of the Panga Munthu Test, described above, is an example of this type of bias.
- ***Item bias*** occurs when individual test items do not measure the same way across groups. Sources of item bias include poor translation and culturally inappropriate content (van de Vijver and Hambleton 1996). For example, a vocabulary test in English and Spanish with items matched for meaning (directly translated) yielded different means and standard deviations for two groups of students matched on grade, age, sex, and academic achievement. However, when items with similar frequencies (rather than similar meanings) were used, the two versions yielded similar means and standard deviations in both languages (Tamayo 1987). While translating the items directly resulted in item bias, adapting the items based on language-specific criteria—that is, frequency of use in each respective language—resulted in measures of vocabulary knowledge that were appropriate for each linguistic group.

These biases threaten the validity of tests' capacity to produce "true" scores of children's abilities (Pena 2007). Bias can be reduced, however, by examining how *equivalent* the adapted test is to the original. Four types of equivalencies can be considered (Pena 2007):

- ***Linguistic equivalence,*** or is the translation accurate? This can be accomplished by translation and back-translation (when the translation is translated back into the original language), but does not ensure the appropriateness or validity of the tool in a new context.
- ***Functional equivalence,*** or do the instructions and items have the same functional meaning (i.e., do they get at the same idea and produce the desired response) in the two cultures? When we establish functional equivalence, our goal is to assess the same underlying ability or construct as the original test in a way that is appropriate in the local setting.
- ***Cultural equivalence***, or do the instructions and items have the same relevance or meaning across different cultures?
- ***Metric equivalence***, or do the items have the same level of difficulty? Metric equivalence must be established to compare raw scores across contexts. However, this is the most difficult type of equivalence to establish.

While the International Test Commission (ITC) has published broad guidelines concerning the use and adaptation of psychological and educational tests internationally,[1] no universally recognized minimum standards lay out what test adaptation should entail (Carter et al. 2005; Pena 2007; van Widenfelt et al. 2005; Malda et al. 2008). Several aspects of the adaptation process are repeatedly cited, however, as indispensable to producing a valid adaptation (Carter et al. 2005; van Widenfelt et al. 2005; Malda et al. 2008; Hambleton and Patsula 1998). These include translation; the selection and adaptation of culturally sensitive content; confirmation that test stimuli are culturally relevant; and identification of presentation, administration, and scoring procedures that maximally reduce cultural-based differences in response or performance (Bracken and Barona 1991; Mwamwenda and Mwamwenda 1989). A discussion of these aspects is included below, and Table 6.1 presents the pros and cons of modifying assessment items.

Examples based on some of the authors' experiences adapting the Ages & Stages Questionnaires (Bricker and Squires 1999) in various countries are also provided.

[1] For additional information, please go to: https://www.intestcom.org/files/guideline_test_use.pdf and https://www.intestcom.org/files/guideline_test_adaptation.pdf.

Modification and adaptation guide

STEP 1 *Preparatory work*

- ***Form a panel of local professionals.*** Local professionals who work with young children and their families, including psychologists, social and community health workers, early child education teachers, and doctors, can add important perspective to the development and adaptation of measures for young children (Malda et al. 2008). They can help maximize the cultural appropriateness of the tests to be used, and help contextualize the measurements. Local professionals can also play an essential role in gathering both general and specialized information needed to help adapt tests to local linguistic and cultural norms and rules, for example through focus groups and interviews. The panel should meet periodically to review and provide input on item translations, pilot data, and results of reliability and validity testing.
- ***Conduct preliminary interviews or focus groups.*** Engaging small groups of local key informants (e.g., parents, teachers, and others working with young children) is an ideal way to collect information on the test content and procedures. This process includes using a somewhat structured interview to ask groups of respondents to rephrase the items and responses to ensure they are understood accurately. Respondents should also be asked which response they would select and to explain how they arrived at that answer (Alaimo, Olson, and Frongillo 1999). This interview technique can also be used to get feedback on the test stimuli (e.g., "What does this picture mean to you?") and on various response formats (e.g., multiple choice, scales) to assess their suitability.

STEP 2 *Translation*

- ***Produce an accurate translation that has linguistic and functional equivalence.*** Ideally, the translation process should involve two to four individuals who are bilingual and bicultural. Multiple team members enable identification of problematic translations (van Widenfelt et al. 2005; Solarsh and Alant 2006). While it is generally preferable to keep the translation as close as possible to the original test, word-for-word translations may not retain the original meaning of an instruction or item (van Widenfelt et al. 2005). In such cases, the team needs to develop and test alternative translations to identify the one that best captures the meaning of the original phrase. For example, the piloting of a bilingual language test used with four- to six-year-olds found that instructions to Spanish speakers to "Describe…" a particular object were equivalent to (i.e., got the most similar responses as) the English instructions, "Tell me three things about…" a particular object (Pena 2007). Similarly, a translated and adapted version of the Denver Developmental Screening Test used in Costa Rica altered the instruction "draw a man" to "draw a doll" to produce a response most similar to the original item (Howard and de Salazar 1984). It is also possible that literal translations will result in language that is too complex for the respondents. In populations where literacy levels are low, exact translations may need to be simplified to increase respondents' comprehension of the test (Pena 2007). Translations should strive to be at the most basic level possible.

  Several steps are key to producing an accurate translation (Solarsh and Alant 2006). These include:
  1. Translation and back-translation (by two different individuals) of all test instructions and materials
  2. Review and comparison of back-translated test with original language test
  3. Corrections of the translated version as necessary
  4. Confirmation of the translation by another bilingual adult living in the community
  5. Trial test of the instructions for children in the target community. Often when local variations exist in a language, young children are only aware of the local words. Also, instructions that do not present any difficulty for adults may still be misunderstood by children.

The team should also check for poor or incomplete translations that may occur when a translator is unfamiliar with the underlying concepts of the items or tests. For example, when the (English) ASQ item, "When playing with sounds, does your baby make low-pitched noises?" was translated into Mexican-Spanish, it became, "When you play with your baby, does s/he make low-pitched noises?" The translation changed the meaning of the original item and had to be adjusted.

STEP 3 *Review for appropriateness*

- ***Involve a local panel.*** Invite the same local panel of researchers, psychologists, pediatricians, and teachers formed in Step 1 to review the tests for cultural appropriateness and suggest modifications.
- ***Consider the content for functional, cultural, and metric equivalence.*** The test content may need to be altered to ensure items elicit behaviors or responses similarly across cultures (Pena 2007). To accomplish this objective, the ideas or situations expressed in the item should be relevant, easily recognized, and readily understood in the local context, and they should also match the difficulty level of the original item (Solarsh and Alant 2006). For example, in adapting the Ages & Stages Questionnaires for use in Mexico, an item about whether a child asks a caregiver to wind up a toy was replaced with whether a child asks a caregiver to open something (such as a bottle) or peel something (piece of fruit). In addition, test stimuli such as balls, blocks, or dolls may need to be replaced with objects that are found locally, and pictures and drawings should depict people, houses, trees, animals, and other objects that are familiar to the local setting (Carter et al. 2005).

  Where child development tests require caregiver responses, consideration should be given to cultural norms that may affect how adults understand and answer questions. Where formal education is not universal, caregivers may lack experience reflecting on their thoughts or making relative comparisons. In cultures where thoughts are not distinguished from what is "real" and observed, caregivers may not be able to respond to items asking them to imagine hypothetical situations or make speculations (Greenfield 1997).

  Response sets may also need to be changed to make certain that the response choices are unambiguous and represent the desired complexity. For example, multiple choice tests should include possible responses that are similar in difficulty to the originals, ensuring that there is one clearly correct answer but that it is not too obviously correct. Gradient scales using numbers or phrases may need to be substituted with illustrations or objects that represent the response options, or with hand gestures that indicate more or less.

  In some cases, suitable cultural equivalence may not exist for an item for the age being tested. In our experience adapting the Ages & Stages Questionnaires, we found children do not frequently use forks in Peru, Indonesia, or Tanzania. As a previous item asked about use of a spoon, no suitable substitute items could be found, so this item was dropped from the test. This does not mean, however, that shortening tests at will is appropriate. Some assessment teams may be tempted to abbreviate standardized tests during adaptation to better suit the project demands (e.g., large samples; limited time and resources). Snow et al. (2008) warn against this, as shortening a measure may threaten its reliability, validity, and equivalence with the original test (Snow and Van Hemel 2008).
- ***Consider the materials.*** Before modifying or adding assessment items, it will be important to assess the pros and cons of those actions. Table 6.1 discusses the aspects to be taken into account. Many commonly used tests have pictures or figurines, objects such as bells or staircases, or materials such as brightly colored plastics, which are unfamiliar to many children living in low- or middle-income countries, especially those in rural areas. These items often need to be replaced with locally produced materials. Similarly, the text or pictures may describe practices (such as sitting around a table having a meal together) that are not part of the local culture and that will need to be replaced. Many of the pictures in the Peabody Picture Vocabulary Test or Kaufman Assessment Battery for Children (KABC-II), for instance, depict objects that are not available in rural communities in low- and middle-income countries. Such necessary adaptations may be costly in terms of time or money and may constitute constraints in selecting instruments.
- ***Consider the administration procedures for functional and cultural equivalence.*** Tests standardized in the United States or United Kingdom typically identify the range of items to be used with children of a particular age. These age-specific item sets reflect how items work in the country in which they were developed and may not be appropriate in other countries. For optimal test item development, teams can explore which set of items most accurately assesses development at particular ages by piloting a larger range of items (i.e., from younger and older item sets) in a representative sample. Reordering of individual items may also prove necessary, based on their performance in the piloting. For example, in Indonesia, a child's use of the pronouns "I" and "me" occurs at a later age than when reflected in the questions used in the ASQ test. Further testing would have to be done to determine at what child ages the mother should be asked about the child's use of these pronouns. In addition, many adults and children will be

unfamiliar with test taking, and therefore the very situation of being asked questions and responding to a stranger will be foreign, which could interfere with test performance. In addition, women and children may be very shy.

Test administrators can address these issues from several different angles:

1. ***Tester.*** The tester should understand the test materials well, be from the community, and be fluent in the language spoken by the respondent. An open, engaging, non-judgmental approach toward the testing will be less likely to intimidate the respondent, especially young children. It may be important to alter the pace according to the child and culture. Special training is needed to make sure that assessors can encourage a child to try to answer difficult questions. Training should incorporate guidance on how to deal with children's shyness, particularly in cultures where children are not encouraged to speak to unfamiliar adults or to voice opinions in the presence of adults. Similarly, the manuals associated with assessment tests should include standardized procedures for working with difficult-to-test children.
2. ***Test environment.*** There are two issues to consider. In the absence of a standardized setting, testers should attempt to simulate ideal testing conditions (a fairly quiet place that provides some privacy to respondents; a place with sufficient light and space to complete all items) as best possible across all test administrations. The second issue involves creating a friendly, non-threatening atmosphere. This begins with ensuring that a caregiver or other familiar adult is present with the child throughout the testing. Other steps could involve changes to the procedures by the tester, including sitting next to, and at the same level as, the child (Pena 2007); not asking questions directly to the child if culturally inappropriate (Snow and Van Hemel 2008); spending additional time chatting with the child or household members to establish rapport; or providing toys or materials for the child to play with before beginning the test.
3. ***Test procedures.*** The instructions or procedures may need to be altered to elicit the best performance possible from the respondent. These changes should be discussed with the local team, and may include: allowing extra time for a child to become sensitized to test stimuli prior to administering an item using the stimuli; allowing additional practice trials for items that contain unfamiliar stimuli or activities, such as engaging in grouping or sorting tasks or working on puzzles; allowing extra time than recommended in the original test for completion of timed tasks (understanding that the importance of time should be explored as it may differ cross-culturally); and adjusting the types and frequency of praise, encouragement, feedback, or probes used throughout the testing. It is important to explore which types of praise and encouragement (words or gestures, or both) work best with the target child. The tester should use praise at the beginning of each test or section, tapering off to active, interested attention. If the tester offers verbal praise after each response, children notice when he or she does not praise. Children should always be praised for effort rather than success on an item. The effectiveness of probes such as "Tell me more" should also be explored with both children and adults to ensure their use has the desired effect of enhancing test performance (Pena 2007). Additional, clarifying instructions may also be required.

STEP 4 *Pilot testing*

- ***Conduct a pilot test.*** Pilot testing in a representative sample of the population where the test will be used can help researchers understand the way that the measurement functions. A debriefing with respondents (adults) and data collectors after the pilot testing can provide additional information on aspects of the test procedures.
- ***Analyze pilot data.*** The psychologists involved with the adaptation should examine basic psychometric properties of the test. To do this, analyses should include the following steps:
  - ➔ *Calculate the percentage of scores that are missing by item.* If any item has a high percentage missing, a problem may exist with that item.
  - ➔ *Ensure the items show variability* (e.g., not all children got an item correct or wrong). If all children get an item correct, it may be useful to modify it so that it is more difficult. Conversely, if all children get an item wrong, it may be useful to modify it so that it is easier. However, whether to make these changes depends on the goal of the assessment. If the goal is for the adapted test to show as much

TABLE 6.1 Pros and Cons of Changing or Adding Assessment Items

	PROS	CONS
Eliminating an inappropriate item	Avoids asking parent or child an irrelevant or confusing question.	The total number of items is not equivalent to the original scale; therefore, (1) the raw scores on the adapted test are not equivalent to those for the original test, and (2) the standard norm scores cannot be applied.
Maintaining an inappropriate item	Maintains the same total number of items as the original scale.	Parents and children may be confused by the question. This could affect their attitude toward the interview or test, and their responses on other questions or items. The raw scores on the adapted test are not equivalent to those for the original test because the item has a different meaning in the two contexts.
Modifying an inappropriate item	Avoids asking parent or child irrelevant or confusing question. Maintains the same total number of items as the original scale. If the modified item is equivalent in difficulty to the original item, then the scores on the adapted test are equivalent to those for the original test.	The modified item might not be equivalent in difficulty to the original item, which would mean the scores on the adapted test are not equivalent to those for the original test.
Adding a locally appropriate item	Adds content that is meaningful for assessing child development in the local context. Increases variance in scores and therefore, the possibility of detecting intervention effects between groups in the same local context.	The total number of items is not equivalent to the original scale, therefore the raw scores on the adapted test are not equivalent to those on the original test, and norms cannot be applied.

variance as possible—for example, to maximize the possibility of showing group differences between intervention groups—then these changes should be made. If the goal is to keep the adapted test as similar as possible to the original test, then these changes should not be made.

- ➔ *Examine whether expected, age-related differences are evident.* If scores show expected correlations with age, then it is reasonable to assume that higher scores are indicative of developmental advance.
- ➔ *Check correlations with other variables expected to be associated with child development*, for example, maternal education, home environment, and height-for-age z-score.
- ➔ *Determine the internal consistency of the measure* (i.e., how well the items work individually and together as a test), with a test such as Cronbach's alpha (Cronbach 1951). If Cronbach's alpha is low (generally a rule of thumb is < 0.7), then examine the item-total correlations to see which items are not correlated with the others. These may be problematic items.
- ➔ *Determine the dimensionality of the measure.* For many tools, the items are divided into subscales, which assess different sub-domains within a domain. Exploratory and confirmatory factor analysis can be used to determine whether the translated and adapted tool conforms to the original dimensional structure. This includes whether it is measuring as many separate domains as the original tool or whether it is measuring fewer or more, as well as whether the same items load on the same subscales.

- ***Check the item characteristic curves following item response theory (IRT).*** Most of the principles for test evaluation presented in this Toolkit are based on classical test theory, which posits that a person's score on a test is determined by a combination of his or her true ability and some level of measurement error. Item response theory is a more recent and robust theory for evaluating the properties of psychological tests. An item characteristic curve can be generated for each item, showing the probability of a correct response (on the y-axis) across individuals with different ability levels (on the x-axis). An individual's ability level is calculated as a latent variable derived from all items in the test. A good item should show that the probability of passing the item increases with ability level, demonstrated by a good fit of the observed data to the item characteristic curve. Evaluation of items using IRT requires more complex statistical models, as compared to classical test theory, and we recommend consulting a statistician with expertise in this area to conduct these analyses.

- ***Allow time and resources for iterative adaptation and testing of the tool.*** The adaptation of tests is likely to require multiple "rounds" for each step outlined above to ensure the test is valid. For example, the test may then require several rounds of piloting, adaptation, and re-piloting to reach the best version. Ideally, researchers should allow at least three months for completing this process from start to finish. The time needed will vary by number of tests being adapted, the access to samples who can be used for pilot testing and who are similar to those who will be examined, and the availability of adaptation team members, among other aspects.

STEP 5 *Documenting changes*

All changes that are made to items, materials, and procedures should be documented. This includes modifications suggested by panel members and informed by pilot data.

Example 1. A table to keep track of item changes made over several iterations of piloting and review with the following information:

- Original item in source language
- Initial translation to target language
- Result of panel review (e.g., exclude, modify)
- Revised translation
- Pilot results (e.g., percent missing: out of all children tested in the pilot test, percentage of children for whom the item was missing; percent passed: out of all children tested in the pilot test, the percentage who passed the item)
- Final item

Example 2: A table to summarize multiple types of changes that an item could undergo might include the following columns:

- Item
- Changes made to item content
- Changes made to item materials (e.g., toys, pictures)
- Changes made to item scoring
- Changes made to item administration

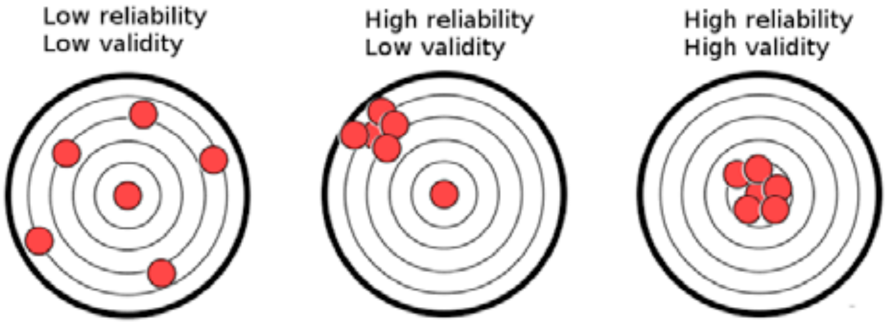
Evaluation of test scores: reliability, validity, and norms

Psychometrics is the area of psychology concerned with evaluating the design and effectiveness of measures to assess psychological characteristics (or domains, such as language or cognitive development). Psychometric analyses are primarily used to determine the *reliability* and *validity* of an assessment. *Reliability* refers to how consistently a measure produces similar results for a child or group of children with repeated measurements over a short period of time. This is based on the assumption that individuals (or groups of individuals) show some stability in how they exhibit the behaviors under evaluation. However, there is typically some variation in scores on successive tests. The reliability of tests can be increased by ensuring that tests are administered uniformly and under conditions where individuals have the capacity to produce their "best" performance. *Validity* refers to the degree to which a measure accurately assesses behaviors or abilities that reflect the underlying concept being tested. For example, do the items included in a language test accurately "tap" a child's capacity to produce a certain number of words or to understand what is being said to him or her at a given age (Cueto et al. 2009)?

Figure 6.1 illustrates the difference between reliability and validity. When a tool is reliable and valid, the score on the test, represented by the red dots, reflects the true ability of an individual, represented by the center of the target, every time that the tool is administered to that individual. This situation is represented by the circle on the right. When reliability is poor, an individual receives varying scores from one time point to another due to measurement error, represented by the circle on the left. It is also possible that a tool is highly reliable, but it does not reflect an individual's true ability, or it does not reflect the construct the tool was intended to measure, shown by the circle in the center. While it is possible for a test to be reliable but not valid, the converse is not true—it is not possible for a test to be valid but not reliable (Rossiter 2011). If every time you stepped onto a scale, the scale displayed a different weight, that scale could not be accurately measuring your weight. However, if the scale was not calibrated properly, it could display the same weight every time even if this was not your actual weight. In this case, the scale would have high reliability but low validity, represented by the central target in Figure 6.1. In practice, reliability for behavioral assessments can be difficult to determine because a person's true ability may change from one time point to another. This is like aiming at a moving target. In this case, instability in scores is due to a change in true ability rather than measurement error.

The majority of published tests developed in developed countries (e.g., the United States, the United Kingdom, European Union countries) have undergone rigorous examination to ensure the assessments are both reliable and valid in the populations in which they were developed; however, reliability and validity need to be determined within each cultural context. Sidebar 6.1 discusses how to assess reliability of a test while Sidebars 6.2 and 6.3 present how to assess validity of ability tests and screening tests, respectively.

FIGURE 6.1 Reliability and Validity



Many tests, especially tests of cognitive development, have been "normed," meaning that test producers have collected data from a large representative sample of a population, usually from developed countries, to draw a normal distribution of scores. When transferring tests from one context to another, an important issue is whether it is valid to use standardized norms from one country when calculating scores of children in a different country. Standard norming guidelines require creation of a nationally representative sample that reflects the characteristics of an entire country (e.g., ethnicities, levels of education, primary languages spoken at home, poverty rates, geographic regions, children with disabilities, and giftedness) (Glascoe et al. 2013). When the test score of a child in the United States is calculated based on U.S. norms, the score can be interpreted in relation to the average score of his or her peers of a similar age. Norms are often standardized to a mean of 100 and a standard deviation (SD) of 15. Therefore, a child who scored 100 performed at an average score for his or her age, while a child who scored 85 performed about one SD below other children his or her age. When the test score of a child in another country is calculated based on U.S. norms, this interpretation is not valid: It is not accurate to say that a child in a non-U.S. population who scores two SD below the mean is "delayed" using a set of norms developed in the United States. Even high-income countries, such as the United States, Canada, the United Kingdom, and Australia, usually have their own separate norms.

Standardized scores on these tests can be used to calculate differences between groups, however. For example, when comparing an intervention and control group in a research study, calculating the age-standardized norm score for each child can be a useful way to create a standardized scoring system. Another way to create such standardized scores is to calculate z-scores on the research sample, if the sample includes sufficient children per age group. Standard norming guidelines state that a sufficient sample size is 75–200 per age group. Age bands (or intervals) are expected to be smaller in the first year of age (one-month periods)

SIDEBAR 6.1 How to Evaluate Reliability of Test-Retest

- Assess the same group of children twice, usually with a test-retest interval of about one to two weeks.
- For items rated on a scale with fewer than five levels (categorical ratings), use Cohen's kappa coefficient or a weighted kappa statistic to assess test-retest reliability.
- For scores that can be considered approximately continuous (for example, the response is on a rating scale with five or more ordinal categories), reliability can be assessed using the intraclass correlation coefficient (ICC) or Pearson's r. ICC values from 0.40 to 0.75 are considered as fair to good. Excellent values are greater than 0.75, although these, as with any cutoffs, are guidelines only. For Pearson's r, good reliability corresponds to values of 0.70 or greater.

SIDEBAR 6.2 How to Evaluate Validity of Ability Tests

- **Criterion validity.** Compare the test score to a gold standard measure. Usually this is not possible for ECD assessments in low- and middle-income countries because a gold standard measure does not exist.
- **Convergent validity.** Check whether test scores are associated with factors that are expected to be related to them. For example, language and motor development would be expected to be associated with each other and with factors such as maternal education, home stimulation, and linear growth. The purpose of this evaluation is to generate confidence that the tool is performing as expected. However, if an expected correlation is not found, there may be a plausible reason. For example, in Burkina Faso, motor scores on the Developmental Milestones Checklist-II were not associated with maternal education. The authors speculated that the weak association with maternal education may be explained by the generally low levels of education in this sample. Of the 1,123 mothers in this study, only 155 (14 percent) had been to school and only 19 (2 percent) had attended school beyond grade six. A few years of elementary education may not be enough for the emergence of an association between maternal education and children's developmental attainment (Prado et al. 2014).
- **Discriminant validity.** Check whether test scores are not associated with factors not expected to be related to them. For example, behavior problem scores on the Brief Infant-Toddler Social and Emotional Assessment (BITSEA) were not associated with age in the norm sample (Briggs-Gowan and Carter 2002). An adaptation of this tool in Indonesia also showed no correlation with age (Prado et al. 2010).

SIDEBAR 6.3 How to Evaluate Validity of Screening Tests

- **Establishing a cutoff.** Cutoffs used in one population to classify children as "delayed" or "normal" cannot be applied to another population, though screening tests can be useful for examining developmental differences in raw scores between groups of children. However, a difficulty with screening tests is that the raw scores are not usually normally distributed, since they are not designed to capture variance between typically developing children; because the data cannot be analyzed as a continuous score, a categorical score must somehow be created or a non-parametric statistical procedure used.
- **Sensitivity and specificity.** The validity of screening tests is usually defined by the sensitivity and specificity of the tool to identify children diagnosed with developmental delay or other conditions. Evaluating sensitivity and specificity requires testing a group of children, some of whom have been diagnosed by a clinician with developmental delay and some of whom have been judged by a clinician to be typically developing. A true positive is a child who is delayed and screens positive for delay. A false negative is a child who is delayed and screens negative for delay. A false positive is a child who is not delayed and screens positive for delay. A true negative is a child who is not delayed and screens negative for delay. The *sensitivity* of a tool is the percentage of true positives out of all children who are delayed. The *specificity* is the percentage of true negatives out of all children who are not delayed.

and larger at later ages (two- to three-month periods for toddlers, and six-month or one-year periods for school children). (Glascoe et al. 2013). As an alternative to calculating standard norm scores, groups can be compared on the basis of raw scores, adjusting for child age.

Another use for child development scores is to describe a population, if researchers are using representative samples and have a clear priority placed on describing the population, as opposed to explaining differences between groups. In the context of low- and middle-income countries, where half of all children may not be fulfilling their developmental potential (Black et al. 2016), it may be useful to distinguish between descriptive norms and prescriptive norms. Descriptive norms reflect the scores of a representative sample of the country's population, as described above. Prescriptive, or "reference" norms, reflect the developmental achievement children would be able to attain in that country in the absence of any environmental constraints on development (Serpell 2015). This concept is similar to the rationale for the development of the WHO growth standards (WHO Multicentre Growth Reference Study Group 2006). Developmental delay is often determined by a cutoff score based on the standardized norm sample, for example, a score below 2 standard deviations below the mean in a given age group. Using descriptive versus prescriptive norms would result in a different definition of delay. Even using prescriptive norms, the definition of delay might not be comparable between one country and another.

However, for most tests, neither descriptive norms nor reference norms currently exist for children in many low- and middle-income countries. Whether it is possible to specify universal, prescriptive norms that apply across countries is highly controversial; several initiatives are currently underway to explore this possibility.

Table 6.2 provides a summary of the aspects to consider when adapting a test and analyzing pilot data.

TABLE 6.2 Checklist for Test Adaptation

✓ Form a panel of local professionals who meet periodically throughout the test adaptation process to review the test materials, translations, pilot data, and results of reliability and validity testing.

✓ If necessary, conduct preliminary interviews or focus groups.

✓ Produce an accurate translation.

✓ Review the content, materials, and administration procedures with the panel.

✓ Conduct an iterative series of pilot tests, making modifications based on the results of each round, and then piloting the modified tests.

✓ Analyze the pilot data to check:

- The percentage of missing item scores
- Item variability
- Expected age-related associations
- Associations with other variables that are expected to be related
- Internal consistency
- Dimensionality (factor structure)

✓ Keep track of changes.

✓ Evaluate test-retest reliability.

✓ If appropriate, evaluate sensitivity and specificity.

Ensuring quality in test administration

To achieve the highest possible quality measurement of outcomes, the research team should provide adequate training to testers and supervisors. Sidebar 6.4 summarizes the various strategies for training testers, from initial training to practice sessions to recertification. Trainees should have completed schooling in related disciplines (social sciences, psychology, child development, education) or have relevant experience (interviewing, community work). If trainees have no previous experience in child assessment or community work, they may require a more extended period of training and practice. It is essential that all testers receive the same training by the psychologists and team on all aspects of the testing situation: approaching families and establishing rapport, introducing the test to families, giving instructions, administering items and recording responses, offering praise and encouragement, using probes during the administration, and providing feedback on test performance or results.

Local psychologists can add an important perspective to the adaptation and training process when developing a measurement tool. In addition to their inputs during adaptation and training, they may be able to provide continued follow-up training as needed, as well as supervision. Universities and local non-governmental organizations or government agencies can be good sources for finding psychology-trained personnel to assist with adaptation and supervision. That said, there can be international collaborators who have gained sufficient contextual knowledge and cultural insights, perhaps through working with local collaborators.

Inter-rater reliability

Trainees should also undergo some standardization exercises. For the exercises described below, a reference or optimal interviewer should be established. This person should be trained and efficient with the questionnaire and fluent in the local language. The goals of standardization are to compare each of the trainee interviewers with this reference standard to ensure accuracy and reliability, with the process consisting of two parts.

Inter-rater reliability is how much scores among raters agree. This type of reliability is important to ensure that all personnel are administering the assessments in the same way and to subsequently reduce measurement error or bias due to a particular assessor. To test inter-rater reliability, all interviewers should be present at the same session with the same child or interviewee. The trainees, who will follow along silently, should record the responses on their own forms, based on their observations of the assessment (see Figure 6.2).

SIDEBAR 6.4 Tips for Training

- **Train more testers than you need.** Individuals learn at different rates and some naturally interact better with children and their caregivers than others. Train more testers or interviewers than needed for the project, then hire the top performers.
- **Train testers on general strategies for interacting with children, as well as specific testing procedures.** For example, strategies on how to deal with a child who refuses to do the activity include:
  - Ask the mother to demonstrate the task and to encourage the child to do it.
  - Involve other children present or an older sibling. Have the other child do the task and see if the child will do it along with the other child.
  - Give the child a drink and a break, or let the child rest or play for a few minutes.
  - Promise to give the child a reward after completing the task.
- **Read the entire test manual aloud section by section.** After reading each short section, use various strategies to encourage the trainees to engage with the material:
  - Ask a few comprehension questions.
  - Have someone explain the material in his or her own words.
  - Show a video.
  - Demonstrate ideal interactions.
  - Expand or explain further.
- **Role play with one person playing the child or the mother.** This could be done with one group role playing while others watch and give feedback. Or the trainees could divide into groups and everyone role plays simultaneously.
- **Practice with community children.** For interviews, each interviewer should practice with at least five caregivers. For direct assessments, each tester should practice with at least 10 children before inter-rater reliability is evaluated.
  - Trainers should observe and give immediate feedback. They should interrupt if trainees are doing something wrong and ask them to correct it right away. It's not just any practice that makes perfect, but perfect practice makes perfect.
  - Take videos of practice sessions.
- **Review videos of practice sessions.**
  - Pause after each item to discuss what the tester did well and what could be improved.
  - Practice scoring. Each person has a form and practices scoring independently, or the trainees can answer out loud and discuss together.
  - Review the forms from the practice sessions and give feedback on scoring and any aspect of the form that was not completed correctly.
- **Review key interview tips with field workers.**
  - Stress the importance of building in time for the child and caregiver to become comfortable with the assessment situation.
  - Go over how the field worker can build rapport with the child and caregiver so they feel comfortable.
  - Recommend stopping an assessment if the child appears tired or disinterested in the activity and rescheduling for a later time or day.
  - Require testers to revisit the child's household as many times as necessary and feasible to obtain a response from the child.
- **Give knowledge-based evaluations.** Developmental assessment involves learning content related to the test or interview, such as administration and scoring procedures. Throughout training, give multiple choice or short-answer tests to evaluate whether each trainee has learned the content required to administer the assessment correctly. Trainees should achieve high scores before being certified to administer tests.
- **Give practice-based evaluations.** Developmental assessment also involves learning skills. The content that is learned must be put into practice. Create a checklist of everything that must be done to administer the test correctly. One way to do this is to translate the manual into a point-by-point checklist. For long tests, this could result in a checklist with hundreds of items. Trainers should observe and score whether the trainee completed each instruction and action correctly. Give feedback on anything that was not done correctly. Trainees should achieve high scores before being certified to administer tests.
- **Recertify every three to six months.** Over time, testers can forget the correct procedures and form poor habits. Evaluations, retraining, and recertification should be conducted every few months.

Source: Shankar et al. 2009.

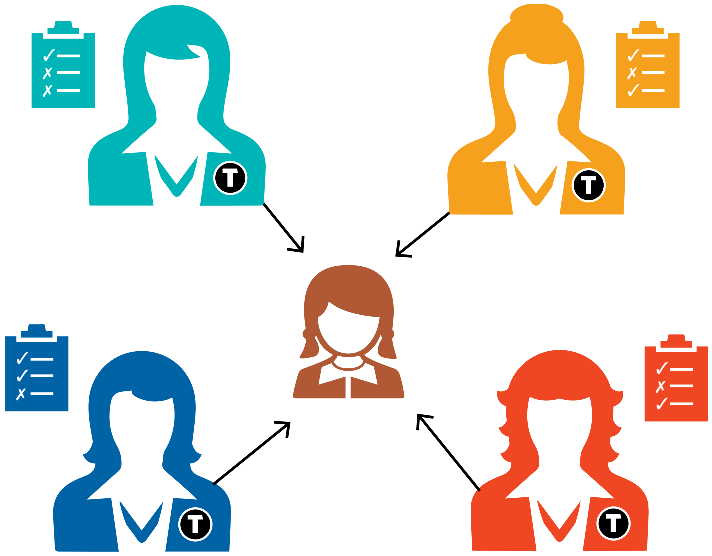
Note: These principles are based partly on the "head, heart, and hands" system for evaluating community workers developed in the Supplementation with Multiple Micronutrients Intervention Trial (SUMMIT) in Indonesia.

Cohen's kappa coefficient (κ) is a statistic which is used to measure inter-rater reliability by taking into account the observed agreement between raters (P_0), as well as the expected agreement between raters due to chance (P_e). To compute the kappa statistic, responses to each item should be compared, and Formula 6.1 should be applied. Each trainee's responses should be compared with every other to ensure a kappa statistic of at least 0.80, though as for all such cutoffs, this is a guideline only.

Formula 6.1 Cohen's Kappa Coefficient

$$\kappa = \frac{(P_0-P_e)}{(1-P_e)}$$

FIGURE 6.2 Assessing Inter-Rater Reliability



Inter-rater reliability is how much scores among raters or interviewers agree. To test inter-rater reliability, all trainee interviewers (T) should be present at the same session with the same child or interviewee. The trainee interviewers follow along silently and record the responses on their own forms, based on their observations of the assessment. Then the degree to which their responses agree can be measured by Cohen's kappa coefficient.

For example, let's say that the trainees assessed a child with a 20-item measure of language development. Each item can be scored as 1 (Pass) or 0 (Fail). Record in a spreadsheet column or on a piece of paper each trainee's response for item 1, item 2, and so on, through item 20. Out of the 20 items, count the number of times in which each pair of trainee's responses agrees, and divide by the total number of items to obtain the P_0. To calculate the P_e, determine the probability that, given the responses the pair of trainees gave, they would have agreed by chance. If both trainees scored "Pass" 50 percent of the time, the probability of both scoring "Pass" would be 0.50 x 0.50 = 0.25. The probability of both scoring "Fail" would also be 0.25. As such, the chance probability of agreement would be 0.25 + 0.25 = 0.50. Use the P_0 and P_e to calculate the κ. For further discussion of common issues and appropriate practices for calculating inter-rater reliability, see Hallgren (2012).

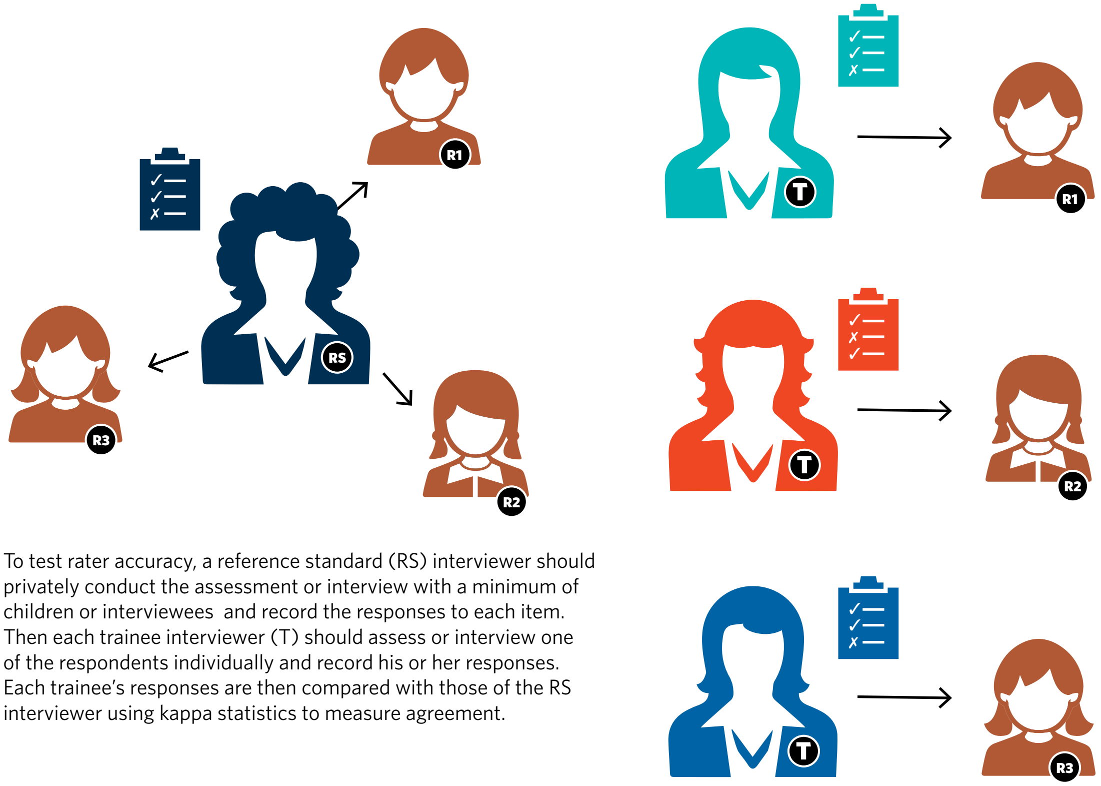
Estimating the number of participants needed to establish inter-rater reliability requires a power analysis for Cohen's kappa (Cantor 1996; Gwet 2014). An alternative method for assessing inter-rater agreement has been proposed by Bland and Altman (1986).

Rater accuracy

In addition to how much trainees agree with one another, we are also interested in ensuring that each rater is accurate in his or her assessments for a given measurement tool or for a given domain. To do this, a reference standard (RS) interviewer should conduct the assessment or interview with a minimum number of children or interviewees privately, and record the responses to each item (see Figure 6.3). Subsequently, each of the trainees should assess or interview one of the three respondents (R1, R2, and R3 in Figure 6.3) individually and record his or her responses. Each trainee's responses are then compared with those of the reference standard interviewer, and kappa statistics for agreement are computed as described above. A correlation of 0.70 or above is desirable, though as for all such cutoffs, this is a guideline only.

Many projects involve an extended period of data collection over several months or years. Data collectors should be reevaluated periodically, for example every three or six months, to ensure continued adherence to the correct test administration and scoring procedures. Retraining should be conducted, focusing on any items or testers that show poor performance. Table 6.3 provides a checklist for ensuring quality of test implementation.

FIGURE 6.3 Testing Accuracy



To test rater accuracy, a reference standard (RS) interviewer should privately conduct the assessment or interview with a minimum of children or interviewees and record the responses to each item. Then each trainee interviewer (T) should assess or interview one of the respondents individually and record his or her responses. Each trainee's responses are then compared with those of the RS interviewer using kappa statistics to measure agreement.

TABLE 6.3 Checklist for Ensuring Quality of Test Implementation

✓ Involve local psychologists, clinicians, early education specialists, public health professionals, and others, as relevant.

✓ Train more testers than you need and hire the top performers.

✓ Read the entire manual out loud section by section.

✓ Role play with one person playing the child or mother.

✓ Practice with community children while trainers give real-time coaching and feedback.

✓ Review videos of practice sessions and provide feedback.

✓ Review the forms from the practice sessions and provide feedback.

✓ Require testers to pass knowledge-based evaluations (e.g., multiple choice tests) before being certified to administer tests.

✓ Require testers to pass practice-based evaluations before being certified to administer tests.

✓ Require testers to achieve inter-rater agreement above 80 or 90 percent.

✓ Conduct retraining for testers and items that show low agreement.

✓ If data collection continues over a long period, reevaluate testers periodically, for example ever three or six months, on knowledge and practice-based evaluations and inter-rater agreement.

7

Creating New Assessments

KEY MESSAGES:

- Developing a new assessment involves a long list of procedures for modification and adaptation, as well as a detailed examination of how the new assessment functions.
- Many new assessments have been created and hold great potential for use in low- and middle-income contexts.
- Creating a new assessment requires a great deal of time, energy, and resources (financial and human), and thus is generally not recommended.
- A developmental psychologist or someone with equivalent training should lead the process of creating a new assessment.

RATHER THAN ADAPTING AN EXISTING TEST, RESEARCH TEAMS OCCASIONALLY elect to create their own tests. This may be done when previously adapted measures are not available, or when copyrighted tests are too expensive. However, the process of developing a new test can also be expensive and time-consuming, and the cost of using a copyrighted test could ultimately prove to be cheaper.

The great advantage of creating a new test is that it can be tailored to the local context. Often, this process involves compiling items from existing tests that include items known or believed to validly measure concepts in the population under study (Gladstone et al. 2008; Stoltzfus et al. 2001; Holding et al. 2004). The great disadvantage of creating local tests is that a large amount of time and resources is required.

Some researchers may be interested in identifying and measuring locally defined concepts of child competence (Lansdown et al. 1996). Before undertaking the development of such tests, researchers should have a clear idea of how this measure would provide information that would discriminate between groups of children under study (i.e., treatment versus control) and how these measures would relate to intervention goals, such as school achievement or adult productivity.

The development of any new test requires employing the procedures outlined in Chapter 5 for modifying and adapting tests, as well as a more detailed examination of how the new test works. Ultimately, scores on the new instrument should measure the domains similarly to other assessments (if possible) (Hambleton and Patsula 1998), or correlate with factors (e.g., physical growth, caregiving practices, maternal education, socio-economic status) known to be predictive of outcomes being measured. Several textbooks provide comprehensive guides on test construction (Schweizer and DiStefano 2016; Franzen 2011).

Recommendations for creating a new test

Sidebar 7.1 lists best practices for creating a new test. There are many examples of new tests that have been developed for a particular cultural framework. Sidebar 7.2 provides some examples of new, country-specific tests. In each case, the tests were developed to be appropriate for the cultural context or specific assessment need.

One elegant example of this process is the study undertaken by the World Health Organization in the 1990s to produce culturally relevant developmental checklists (for screening) for use in the home, community, or in primary care centers (Lansdown et al. 1995). The tests were developed in several phases in China, India, and Thailand. A total of 28,115 children aged 0–6 years were tested during the process of creating and selecting the motor and mental milestones. While the countries maintained longer versions, each ultimately selected 13–19 key milestones for use in health clinics and community centers. The inclusion of overlapping behaviors enabled the authors to create norms (median age at attainment) for comparison within and across sample sites. Examples include “sits” (range 5.4 months in Thailand to 6.9 months in rural China); “uses cup” (9.5 months in Thailand to 35.4 months in urban India); and “says one word” (9.7 months in urban India to 15.0 months in rural India). Each country also included culture-specific items, such as “use of chopsticks with small foods” (31–33 months in China), “ties sticks together with string” (45.7 months in Thailand), and “carries wooden block on head for 5 steps” (45–47 months in India).

Another more recent example is the CREDI, the Caregiver-Reported Early Development Index, which is a population-level measure designed to capture the early developmental skills of children under age three years.[1] The CREDI uses caregiver reports to assess children's motor, cognitive, language, social-emotional, and mental health development. The CREDI tool was tested and validated in 17 countries to ensure reliability and validity, as well as metric invariance across country income status (high-, middle-, and low-income). However, metric invariance has not been tested within country (McCoy et al. 2017). Because it was designed to be culturally neutral, CREDI scores can be compared across context, and local adaptation work should be minimal. The CREDI comes in two forms. The Short Form provides a single, continuous score representing children's overall development based on a set of 20 age-specific questions. The CREDI Short Form is intended for use in large-scale, population-level surveys, as well as large-scale monitoring efforts

[1] See: https://sites.sph.harvard.edu/credi/.

SIDEBAR 7.1 Recommendations for Creating a New Test

- **Involvement of an inter-disciplinary research team.** The team should include bilingual psychologists or other knowledgeable professionals who are able to ensure a psychometrically sound process is employed in the development of the test and (if different) local psychologists who are able to provide insight into the constructs being defined and instrumentalized.
- **Adequately representative sample for testing items and test cohesion.** New assessments should be piloted with a sample similar in age, sex, ethnicity, and socio-economic status as the target population to make the pilot as relevant as possible.
- **Engagement of a statistician with expertise in psychometric evaluation.**
- **Detailed analyses of the instrument's psychometric properties,** so a thorough examination of how the measure “works” can be made. Issues to consider include:
    - Does the instrument adequately cover the entire domain or concept intended to be measured? If a test is measuring language, for example, do items address both receptive and expressive language abilities?
    - Are the items ordered to reflect age-related progression in the domain under study?
    - Does internal reliability demonstrate that the items measure the same construct?
    - Is the test reliable, or do the items assess the concept the same way over time (test-retest scores are highly correlated)?
    - Do the items measure the same way in different groups (e.g., poor versus less poor) of children? (For example, there should not be items on the test that only children of higher socio-economic status or from a rural region can pass.)
    - Do scores on the scale vary meaningfully by subgroups of children in the sample? If it is of interest to create a national tool, is the pilot sample nationally representative and of sufficient number to detect developmental differences?
    - Is there evidence for item discrimination or difficulty through Item Response Theory analyses (see Chapter 6)?
- **Development of norms or standards** that represent typical development in the population under study so that recommendations for services or meaningful interventions can be made. This can be much more demanding in time, effort, and resources and required expertise, and the resulting measure may not be comparable with other measures of similar constructs.

SIDEBAR 7.2 Examples of New Country-Specific Tests by Region

Africa

- **The Kilifi Developmental Inventory (KDI)** (Abubakar et al. 2007; Abubakar et al. 2008; Abubakar et al. 2008) was developed to assess psychomotor development in a resource-limited setting. The KDI is a continuous measure and was originally designed to assess effects of malaria on functioning.

- **The Grover–Counter Scale of Cognitive Development** (Sebate 2000) was developed in South Africa to assess the level of cognitive functioning of children aged 3–10 years with impaired verbal skills, whether receptive, expressive, or both. It is language-free and based on Piagetian concepts of development. This test was designed to facilitate diagnosis of, and treatment for, mentally handicapped children, but may also be used in populations where many languages are represented or where children are very shy.[a]

- **The Malawi Developmental Assessment Tool (MDAT)** was created in Malawi by combining items from the Denver Developmental Screening Test (Frankenburg et al. 1992; Frankenburg 1985), the Griffiths Mental Development Scale (Griffiths 1984) and some new items drawn from culturally sanctioned behaviors (Gladstone et al. 2010; Gladstone et al. 2008).

- **The Parent Report Scales of Motor and Language Development** (Stoltzfus et al. 2001) measures gross motor and language milestones via parent report for children 6–59 months of age. It has been used in Tanzania and Nepal.

Asia

- **The Indian Council of Medical Research (ICMR) Psychosocial Development Screening Test** has been used both as a screening instrument and as a tool for assessing group differences in intervention research (Vazir and Kashinath 1999).

- **The Cambodian Developmental Assessment Test** (Rao et al. 2012) measures the level of cognitive, social, motor, and academic development for program evaluation based on country-specific standards.

Latin America

- **Test de Desarollo Psicomotor (TEPSI)** (Haeussler and Marchant 1980), developed in Chile, evaluates child development in three basic areas—motor function, coordination, and language—by observing behavior in certain situations set up by the examiner.

- **Escala de Evaluación del Desarrollo Psicomotor (EEDP)**, developed in Chile (Rodriguez 1996), is a screening measure for language, social, coordination, and gross motor skills. Norms and cutoffs have been determined to classify children as normal, at risk, or delayed.

- **Escala Argentina de Inteligencia Sensorio-motriz (EAIS)** (Oiberman 2005; Oiberman 2006) is a diagnostic, qualitative measure of practical intelligence in the sensory-motor period. The test is based on observation of the child's behavior in a variety of tasks.

Multinational

- **The IDELA (International Development and Early Learning Assessment)** was developed by Save the Children in rural, impoverished communities across 11 low- and lower middle-income countries, largely because these are the communities the group serves. A primary goal of the tool was to support program evaluation, early childhood care and development, and evidence building in low-income countries.[b]

- **The International Association for the Evaluation of Educational Achievement (IEA)** developed cross-national tests of language and cognitive development, as well as child observation tools, for use in 15 different countries with children at age four years and seven years (Montle, Xiang, and Schweinhart 2006).

- **Regional Project on Child Development Indicators (PRIDI)** is an initiative launched by the Inter-American Development Bank that aims to generate high-quality and regionally comparable data on child development.[c]

- **Measuring Early Learning Quality and Outcomes (MELQO)** modules were developed by a consortium led by Brookings Institution, UNESCO, UNICEF, and the World Bank.[d]

- **The Early Human Capability Index (EHCI)** was originally developed in Tonga, and has been further developed in China, Lao PDR, Samoa, Tuvalu, Kiribati, Brazil, Peru, and Australia.[e]

a. http://www.hsrc.ac.za/ECD-Measure-158.phtml

b. Test material is available at https://idela-network.org.

c. http://steinhardt.nyu.edu/global-ties/early_childhood/melqo

d. http://www.iadb.org/en/topics/education/initiative-pridi/ home,20387.html.

e. https://www.telethonkids.org.au/our-research/brain-and-behaviour/development-and-education/child-health-development-and-education/the-early-human-capability-index-ehci/

more generally. The CREDI Long Form includes an expanded set of caregiver-reported questions and provides both overall and domain-specific scores. The CREDI Long Form is intended for use in large-scale research and evaluation studies.

The "standards" approach

Another approach to child assessment is for a country to develop a set of "standards" or expectations about what every child should know and be able to do at a certain age (often four years, before the child enters school) (Kagan and Britto 2005). These standards can then be translated into assessments; a notable example is the East Asia-Pacific Early Child Development Scales (EAP-ECDS),[2] which demonstrate how standards can be used to build assessments that generate specific, measurable indicators across countries (Rao et al. 2014; Rao et al. 2016). For the EAP-ECDS, an 85-item test was developed, which includes questions about seven domains: cognitive development, social-emotional development, motor development, language and emergent literacy, health and hygiene, cultural knowledge and participation, and approaches to learning. This test was administered to children ages 3–5 years in six countries (Cambodia, China, Mongolia, Papua New Guinea, Timor-Leste, and Vanuatu). Table 7.1 shows how the EAP-ECDS utilizes parents' feedback to rate their children's competence in a range of domains. Analyses indicated that the EAP-ECDS is a reliable and valid measure of developmental functioning and school readiness in each of the six countries.[3]

TABLE 7.1 Examples of Parent Rating Items from EAP-ECDS

DOMAIN	SKILL
Cognitive Development	Ability to learn new things and solve new problems
Social-Emotional Development	Display of social skills, such as showing consideration for others and ability to manage emotions
Motor Development	Ability to run and jump Ability to hold chopsticks, spoons/pencils/pens
Language and Emergent Literacy	Language kkills
Health, Hygiene, and Safety	Practice of healthy and hygenic habits (e.g. washing hands independently) Ability to follow safety rules (e.g., not touching hot/dangerous things)
Cultural Knowledge & Participation	Participation in important community events (including festivities)
Approaches to Learning	Ability to concentrate on learning new tasks (excluding watching TV)

Source: Adapted from Rao et al. 2014; Rao et al. 2016.

These standards, or desired results, can be linked with program standards for a health care or childcare center program, resulting in a system of childhood assessment in which the expectations for children and programs are aligned for maximum effectiveness. For example, if a standard stipulates that children should be able to understand the concept of sequence by age four, then the program should be assessed in terms of its ability to provide opportunities for learning how to sequence. In the case of the EAP-ECDS, for example, the early childhood development results and the program participation results were analyzed together. The findings showed that children who attended ECD programs had significantly higher scores on ECD-related tasks.

In developing standards for early learning and development, domains are defined, and within each domain, a set of standards or goals for children is established. For each standard, a set of specific objectives is outlined for the age level, and indicators for each are specified. Indicators are often broad descriptions of behaviors and may lack the specificity needed to develop a test, but are intended to help a teacher or parent observe a child's behavior.

[2] For more details, please see: http://www.arnec.net/ecd-arnec-resources/eap-ecd-scales/.

[3] A full report is available describing results: http://www.arnec.net/wp-content/uploads/2015/07/EAP-ECDS-Final-Report1.pdf.

SIDEBAR 7.3 **Examples of Standards Adaptation from Vietnam**

DOMAIN: COGNITIVE DEVELOPMENT AND APPROACHES TO LEARNING

Sub-domain: Cognitive Development

Standard 3.2.3. Children demonstrate initiative in daily activities (3 indicators)

Performance Indicators:

- Undertakes activities in his/her own ways
- Displays his/her experiences in various ways (role playing, acting, stories telling, drawing, collaging, movement, etc.)
- Suggests new activities

DOMAIN: SOCIAL AND EMOTIONAL DEVELOPMENT

Sub-domain: Emotional Development

Standard 2.1.1. Children are able to perceive themselves (4 indicators)

Performance Indicators:

- Child tells important information about his/her self and family members (e.g., full name, birthday, address, telephone number, father and mother's full names, occupation, etc.)
- Child expresses his or her own preferences, for example favorite activities, foods, etc.
- Child suggests activities that demonstrate his or her perceived abilities, for example, playing an outdoor game.
- Child differentiates between his or her own preferences and preferences of others, for example "I like sweet potato, she likes eating corn".

Beyond relevance for measurement, the process of developing national-level standards can be of value for a country, as it brings all stakeholders together and makes them define goals and actions for children. However, this process takes time. The advantage for a country developing its own standards is that these standards cover items and domains important to the country. If governments have not developed their own child and program standards, they may find it more convenient to simply adopt standards from another country. This could lead to inappropriate standards unless they are modified for the setting.

Therefore, a major effort was initiated, beginning in 2003 and led in part by UNICEF, to help countries define what they expect children of various age groups to know and be able to do (Kagan, Britto, and Engle 2005). Country teams (experts, policymakers, teachers, and families) first define the most appropriate domains for their country, along with possible sub-domains, and the age groups for which they wish to define standards. The next step is for the country to develop a set of standards, or expectations for learning, that are appropriate to their cultural context. In this setup, standards are statements that specify an expectation for achievement of skills or knowledge. Within each standard are several indicators that can be used to assess the standard. Domains may have sub-domains defined as well, along with a standard and several indicators. A complete set of standards would include suggestions for activities for achieving these standards.

In sum, for each domain or sub-domain of development (e.g., cognitive, language, social, or physical), there are a set of statements that stipulate what children should be able to do, and a series of indicators that define what percentage of children should be able to do the defined task by a certain age. Researchers tend to use the top 50% to define children who pass the item, although others use the top 75% to avoid mislabeling children as "slow" when they are potentially performing within a normal range.

Sidebar 7.3 shows two examples from Vietnam for children aged 5–6 years. Not all of the performance indicators are specific enough for testing, but it may still be possible to test some of them using easily administered items. Because these indicators are often used to help teachers of young children to plan curricula, improve teaching, and develop awareness of children's skills, these performance indicators can be translated into items that are suitable for testing in situations where they will be used for assessment in a systematic way. Typically, some "gap" exists between the full intent of the standard or indicator, and what is possible to test. For example, in a short observation with a child, it may be more feasible to assess whether he or she is able to state important information about himself or herself, rather than to describe how the child initiates activities. However, by starting with the standards, it is possible to generate a list of items that may come closer to policy and cultural expectations for children's development than by starting with an existing assessment, in the absence of a careful review of standards.

The standards approach requires each country to develop its own set of early learning standards that are culturally appropriate. It can be better to develop standards that are appropriate to the national environment than to use a measure developed somewhere else that has no relationship with the country's values for its children. However, experience has shown that it is helpful for countries to see what others have done and to reference those standards to help define their own. Table 7.2 lists some of the benefits and drawbacks of taking a standards approach in assessing early child development.

UNICEF's team has been working with more than 40 countries to develop standards. Many of them are now being validated for each age group. This process can take between three months and one year, depending on interest and the breadth of the effort. The more ages selected and the more domains included, the longer it will take. The process should be participatory and country-specific.

ECD standards can be utilized for numerous purposes, as summarized in Sidebar 7.4. If standards are used effectively in classrooms, they assist teachers in focusing on goals for individual children, planning activities to achieve those goals, monitoring the child's progress toward the goal, and assessing the child's progress periodically. This approach to preschool education should result in individualized, age-appropriate, and effective learning experiences for children. Given the constraints faced by many programs for disadvantaged children in low- and middle-income countries, however, only a small portion of these activities may be possible.

To use standards for population- or individual-level assessment, it is necessary to translate them into an assessment form. They can be collated at the individual item level, to assess learning and progress on each item. Creating a single scale or test from these standards requires a second step of test creation.

TABLE 7.2 Pros and Cons of Standards Approach

+ PROS	− CONS
Standards are culturally appropriate. These measures have been defined by each country, and therefore are appropriate for that specific country.	Development is time-intensive and requires long-term follow-up. It can take as long as a year to develop the standards and complete an age validation (to see if indeed children are able to perform as the standards recommend).
The process increases understanding of early child development. For countries that have developed their own local standards, the process of reaching consensus within a group about what children should know and be able to do before entering school is valuable for planning, program development, and policy development.	Indicators are not easily translated into a test. Indicators as developed by a standards-writing team often tend to be too vague to use as a test item. To adapt the standards to a test, more work needs to be done to clarify and specify the indicators clearly enough to justify a test.
	The modified item might not be equivalent in difficulty to the original item, which would mean the scores on the adapted test are not equivalent to those for the original test.
	The total number of items is not equivalent to the original scale, therefore the raw scores on the adapted test are not equivalent to those on the original test, and norms cannot be applied.

SIDEBAR 7.4 Uses for Standards

Early learning and development standards are used for many purposes:

- **Individual child development.** Used by teacher or health worker to assess what the child can do and to decide on a learning plan for his or her development
- **Curriculum development.** Used by policymakers or other experts to decide what kind of lessons and experiences should be included
- **Program quality.** Used by experts to design teacher training methods and supervision criteria, aimed at helping teachers and schools recognize what should be in their curriculum and at developing systems for program accountability
- **Planning.** Used to determine where resources are most needed and to make allocations based on the findings
- **Advocacy.** Used to provide the public with greater understanding of child development and to help them recognize what percent of children might be "ready for school"
- **Monitoring and program evaluation.** Used to develop a monitoring or assessment system, as was done in Cambodia (Rao and Pearson 2007)

8 Children's Home and Early Learning Environments

KEY MESSAGES:

- The quality of children's early environments has a large influence on their development and performance on developmental assessments.
- There are many ways to measure the home and learning environment, including the "gold standard measure" (the HOME), and several measures that have been derived from the HOME, including the HSQ, FCI, MICS, and PROCESS.
- When using the HOME or any related measure, it is critical to adapt the test to the particular culture and context where the measure is being used.
- HOME scores may not be compared across cultures.
- There are many options for measuring classroom environments.

AS DISCUSSED IN PREVIOUS CHAPTERS, BRAIN DEVELOPMENT AND THE ACQUISITION of skills and capacities are built through children's interactions with their environments. Thus, the quality of children's early environments has a large influence on their development (Walker et al. 2007; Walker et al. 2011). In the first two years of life, a child's explorations of the world typically take place close to home, but as children become more independent, their social and physical settings broaden, especially when they start to attend school (Rimm-Kaufman and Pianta 2000; Whiting and Edwards 1988; Bronfenbrenner 1986). This chapter reviews validated tools for measuring two important influences on children's early development: the home and early learning environments.

Measuring the home environment and family functioning

Associations between household socio-economic status (SES) and cognitive, behavioral (Bradley and Corwyn 2002), and language (Hart and Risley 1995) development are well established. Over the past 50 years, social scientists have focused on unpacking the mechanisms for understanding how SES components (e.g., household education level, occupational status, income, and social position) affect children's development (Conger and Donnellan 2007).

There are two major perspectives for how socio-economic status impacts development (Bradley and Corwyn 2002; Conger and Donnellan 2007). *Social causation* proposes that household socio-economic status influences child outcomes through its effects on parenting behaviors. These behaviors include the extent to which parents invest in a child's learning (e.g., providing stimulating materials and experiences, engaging child in learning activities), as well as the extent to which parents maintain warm and supportive relationships with the child.

In contrast, *social selection* posits that the SES-child development association is spurious because of a largely unmeasured third set of variables: individual (parental) characteristics. Some argue that the traits and dispositions of individuals (e.g., intelligence, interpersonal skills, motivation, diligence) that determine household socio-economic status also determine child outcomes. The literature shows support for both frameworks, but little work has been done that directly tests the two perspectives using the same dataset (Conger and Donnellan 2007; Conger, Conger, and Martin 2010). Parenting behaviors appear to be part of the pathway connecting parent characteristics and child outcomes in the United States (Schofield et al. 2011; Schofield et al. 2012).

Several studies conducted in various parts of the world have used mediation, or pathway, analyses to unpack the correlations between socio-economic status and child well-being and development. Research from the United States (Noble, McCandliss, and Farah 2007), Bangladesh (Hamadani et al. 2014), India, Indonesia, Peru, Senegal (Fernald et al. 2012), Mexico (Knauer et al. 2016), and Colombia (Rubio-Codina, Attanasio, and Grantham-McGregor 2016) indicate that parenting behaviors partially underlie or mediate this correlation. Formal educational attainment by parents can also be significantly related to child outcomes and thus is important to measure. Measuring family factors can help researchers explore the ways in which socio-economic status contributes to outcomes of interest. The tools reviewed focus on parent and family functioning.

The Home Observation for Measurement of the Environment (HOME)

The most widely used measure of the household environment is the Home Observation for Measurement of the Environment (HOME), which includes inventories for assessing the households of children from infancy to adolescence (Caldwell and Bradley 1984; Caldwell and Bradley 2003a; Caldwell and Bradley 2003b). Each age-related inventory includes 6–8 subscales to assess various dimensions of the home environment (Figure 8.1), including parental responsiveness, organization and safety of the household, and support for learning. It is used primarily in research settings, not as a diagnostic tool for individuals.

For each age group, the HOME has subscales that are ordered by factor analyses loadings, with the first subscale accounting for more variance than the second, and the second accounting for more variance than the third, and so on. The authors have also adapted versions for use in households caring for children with various disabilities, and there are two versions for assessing family (not center-based) childcare settings serving infants and young children.

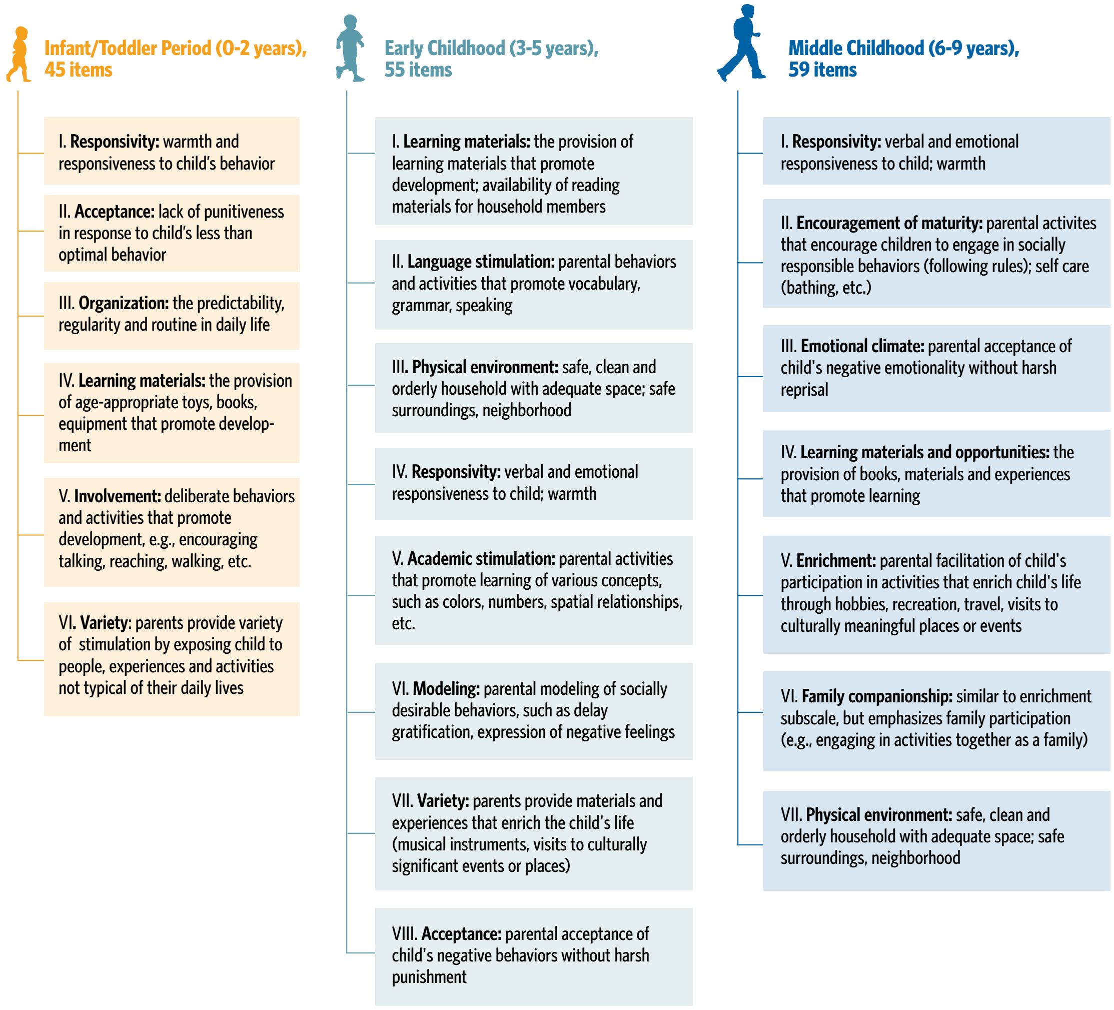
HOME inventories are completed in the home by trained personnel using both interview and observation techniques and take about 45–90 minutes to administer. Inventories for younger children tend to have more observation items than do those for older children. Response options are "Yes" or "No," and scores are computed for subscales and the total inventory. The HOME and training materials can be obtained, for purchase, from Rfrom the distribution center.[1]

The HOME interview is more of a structured interview, where the interviewer does not ask the mother direct questions but rather engages in a conversation with her. The interviewer must gather enough information to score the HOME through the course of this conversation, while observing what is going on around him or her (e.g., how the mother relates with the child, engages with the child, refers to the child). Thus, both the training for administering the HOME and the administration process for the test itself are complex and are required in addition to the training for and actual administration of a household survey.

The HOME has been used in research studies in more than 50 countries to examine parental responsiveness and the home environment. Studies using the HOME have found some clear cross-cultural inconsistencies worldwide, reinforcing the message that scores on the HOME should not be compared across cultures. Research from the United States (Brooks-Gunn et al. 1995; Bradley et al. 2001; Bradley et al. 2001; Bradley et al. 1989) and low- and middle-income countries (Bradley and Corwyn 2005; Bradley, Corwyn, and Whiteside-Mansell 1996; Bradley 2015) showed: (1) positive associations between socio-economic status and HOME scores; (2) low-moderate correlations between total HOME scores and cognitive func-

[1] Home Inventory LLC, Distribution Center, 2627 Winsor Drive, Eau Claire, WI 54703, USA.

FIGURE 8.1 Subscales of HOME Inventories for Use with Children Aged 0–8 Years



Infant/Toddler Period (0-2 years), 45 items

I. **Responsivity:** warmth and responsiveness to child's behavior

II. **Acceptance:** lack of punitiveness in response to child's less than optimal behavior

III. **Organization:** the predictability, regularity and routine in daily life

IV. **Learning materials:** the provision of age-appropriate toys, books, equipment that promote development

V. **Involvement:** deliberate behaviors and activities that promote development, e.g., encouraging talking, reaching, walking, etc.

VI. **Variety:** parents provide variety of stimulation by exposing child to people, experiences and activities not typical of their daily lives

Early Childhood (3-5 years), 55 items

I. **Learning materials:** the provision of learning materials that promote development; availability of reading materials for household members

II. **Language stimulation:** parental behaviors and activities that promote vocabulary, grammar, speaking

III. **Physical environment:** safe, clean and orderly household with adequate space; safe surroundings, neighborhood

IV. **Responsivity:** verbal and emotional responsiveness to child; warmth

V. **Academic stimulation:** parental activities that promote learning of various concepts, such as colors, numbers, spatial relationships, etc.

VI. **Modeling:** parental modeling of socially desirable behaviors, such as delay gratification, expression of negative feelings

VII. **Variety:** parents provide materials and experiences that enrich the child's life (musical instruments, visits to culturally significant events or places)

VIII. **Acceptance:** parental acceptance of child's negative behaviors without harsh punishment

Middle Childhood (6-9 years), 59 items

I. **Responsivity:** verbal and emotional responsiveness to child; warmth

II. **Encouragement of maturity:** parental activites that encourage children to engage in socially responsible behaviors (following rules); self care (bathing, etc.)

III. **Emotional climate:** parental acceptance of child's negative emotionality without harsh reprisal

IV. **Learning materials and opportunities:** the provision of books, materials and experiences that promote learning

V. **Enrichment:** parental facilitation of child's participation in activities that enrich child's life through hobbies, recreation, travel, visits to culturally meaningful places or events

VI. **Family companionship:** similar to enrichment subscale, but emphasizes family participation (e.g., engaging in activities together as a family)

VII. **Physical environment:** safe, clean and orderly household with adequate space; safe surroundings, neighborhood

Source: Based on Cadwell and Bradley 1984; Cadwell and Bradley 2003a; and Cadwell and Bradley 2003b

tion, language abilities, and academic achievement from early childhood to adolescence; and (3) evidence that the stimulation and parental responsiveness subscales were particularly important for child outcomes.

Cultural differences have also been noted (Bradley and Corwyn 2005; Bradley, Corwyn, and Whiteside-Mansell 1996). These included the degree to which parents engage in certain types of behaviors (e.g., the value placed on engaging children in academically enriching activities); attitudes on the use of physical punishment; accessibility of materials (books, toys) and of experiences that promote different kinds of developmental growth; and specific relationships between subscales and social-emotional and motor outcomes. Additionally, total HOME scores are not always associated with the same health outcomes: Low scores may correlate with undernutrition in some contexts, but obesity in others. This pattern suggests that there may be a quadratic (inverse U-shaped) relationship, instead of a linear relationship between HOME scores and weight-for-age, with low HOME scores associated with very low weight-for-age and very high weight-for-age, but high HOME scores in the middle range of weight-for-age.

Many places have adapted or supplemented HOME inventories to reflect cultural values. For example, in areas where respect for elders and adult authority is integral to socialization, and punishment is considered necessary for teaching children to be respectful, items relating to acceptance and physical punishment were altered or dropped altogether. Researchers in Japan and Kenya added items to assess support for development of valued social skills not present in the original HOME. For these reasons, it is strongly recommended to conduct careful adaptation of the HOME before use (Bradley 2015); furthermore, HOME scores should not be compared across countries.

Tools derived from the HOME

Many of the home environment tools developed after the publication of the HOME draw heavily from the HOME inventories. Most were developed as shorter questionnaires that include fewer observational items and require less training than the HOME. Some of the most well-known measures and their key characteristics are outlined in Table 8.1 and described in more detail below.

TABLE 8.1 Home Environment Questionnaires Derived from the HOME

TOOL	AGES	COST	ADMINISTRATION TIME	TRAINING REQUIRED	WHERE TO FIND MATERIALS
HOME-SF	0-14 years; 20-30 years	Items are free; HOME manual is $50	15-20 minutes	HOME manual recommended	https://www.nlsinfo.org/sites/nlsinfo.org/files/attachments/12127/mothersup1986.pdf For observation items: https://www.nlsinfo.org/sites/nlsinfo.org/files/attachments/12127/childsup1986.pdf For manual: http://fhdri.clas.asu.edu/home/contact.html
Family care indicators (FCI) / UNICEF MICS Early Child Development Index	0-3 years, 3 items; 0-5 years, 4 items	Free	10 minutes or less	None	http://mics.unicef.org/tools
Home Screening Questionnaire (HSQ)	0-3 years, 30 items; 3-6 years, 34 items.	Out of print	15-20 minutes	HSQ manual recommended	No longer available from the publisher (Denver Developmental Materials)
Pediatric Review and Observation of Children's Environmental Support and Stimulation (PROCESS)	2-18 months: 24-item questionnaire, 40-item toy checklist, 20-item observational tool	Free	30 minutes	Administration and scoring manual recommended	Robert Bradley School of Social & Family Dynamics Arizona State University 951 S. Cady Mall Tempe, AZ 85287

■ HOME-Short Form (HOME-SF)

The HOME-SF is an abbreviated version of the inventories (Infant/Toddler, Early Childhood, Middle Childhood, Early Adolescence) created by the HOME authors for use in the U.S. National Longitudinal Survey of Youth (NLSY).[4] The HOME-SF includes items grouped into two subscales (emotional support and cognitive stimulation), each of which contains roughly half the number of items per age group as does the original HOME. Most items are administered through an interview, but some involve observation as well. Response options for the interview portion are in multiple-choice format, but are converted to binary variables for scoring. Normed scores were computed for the NLSY and are appropriate for use in U.S. samples. The measures function similarly to the full HOME inventories in terms of detecting differences in parenting behaviors in relation to socio-economic status and in associations with child outcomes (Bradley et al. 2001). The scales showed good reliability and validity and are suitable for use in large U.S. field studies (Mott 2004), but have not been studied extensively in other countries. The HOME-SF is best suited for administration during a visit to the household and can be completed in 15–20 minutes.

[4] For more information, please see https://www.nlsinfo.org/.

■ UNICEF Multiple Indicator Cluster Surveys (MICS) Early Childhood Development Module and Family Care Indicators (FCI)

The UNICEF Multiple Indicator Cluster Surveys Early Childhood Development Module and Family Care Indicators (FCI) are both shortened versions of the HOME. They measure access to play materials and books, the availability of alternative caregivers, and whether caregivers recently engaged in any of six stimulating activities with the child. The FCI items were developed through quantitative and qualitative methodology and tested in multiple sites (Kariger et al. 2012), and then the UNICEF Multiple Indicator Cluster during the last three days incorporated the Family Care Indicators into its survey for use with parents of children ages 0–4 years.[5] The two tools are similar to each other, but are not identical. The FCI includes a list of play materials by type but the UNICEF MICS only classifies play material by source and it does not include a list of play materials by type. The Family Care Indicators survey requires that information about play materials be collected through direct observation, not direct report, and hence is subject to less reporting biases.

Analyses from 28 countries using the Multiple Indicator Cluster Surveys data showed that higher country gross domestic product (GDP) was positively associated with provision of learning resources, such as toys and books, as well as engagement in the six activities (Bornstein and Putnick 2012). Three items in the MICS can be used with all children under five years, and one extra item is recommended for use with children 3–4 years of age. The module has been widely used throughout the world and provides information on caregiving practices in countries previously understudied (Bornstein and Putnick 2012). These items have been associated with child development outcomes in a variety of countries (Fernald et al. 2012; Knauer et al. 2016; Hamadani et al. 2010). Administration is brief, and the module has been translated into multiple languages.[6] Data are available from more than 30 countries, allowing for cross-country comparisons. An expanded version of these and related items has been used in Bangladesh (Hamadani et al. 2010), Burkina Faso, Ghana, and Malawi (Prado et al. 2016). In Bangladesh, scores on this expanded version were moderately correlated with HOME scores and cognitive and language outcomes.

■ Home Screening Questionnaire (HSQ)

The Home Screening Questionnaire (HSQ) is a parent-completed questionnaire for use with children 0–6 years of age (Frankenburg and Coons 1986). There are two forms: one for use with parents of children aged 0–3 years (30 items), and the other for use with parents of children aged 3–6 years (34 items). Unlike with the HOME-SF, there are no observation items, and the measure can be completed without a home visit (e.g., during pediatric check-ups). The items focus on activities and materials that promote learning in young children, but do not assess the emotional environment of the home. Validation studies with the full HOME (administered during subsequent household visits) showed the HSQ scores were accurate in predicting low HOME scores (less than the 50th percentile) more than 80 percent of the time in U.S. samples (Frankenburg and Coons 1986). The Home Screening Questionnaire has been used in some non-Western countries, including Turkey (Kesiktas et al. 2009), South Africa (Richter and Grieve 1991), and India (Nair et al. 2009). The questionnaire can be administered in about 15–20 minutes.

■ Pediatric Review and Observation of Children's Environmental Support and Stimulation (PROCESS)

The PROCESS was created for use with parents of children 2–18 months of age and can be administered in a clinic or home setting (Casey et al. 1988). The PROCESS consists of three sections: a parent-completed questionnaire of 24 items; a 20-item observational tool; and a 40-item toy checklist. Some items were drawn from the HOME, but others were newly generated and tested in an iterative process. The parent questionnaire includes items about the physical environment, household organization, and stimulation for development. The observational items focus primarily on the emotional quality of parent-child interactions, as observed during a clinic or home visit. Total scores are summed across the three sections. In the United States, validation studies found high correlations with the HOME, and low PROCESS scores predicted low HOME scores in 77 percent of the sample (Casey et al. 1988). The PROCESS has not been used widely outside of the United States. The total time for administration and observation is about 30 minutes.

[5] Please see http://mics.unicef.org/tools.
[6] Please see http://mics.unicef.org/contents-by-survey - MICS5.

Tools for observing parent-child interactions

One criticism of the HOME and related measures is that they rely on parental report (at least partially) and therefore may be biased. Objective observation of select parenting behaviors may offer an alternative, valid option for estimating the quality of parent-child interactions associated with better child outcomes. Observational systems, however, can be cumbersome to learn and implement reliably (Fuligni and Brooks-Gunn 2013), and, as many were developed for use with middle-class U.S. samples, could lack validity when applied to multi-ethnic groups (Ispa et al. 2013). Two recently developed, quick and easy-to-use observation tools appear promising for use in diverse environments and are described below.

■ Parenting Interactions with Children: Checklist of Observations Linked to Outcomes (PICCOLO)

The PICCOLO (Roggman et al. 2013) is a coding scheme that includes 29 items across four parenting domains (affection, responsiveness, encouragement, and teaching) and that is designed for use with parents and children under the age of three years. Coding is completed while watching a parent-child pair interact around play materials or books, conducted either live or via videotaped sessions. For each item, the observer rates if it is clearly present, barely present, or not observed.

To create the PICCOLO, early child development practitioners identified 80 parenting behaviors important for child development. These were then evaluated for their psychometric properties using data collected on more than 2,000 children attending Early Head Start programs, which target low-income families in the United States and provide child development and family support services for pregnant women and children under the age of three years. The resulting 29 items were those that showed the strongest inter-rater reliability, factor structure, and construct validity (i.e., strong associations with related measures). Predictive validity was demonstrated by significant correlations between both scores on the domains and the total PICCOLO score and later child cognitive and language outcomes, as well as a school readiness index, although the strength of the associations was low.

The PICCOLO has been adapted for use in Turkey (Bayoğlu et al. 2013) but has not otherwise been widely used outside the United States. It has potential for adaptation in other countries as it was developed using multi-ethnic U.S. samples, can be completed with brief live or videotaped sessions, and complements more subjective measures of the home environment. The PICCOLO is a copyrighted measure and requires purchase, with materials ranging in price from $15–$150. It also requires 1-2 days of training.[7]

■ Observation of Mother-Child Interactions (OMCI)

The OMCI (Rasheed and Yousafzai 2015) is a brief parent-child interaction tool developed for an evaluation of a parenting and nutrition intervention in Pakistan (Yousafzai et al. 2016). The tool includes 19 items, drawn from theoretical and expert review, to code behaviors during a live five-minute book reading session with the mother and child at 12 and 24 months of age. Twelve items focus on the parent's behaviors (e.g., "Is sensitive to child's needs, for example follows child's lead, accepts child's disinterest in book and does not force child to play with it any longer"); six capture the child's behaviors (e.g., "Shows excitement and enjoyment like clapping"); and one measures mutual enjoyment. Coders rate the frequency of each behavior using a scale of 0–3, with higher scores indicating greater frequency and more positive interaction. The tool yields an overall interaction score, along with separate scores for the parent and child.

The OMCI showed good reliability, with high internal consistency, and moderate inter-rater reliability (between trainer and university-level fieldworkers). Pearson's correlations with HOME responsiveness and involvement scores were low but significant, as were those with measures of maternal knowledge and maternal depression, indicating some degree of construct validity. The 12-month OMCI score showed moderate predictive validity of growth, cognitive, motor, and language development at 24 months. Importantly, the OMCI is freely available, requires minimal training, and can be implemented in large field studies, showing great promise for adaptability in other low- and middle-income settings (Rasheed and Yousafzai 2015).

[7] More information is available at http://www.brookespublishing.com/resource-center/screening-and-assessment/piccolo/.

Measuring the quality of early learning environments

The quality of early learning environments has a critical impact on young children's development (Engle et al. 2011; Yoshikawa et al. 2013). Children attending higher-quality preschools show better learning and behavioral outcomes throughout the world (Engle et al. 2011). Factors that have been identified as significant for estimating quality have generally been classified as either "structural" or "process" variables. Structural variables refer to those that assess the quality of the housing, rooms, availability of materials, scheduling of activities, teacher-to-child ratio, classroom size, schooling and payment of teachers, and many other administrative details. Process variables are primarily those that examine how teachers or caregivers interact with the children they supervise, and include teaching styles, responsiveness to child needs, flexibility to adjust teaching and supervision in response to child or classroom needs, and communication style when interacting with children. Factors notably associated with better child outcomes include higher levels of teacher education and training, child autonomy (for instance, choice of activities), more time spent in small group activities, small class size, low teacher-to-child ratio, responsive interactions, high and consistent levels of child participation, language-rich environments, age-appropriate curricula, stimulating materials, and a safe environment (Montie, Xiang, and Schweinhart 2006; Yoshikawa et al. 2013).

Measuring the quality of settings can help researchers identify their influence on child outcomes. Research in the United States suggests that improvement in the quality of process variables is the likely core driver of improved child outcomes, once adequate structural quality has been achieved (NICHD Early Child Care Research Network 2002). The Inter-American Development Bank recently published a toolkit for measuring the quality of early learning environments for children under age three years (Lopez Boo, Araujo, and Tome 2016).

In this section, we briefly review assessments that have been widely used and adapted throughout the world.

■ Infant/Toddler Environment Rating Scale-Revised (ITERS-R) and Early Childhood Environment Rating Scale-Revised (ECERS-R)

The ITERS-R (Harms, Cryer, and Clifford 2006) is designed for use in settings with children up to 30 months of age, and the ECERS-R (Sylva et al. 2006) is suitable for children 30–60 months of age. The ITERS-R includes 39 items that assess quality across seven dimensions: space and furnishings, personal care routines, listening and talking, activities, interaction, program structure, and parents and staff. The ECERS-R consists of 43 items distributed across seven subscales: space and furnishings, personal care routines; language-reasoning; activities, interaction, program structure, and parents and staff. For both scales, items include detailed notes for rating on a seven-point Likert scale. Subscale and total scores are computed. The authors recommend allowing three hours for observation and 30 minutes for scoring. Training materials are available for purchase through the Environment Rating Scales Institute web page.[8]

For the ECERS-R, some evidence suggests that compared to structural items, those measuring process variables related to teaching and interactions were more highly associated with concurrent (Howes et al. 2008) and future (Burchinal et al. 2008) child outcomes. The ECERS-R has been adapted for use in many countries throughout the world, including those in Latin America (Lopez Boo 2016), Asia (Brinkman et al. 2016), and Africa (Malmberg, Mwaura, and Sylva 2011).

The ECERS-Extension (ECERS-E) for children 3–5 years of age supplements the ECERS-R. It focuses on the provision of specific materials and activities that promote language, math, and scientific exploration. Scores on the ECERS-E predicted child outcomes in the United Kingdom (Sylva et al. 2006) and moderated the effects of preschool interventions on cognitive outcomes in East Africa (Malmberg, Mwaura, and Sylva 2011). The tests were specifically designed to reflect each country's national curriculum, but this could be adapted to a different local curriculum.

■ Classroom Assessment Scoring System (CLASS)

CLASS (Pianta, La Paro, and Hamre 2008) is an observational tool that was first developed for the large-scale NICHD Study of Early Child Care and was later improved and extended to cover other age groups (Hamre et al. 2013). The CLASS tool involves four cycles of 15-minute observations of teachers and students by a certified observer. Those observations are then rated using a manual of behaviors and

[8] See http://ersi.info/order.htm.

responses. A recent study found support for both the construct and predictive validity of the teaching through interactions conceptual framework as assessed by the CLASS in Chile (Leyva et al. 2015). A version of the CLASS scale, called the Teacher Instructional Practices and Processes System (TIPPS), has also been developed, and validation work is underway.

■ Measure of Early Learning Environments (MELE)

As part of the Measuring Early Learning Quality and Outcomes project, a set of tools designed to measure classroom environments was created, specifically for adaptation and use in low- and middle-income countries. The MELE tools were developed from a literature review to identify universally relevant aspects of learning environments that predict child outcomes and also from advice from a consortium of experts with experience in measuring learning environments across countries (please see UNESCO [2017] for an overview). Tools were developed to address the application of definitions of "quality" that come from high-income countries and may not be applicable in low- or middle-income settings. Seven domains were identified as having relevance across contexts: interactions, pedagogy, play, inclusiveness, environment, family and community engagement, and personnel. The MELE has a classroom observation module; a teacher survey on teacher characteristics, motivations, compensation, and approach to pedagogy; and a director survey that outlines professional development opportunities and other aspects of schools that have been shown to influence quality in classrooms. The MELE is designed for adaptation to different contexts through discussions with stakeholders and alignment with national standards and has been used in several countries in partnership with the World Bank. Validation work is underway. Tools are open source.[9]

[9] The material is available at http://ecdmeasure.org/.

9 Summary and Recommendations

IN CHAPTER 5, THIS TOOLKIT LISTS 10 CRITERIA OF AN IDEAL EARLY CHILDHOOD development assessment. No current test meets all of these criteria; however, technological advances are rapidly changing the range of possibilities. In the next decade, we expect to see immense progress toward the ideal. We have provided an overview of some of the methods that are now possible with more advanced technology, though many are still expensive and require a high level of expertise.

At the present time, the selection of any assessment will require a trade-off between different aspects of the ideal test. The purpose of the assessment and the budgetary and logistical constraints of the project will inform which criteria to prioritize. The three broad purposes of (1) population monitoring, (2) program evaluation, and (3) hypothesis-driven research require differing depth and detail of assessment (Figure 9.1). The degree of adaptation required also depends on the assessment method selected and project-specific goals (Figure 1.1). Increasing adaptation will strengthen the validity of the assessment in the local context and the probability of detecting intervention effects, but may weaken the comparability to scores on the same test in other studies (Figure 1.2).

Successful program evaluations (e.g., for early childhood education, literacy, or nutrition) hinge on accurately assessing children's development. The accuracy of the data to reflect the child's true ability depends on the validity of the method used to collect the data and the quality of the implementation of that method (Figure 4.1).

In this book, we have reviewed and discussed different approaches for measuring early childhood development, explaining the approaches for adapting these for practical use in different contexts. We have devised a set of recommendations to guide successful and accurate early childhood development measurements. Following these guidelines will advance the field of international early child development, as projects clarify the influences on ECD in low- and middle-income countries and the policies and programs that can support children to achieve their full developmental potential. The recommendations are as follows:

RECOMMENDATION 1 **Decide on the type of outcome measure that is appropriate.** Decide whether the purpose of the assessment is to screen for developmental delay or to have a quantitative measure of development. Decide whether the goal is to have a measure of the population or an individual-level assessment. Decide whether it is more important to make a comparison within a culture (e.g., comparing an intervention and control group in an evaluation) or a comparison across cultures (e.g., developing a global assessment of children's development).

RECOMMENDATION 2 **Consider the cultural context and how it may affect children's development and school readiness.** While the tests recommended in this Toolkit have been used in many countries, much less is known about their validity and reliability in low-income countries. Therefore, it is important for evaluators to have a strong sense of the skills and competencies that are emphasized within each culture to aid in the interpretation of the data. It is also recommended that researchers work closely with child psychologists and education specialists in the culture where the assessment will take place.

RECOMMENDATION 3 **Collect and evaluate pilot data to assess the properties of adapted tests.** Problematic items can be identified by: (1) a high percentage of missing item scores, (2) zero or low variability, (3) low or negative item-total correlations assessed by Cronbach's alpha, (4) no correlation with child age, when age progression is expected, and (5) in factor analysis, no loading on the underlying factor representing the sub-domain the item intends to measure. Test-retest reliability, convergent and discriminant validity, and inter-rater agreement should also be evaluated.

RECOMMENDATION 4 **Look for national-level tests where possible and use parent or teacher report when possible.** National-level tests with evidence for reliability and validity in the local context can be more appropriate than adaptations of tests designed for high-income country settings. Assessing children individually with standardized techniques can be time-consuming and take a lot of training by skilled professionals. Reports made by teachers, parents, or home visitors may be useful as well.

RECOMMENDATION 5 **To assess an indicator of future success, assess children at age three to five years.** Across domains, existing assessments for the age range 0–2 years are generally poor predictors of later performance (e.g., during primary school age), but become stronger predictors when children are tested at age 3-5 years. If interested in an early indicator of later academic achievement, assess pre-academic and cognitive skills, such as language, general knowledge, and executive function. If interested in an early indicator of later social-emotional and behavioral function, assess early development of social-emotional skills and self-regulation.

RECOMMENDATION 6 **Include assessments of home and early learning environments.** The quality of children's early environments has a large influence on development and performance on developmental assessments, and should be measured when possible. In the context of an impact evaluation, measuring the home and early learning environments will allow investigators to understand whether the intervention changes caregiver behaviors or the quality of the home environment in which the caregiver and child interact.

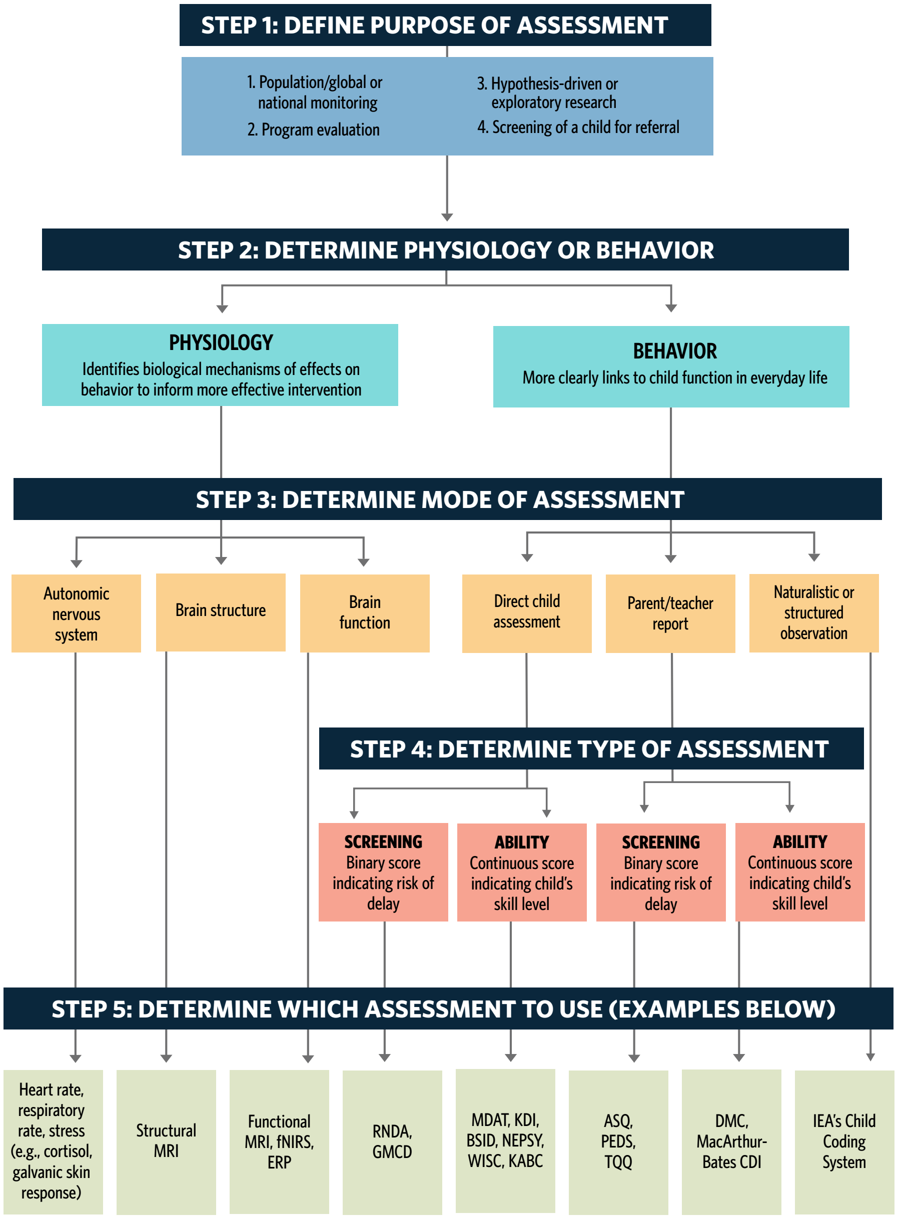
RECOMMENDATION 7 **If possible, rely upon multiple measures of children's development.** In addition to providing a more comprehensive picture of children's development, some measures index children's current development, while others may provide an indication of how children will perform in the future. Some effects of interventions are not apparent until years after the intervention (known as "sleeper effects"). For these reasons, measuring multiple domains of development is especially critical if researchers plan a longitudinal study to examine intervention effects.

RECOMMENDATION 8 **If possible, use computerized tests (administered by laptop, tablet, or smartphone).** Many traditional paper and pencil tasks can be administered in a computerized platform. These tests are generally quick to administer (1–7 minutes per test), minimize verbal instructions (which facilitates transfer from one language to another), and capture small differences in response time, increasing their likely sensitivity to intervention effects. Although this is not possible for the age range 0-3 years, it may be possible beginning at ages 4-5 years, depending on the context.

RECOMMENDATION 9 **For program evaluations, assess characteristics of the child that the intervention is intending to affect and dimensions of a child's development that you expect to be affected at the target age.** It is important to measure behaviors that the intervention is hoping to change. For example, an intervention may focus on literacy, and then the appropriate assessment instrument would be a measure of literacy. Similarly, if an intervention is using iron supplementation to help promote cognitive development, then measures of cognition most directly affected by iron status should be used. Various early childhood development domains develop on different trajectories, with motor and language skills developing rapidly at earlier ages and executive function developing at later ages. A domain that is developing rapidly at the target age is likely to show more variance in scores and therefore to be more sensitive to intervention effects.

RECOMMENDATION 10 **In program evaluations, include the same assessments at both baseline and endline.** Using the same test is likely to account for maximum variance in outcome scores and will increase statistical power to detect the effects of the intervention. The best option is to use the same test at baseline and endline. It might not be possible to use the same test if the age range of the children at baseline is very different from the age range at endline. In this case, use a test at baseline that assesses the same domain(s) as the endline test, since predictive validity within domains is stronger than across domains.

FIGURE 9.1 Flowchart for Identifying a Suitable Assessment Tool


STEP 1: DEFINE PURPOSE OF ASSESSMENT
1. Population/global or national monitoring
2. Program evaluation
3. Hypothesis-driven or exploratory research
4. Screening of a child for referral
STEP 2: DETERMINE PHYSIOLOGY OR BEHAVIOR
PHYSIOLOGY
Identifies biological mechanisms of effects on behavior to inform more effective intervention
BEHAVIOR
More clearly links to child function in everyday life
STEP 3: DETERMINE MODE OF ASSESSMENT
Autonomic nervous system
Brain structure
Brain function
Direct child assessment
Parent/teacher report
Naturalistic or structured observation
STEP 4: DETERMINE TYPE OF ASSESSMENT
SCREENING
Binary score indicating risk of delay
ABILITY
Continuous score indicating child's skill level
SCREENING
Binary score indicating risk of delay
ABILITY
Continuous score indicating child's skill level
STEP 5: DETERMINE WHICH ASSESSMENT TO USE (EXAMPLES BELOW)
Heart rate, respiratory rate, stress (e.g., cortisol, galvanic skin response)
Structural MRI
Functional MRI, fNIRS, ERP
RNDA, GMCD
MDAT, KDI, BSID, NEPSY, WISC, KABC
ASQ, PEDS, TQQ
DMC, MacArthur-Bates CDI
IEA's Child Coding System

CONSTRAINTS TO CONSIDER: budget; copyright issues; time allocated for assessment; training needs and administrator capacities; test setting; capacity of respondents; language and cultural differences requiring extensive adaptation of assessment; materials required for administration.

*Screening test cutoffs must be developed within population.

Notes: MRI, magnetic resonance imaging; fNIRS, functional near-infrared spectroscopy; ERP, event-related potential; RNDA, Rapid Neurodevelopmental Assessment; GMCD, Guide for Monitoring Child Development; MDAT, Malawi Developmental Assessment Tool; KDI, Kilifi Developmental Inventory; BSID, Bayley Scales of Infant Development; NEPSY, Developmental Neuropsychological Assessment; WISC, Wechsler Intelligence Scale for Children; KABC, Kaufman Assessment Battery for Children; ASQ, Ages & Stages Questionnaires; PEDS, Parents' Evaluation of Developmental Status; TQQ, Ten Questions Questionnaire; DMC, Developmental Milestones Checklist; CDI, MacArthur-Bates Communicative Development Inventories; IEA, International Association for the Evaluation of Educational Achievement

References

Aboud, Frances E. 2007. "Evaluation of an Early Childhood Parenting Programme in Rural Bangladesh." *Journal of Health, Population and Nutrition* 25 (1): 3–13.

Abubakar, Amina, Penny Holding, Anneloes van Baar, Charles Newton, and Fons van de Vijver. 2008. "Monitoring Psychomotor Development in a Resource-Limited Setting: An Evaluation of the Kilifi Developmental Inventory." *Annals of Tropical Paediatrics* 28 (3): 217–26.

Abubakar, Amina, Penny Holding, Fons van de Vijver, G. Bomu, and Anneloes Van Baar. 2010. "Developmental Monitoring Using Caregiver Reports in a Resource-Limited Setting: The Case of Kilifi, Kenya." *Acta Padiatrica* 99 (2): 291–97.

Abubakar, Amina, Fons J. R. van de Vijver, Sadik Mithwani, Elizabeth Obiero, Naomi Lewa, Simon Kenga, Khamis Katana, and Penny Holding. 2007. "Assessing Developmental Outcomes in Children from Kilifi, Kenya, Following Prophylaxis for Seizures in Cerebral Malaria." *Journal of Health Psychology* 12 (3): 417–30.

Abubakar, Amina, Fons van de Vijver, Anneloes van Baar, Leonard Mbonani, Raphael Kalu, Charles Newton, and Penny Holding. 2008. "Socioeconomic Status, Anthropometric Status, and Psychomotor Development of Kenyan Children from Resource-Limited Settings: A Path-Analytic Study." *Early Human Development* 84 (9): 613–21.

Adolph, Karen E. 2002. "Babies' Steps Make Giant Strides Toward a Science of Development." *Infant Behavior & Development* 25: 86–90.

Adolph, Karen E., Beatrix Vereijken, and Mark A. Denny. 1998. "Learning to Crawl." *Child Development* 69 (5): 1299–1312.

Adolph, Karen E., Beatrix Vereijken, and Patrick E. Shrout. 2003. "What Changes in Infant Walking and Why." *Child Development* 74 (2): 475–97.

Ahadi, Stephan A. and Mary K. Rothbart. 1994. "Temperament, Development, and the Big Five." In *The Developing Structure of Temperament and Personality from Infancy to Adulthood*, edited by Charles F. Halverson Jr., Geldolph A. Kohnstamm, and Roy P. Martin, 189–207. Hillsdale, NJ: Lawrence Earlbaum Associates.

Ainsworth, Mary D. Salter. 1993. "Attachment as Related to Mother-Infant Interaction." *Advances in Infancy Research* 8: 1–50.

Alaimo, Katherine, Christine M. Olson, and Edward A. Frongillo. 1999. "Importance of Cognitive Testing for Survey Items: An Example from Food Security Questionnaires." *Journal of Nutrition Educatio*n 31 (5): 269–75.

Anastasi, Anne and Susana Urbina. 1997. *Psychological Testing (Seventh Edition)*. Upper Saddle River, NJ: Prentice Hall.

Anderson, Vicki. 1998. "Assessing Executive Functions in Children: Biological, Psychological, and Developmental Considerations." *Neuropsychological Rehabilitation* 8 (3): 319–49.

Andersson, Helle W. 1996. "The Fagan Test of Infant Intelligence: Predictive Validity in a Random Sample." *Psychological Reports* 78 (3 Pt 1): 1015–26.

Araujo, Maria Caridad, Martín Ardanaz, Edna Armendáriz, Jere R. Behrman, Samuel Berlinski, Julian P. Cristia, Luca Flabbi, Diana Hincapie, Analía Jalmovich, Sharon Lynn Kagan, Florencia López Boo, Ana Pérez Exposito, and Norbert Schady. 2015. *The Early Years: Child Well-Being and the Role of Public Policy*. New York: Palgrave MacMillan; Washington, DC: Inter-American Development Bank.

Attanasio, Orazio, Costas Meghir, Emily Nix, and Francesca Salvati. 2017. "Human Capital Growth and Poverty: Evidence from Ethiopia and Peru." *Review of Economic Dynamics* 25: 234–59.

Atwine, Benjamin, Elizabeth Cantor-Graae, and Francis Bajunirwe. 2005. "Psychological Distress Among AIDS Orphans in Rural Uganda." *Social Science & Medicine*, 61 (3): 555–64.

Baddeley, Alan, Julie Meeks Gardner, and Sally Grantham-McGregor. 1995. "Cross-Cultural Cognition: Developing Tests for Developing Countries." *Applied Cognitive Psychology* 9 (7): S173–95.

Bagnato, Stephen J., Janell Smith-Jones, George McClomb, and Jennette Cook-Kilroy. 2002. *Quality Early Learning – Key to School Success: A First-Phase 3-Year Program Evaluation Research Report for Pittsburgh's Early Childhood Initiative (ECI)*. Pittsburgh, PA: SPECS Program Evaluation Research Team.

Bailey, Drew, Greg J. Duncan, Candice L. Odgers, and Winnie Yu. 2017. "Persistence and Fadeout in the Impacts of Child and Adolescent Interventions." *Journal of Research on Educational Effectiveness* 10 (1): 7–39.

Bangirana, Paul, Alla Sikorskii, Bruno Giordani, Noeline Nakasujja, and Michael J. Boivin. 2015. "Validation of the CogState Battery for Rapid Neurocognitive Assessment in Ugandan School Age Children." *Child and Adolescent Psychiatry and Mental Health* 9: 38.

Bayley, Nancy. 1969. *Manual for the Bayley Scales of Infant Development*. New York: The Psychological Corporation.

Bayley, Nancy. 2006. *Bayley Scales of Infant and Toddler Development-Third Edition*. San Antonio, TX: Harcourt Assessment.

Bayoğlu, Birgul, Özlem Unal, Fatma Elibol, Erdem Karabulut, and Mark S. Innocenti. 2013. "Turkish Validation of the PICCOLO (Parenting Interactions with Children: Checklist of Observations Linked to Outcomes)." *Infant Mental Health Journal* 34 (4): 330–8.

Behrman, Jere R., Paul Glewwe, and Edward Miguel. 2007. *Methodologies to Evaluate Early Childhood Development Programs*. Washington, DC: World Bank, Poverty Reduction and Economic Management, Thematic Group on Poverty Analysis, Monitoring and Impact Evaluation.

Belsky, Jay and Michael Pluess. 2013. "Genetic Moderation of Early Child Care Effects on Social Functioning Across Childhood: A Developmental Analysis." *Child Development* 84 (4): 1209–25.

Best, John R., Patricia H. Miller, and Jack A. Naglieri. 2011. "Relations Between Executive Function and Academic Achievement from Ages 5 to 17 in a Large, Representative National Sample." *Learning and Individual Differences* 21 (4): 327–36.

Bisiacchi, Patrizia Silvia, Giovanni Mento, and Agnese Suppiej. 2009. "Cortical Auditory Processing in Preterm Newborns: An ERP Study." *Biological Psychology* 82 (2): 176–85.

Bjorklund, David F. and Kayla B. Causey. 2017. *Children's Thinking: Cognitive Development and Individual Differences. Sixth Edition*. Thousand Oaks, CA: SAGE Publications.

Black, Maureen M., Christine Reiner Hess, and Julie Berenson-Howard. 2000. "Toddlers From Low-Income Families Have Below Normal Mental, Motor, and Behavior Scores on the Revised Bayley Scales." *Journal of Applied Developmental Psychology* 21 (6): 655–66.

Black, Maureen M., Susan P. Walker, Lia C. H. Fernald, Christopher T. Andersen, Ann M. DiGirolamo, Chunling Lu, Dana C. McCoy, Günther Fink, Yusra R. Shawar, Jeremy Shiffman, Amanda E. Devercelli, Quentin T. Wodon, Emily Vargas-Baron, and Sally Grantham-McGregor. 2017. "Early Childhood Development Coming of Age: Science Through the Life Course." *The Lancet* 389 (10064): 77–90.

Blair, Clancy. 2002. "School Readiness: Integrating Cognition and Emotion in a Neurobiological Conceptualization of Children's Functioning at School Entry." *American Psychologist* 57 (2): 111–27.

Blair, Clancy and C. Cybele Raver. 2015. "School Readiness and Self-Regulation: A Developmental Psychobiological Approach." *Annual Review of Psychology* 66: 711–31.

Blair, Clancy and Rachel P. Razza. 2007. “Relating Effortful Control, Executive Function, and False Belief Understanding to Emerging Math and Literacy Ability in Kindergarten.” *Child Development* 78 (2): 647–63.

Bland, J. Martin and Douglas G. Altman. 1986. “Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement.” *The Lancet* 327 (8476): 307–10.

Bloom, Lois. 1998. “Language Acquisition in its Developmental Context.” In *Handbook of Child Psychology, Fifth Edition. Volume 2: Cognition, Perception and Language*, edited by William Damon, Deanna Kuhn, and Robert S. Siegler, 1–50. New York: John Wiley & Sons.

Boissiere, Maurice, John B. Knight, and Richard H. Sabot. 1985. “Earnings, Schooling, Ability, and Cognitive Skills.” *The American Economic Review* 75 (5): 1016–30.

Bolig, Erika E., John Borkowski, and Jay Brandenberger. 1999. “Poverty and Health Across the Life Span.” In *Life-Span Perspectives on Health and Illness,* edited by Thomas L. Whitman, Thomas V. Merluzzi, and Robert D. White, 67–84. Mahwah, NJ: Lawrence Erlbaum Associates.

Bornstein, Marc H. and Diane L. Putnick. 2012. “Cognitive and Socioemotional Caregiving in Developing Countries.” *Child Development* 83 (1): 46–61.

Boyce, W. Thomas and Bruce J. Ellis. 2005. “Biological Sensitivity to Context: I. An Evolutionary-Developmental Theory of the Origins and Functions of Stress Reactivity.” *Development and Psychopathol*ogy 17 (2): 271–301.

Bracken, Bruce A. 2007. “Creating the Optimal Preschool Testing Situation.” In *Psychoeducational Assessment of Preschool Children*, edited by Bruce A. Bracken and Richard Nagle, 137–54. Mahwah, NJ: Lawrence Erlbaum Associates.

Bracken, Bruce A. and Andrers Barona. 1991. “State of the Art Procedures for Translating, Validating and Using Pychoeducational Tests in Cross-Cultural Assessment.” *School Psychology International* 12 (1–2): 119–32.

Bradley, Robert H. 2015. “Constructing and Adapting Causal and Formative Measures of Family Settings: The HOME Inventory as Illustration.” *Journal of Family Theory & Review* 7 (4): 381–414.

Bradley, Robert H., Bettye M. Caldwell, Stephen L. Rock, Craig T. Ramey, Kathryn E. Barnard, Carol Gray, Mary A. Hammond, Sandra Mitchell, Allen W. Gottfried, Linda Siegel, and Dale L. Johnson. 1989. “Home Environment and Cognitive Development in the First 3 Years of Life: A Collaborative Study Involving Six Sites and Three Ethnic Groups in North America.” *Developmental Psychology* 25 (2): 217–35.

Bradley, Robert H. and Robert F. Corwyn. 2002. “Socioeconomic Status and Child Development.” *Annual Review of Psychology* 53: 371–99.

Bradley, Robert H. and Robert F. Corwyn. 2005. “Caring for Children Around the World: A View from HOME.” *International Journal of Behavioural Development* 29 (6): 468–78.

Bradley, Robert H., Robert F. Corwyn, Margaret Burchinal, Harriette Pipes McAdoo, and Cynthia Garcia Coll. 2001. “The Home Environments of Children in the United States Part II: Relations with Behavioral Development Through Age Thirteen.” *Child Development* 72 (6): 1868–86.

Bradley, Robert H., Robert F. Corwyn, Harriette Pipes McAdoo, and Cynthia Garcia Coll. 2001. “The Home Environments of Children in the United States Part I: Variations by Age, Ethnicity, and Poverty Status.” *Child Development* 72 (6): 1844–67.

Bradley, Robert H., Robert F. Corwyn, and Leanne Whiteside-Mansell. 1996. “Life at Home: Same Time, Different Places – An Examination of the HOME Inventory in Different Cultures.” *Early Development and Parenting* 5 (4): 251–69.

Bradley-Johnson, Sharon and C. Merle Johnson. 2007. “Infant and Toddler Cognitive Assessment.” In *Psychoeducational Assessment of Preschool Children*, edited by Bruce A. Bracken and Richard Nagle, 325–58. Mahwah, NJ: Lawrence Erlbaum Associates.

Breitmayer, Bonnie J. and Craig T. Ramey. 1986. “Biological Nonoptimality and Quality of Postnatal Environment as Codeterminants of Intellectual Development.” *Child Development* 57 (5): 1151–65.

Bretherton, Inge, Elizabeth Bates, Laura Benigni, Luigia Camaioni, and Virginia Volterra. 1979. "Relationships Between Cognition, Communication, and Quality of Attachment." In *The Emergence of Symbols: Cognition and Communication in Infancy*, edited by Elizabath Bates, 223–269. Cambridge, MA: Academic Press.

Bricker, Diane and Jane Squires. 1999. *Ages & Stages Questionnaires: A Parent-Completed, Child-Monitoring System, Second Edition*. Baltimore, MD: Paul H. Brookes Publishing.

Briggs-Gowan, Margaret J. and Alice S. Carter. 2002. *Brief Infant-Toddler Social and Emotional Assessment (BITSEA) Manual, Version 2.0*. New Haven, CT: Yale University.

Briggs-Gowan, Margaret J. and Alice S. Carter. 2008. "Social-Emotional Screening Status in Early Childhood Predicts Elementary School Outcomes." *Pediatrics* 121 (5): 957–62.

Brinkman, Sally A., Angela Gialamas, Azizur Rahman, Murthy N. Mittinty, Tess A. Gregory, Sven Silburn, Sharon Goldfeld, Stephen R. Zubrick, Vaughan Carr, Magdalena Janus, Clyde Hertzman, and John W. Lynch. 2012. "Jurisdictional, Socioeconomic and Gender Inequalities in Child Health and Development: Analysis of a National Census of 5-Year-Olds in Australia." *BMJ Open* 2 (5): e001075.

Brinkman, Sally Anne, Amer Hasan, Haeil Jung, Angela Kinnell, Nozomi Nakajima, Menno Prasad Pradhan. 2016. *The Role of Preschool Quality in Promoting Child Development. Evidence from Rural Indonesia*. Policy Research Working Paper WPS7529, World Bank, Washington, DC.

Brinkman, Sally, Sven Silburn, David Lawrence, Sharon Goldfeld, Mary Sayers, and Frank Oberklaid. 2007. "Investigating the Validity of the Australian Early Development Index." *Early Education and Development* 18 (3): 427–51.

Britto, Pia R., Stephen J. Lye, Kerrie Proulx, Aisha K. Yousafzai, Stephen G. Matthews, Tyler Vaivada, Rafael Perez-Escamilla, Nirmala Rao, Patrick Ip, Lia C. H. Fernald, Harriet MacMillan, Mark Hanson, Theodore D. Wachs, Haogen Yao, Hirokazu Yoshikawa, Adrian Cerezo, James F. Leckman, Zulfiqar A. Bhutta, and the Early Childhood Development Interventions Review Group. 2017. "Nurturing Care: Promoting Early Childhood Development." *The Lancet* 389 (10064): 91–102.

Bronfenbrenner, Urie. 1986. "Ecology of the Family as a Context for Human Development: Research Perspectives." *Developmental Psychology* 22 (6): 723–42.

Brooks-Gunn, Jeanne, Pamela Kiebanov, Fong-ruey Liaw, and Greg J. Duncan. 1995. "Toward an Understanding of the Effects of Poverty Upon Children." *Children of Poverty: Research, Health, and Policy Issues*, edited by Hiram E. Fitzgerald, Barry M. Lester, and Barry Zuckerman, 3–41. New York: Garland Publishing.

Burchinal, Margaret, Lynne Vernon-Feagans, Martha Cox, and Key Family Life Project Investigators. 2008. "Cumulative Social Risk, Parenting, and Infant Development in Rural Low-Income Communities." *Parenting: Science and Practice* 8 (1): 41–69.

Bushnell, Emily W. and J. Paul Boudreau. 1993. "Motor Development and the Mind: The Potential Role of Motor Abilities as a Determinant of Aspects of Perceptual Development." *Child Development* 64 (4): 1005–21.

Caldwell, Bettye M. and Robert H. Bradley. 1984. *Home Observation for Measurement of the Environment*. Little Rock, AR: University of Arkansas, Little Rock.

Caldwell, Bettye M. and Robert H. Bradley. 2003a. *Home Inventory Administration Manual*. Little Rock, AR: University of Arkansas for Medical Sciences.

Caldwell, Bettye M. and Robert H. Bradley. 2003b. *HOME Inventory Early Adolescent Version*. Little Rock, AR: University of Arkansas for Medical Sciences.

Cameron, Claire E., Laura L. Brock, William M. Murrah, Lindsay H. Bell, Samantha L. Worzalla, David Grissmer, and Frederick J. Morrison. 2012. "Fine Motor Skills and Executive Function Both Contribute to Kindergarten Achievement." *Child Development* 83 (4): 1229–44.

Cameron, Claire Elizabeth, Elizabeth A. Cottone, William (Hank) Murrah, and David W. Grissmer. 2016. "How Are Motor Skills Linked to Children's School Performance and Academic Achievement?" *Child Development Perspectives* 10 (2): 93–8.

Campbell, Susan B. 2005. “Maladjustment in Preschool Children: A Developmental Psychopathology Perspective.” In *The Blackwell Handbook of Early Childhood Development*, edited by Kathleen McCartney and Deborah Phillips, 358–77. Malden, MA: Wiley-Blackwell Publishing.

Canault, Mélanie, Marie-Thérèse Le Normand, Samy Foudil, Natalie Loundon, and Hung Thai-Van. 2016. “Reliability of the Language ENvironment Analysis System (LENA™) in European French.” *Behavior Research Methods* 48 (3): 1109-24.

Cantor, Alan B. 1996. “Sample-Size Calculations for Cohen’s Kappa.” *Psychological Methods* 1 (2): 150–153.

Carlson, Stephanie M. 2005. “Developmentally Sensitive Measures of Executive Function in Preschool Children.” *Developmental Neuropsychology* 28 (2): 595–616.

Carter, Julie A., Janet A. Lees, Gladys M. Murira, Joseph Gona, Brian G. R. Neville, and Charles R. J. C. Newton. 2005. “Issues in the Development of Cross-Cultural Assessments of Speech and Language for Children.” *International Journal of Language & Communication Disorders* 40 (4): 385–401.

Casey, Patrick H., Robert H. Bradley, Joann Y. Nelson, and Steven A. Whaley. 1988. “The Clinical Assessment of a Child’s Social and Physical Environment During Health Visits.” *Journal of Developmental & Behavioral Pediatrics* 9 (6): 333–8.

Caskey, Melinda, Bonnie Stephens, Richard Tucker, and Betty Vohr. 2011. “Importance of Parent Talk on the Development of Preterm Infant Vocalizations.” *Pediatrics* 128 (5): 910–6.

Cattell, Raymond B. 1963. “Theory of Fluid and Crystallized Intelligence: A Critical Experiment.” *Journal of Educational Psychology* 54 (1): 1–22.

Cole, Michael. 1999. “Culture-Free Versus Culture-Based Measures of Cognition.” In *The Nature of Cognition*, edited by Robert J. Sternberg, 654–64. Cambridge: MIT Press.

Colombo, John, Susan E. Carlson, Carol L. Cheatham, Kathleen M. Fitzgerald-Gustafson, Amy Kepler, and Tasha Doty. 2011. “Long-Chain Polyunsaturated Fatty Acid Supplementation in Infancy Reduces Heart Rate and Positively Affects Distribution of Attention.” *Pediatric Research* 70 (4): 406–10.

Colombo, John, Susan E. Carlson, Carol L. Cheatham, D. Jill Shaddy, Elizabeth H. Kerling, Jocelynn M. Thodosoff, Kathleen M. Gustafson, and Caitlin Brez. 2013. “Long-Term Effects of LCPUFA Supplementation on Childhood Cognitive Outcomes.” *The American Journal of Clinical Nutrition* 98 (2): 403–12.

Conger, Rand D., Kathleen J. Conger, and Monica J. Martin. 2010. “Socioeconomic Status, Family Processes, and Individual Development.” *Journal of Marriage and Family* 72 (3): 685–704.

Conger, Rand D. and M. Brent Donnellan. 2007. “An Interactionist Perspective on the Socioeconomic Context of Human Development.*” Annual Review of Psychology* 58: 175–99.

Couperus, Jane W. and Charles A. Nelson. 2006. “Early Brain Development and Plasticity.” In *Blackwell Handbook of Early Childhood Development*, edited by Kathleen McCartney and Deborah Phillips, 85–105. Oxford: Blackwell Publishing Ltd.

CREDI (Caregiver-Reported Early Childhood Development Index). https://sites.sph.harvard.edu/credi/.

Cronbach, Lee J. 1951. “Coefficient Alpha and the Internal Structure of Tests.” *Psychometrika* 16 (3): 297–334.

Cueto, Santiago, Juan Leon, and Gabriela Guerrero. 2009. “Psychometric Characteristics of Cognitive Development and Achievement Instruments in Round 2 of Young Lives.” Young Lives Technical Note #15, Department of International Development, Oxford. *http://www.younglives.org.uk/pdf/publication-section-pdfs/technical-notes-pdfs/YL-TN15-Cueto.pdf.*

Damon, William, Deanna Kuhn, and Robert S. Siegler, eds. 1998. *Handbook of Child Psychology, Fifth Edition. Volume 2: Cognition, Perception and Language.* Hoboken, NJ: John Wiley & Sons.

Darrah, Johanna, Lynn Redfern, Thomas O. Maguire, A. Paul Beaulne, and Joe Watt. 1998. “Intra-Individual Stability of Rate of Gross Motor Development in Full-Term Infants.” *Early Human Development* 52: 169–79.

Darling, Kristen E. 2016. *Inventory of Measures of Social and Emotional Development in Early Childhood.* Bethesda, MD: Child Trends.

Denboba, Amina Debissa; Leslie K. Elder, Joan Lombardi, Laura B. Rawlings, Rebecca Kraft Sayre, Quentin T. Wodon. 2014. “Stepping up Early Childhood Development: Investing in Young Children for High Returns.” Working Paper 92988, World Bank, Washington, DC.

Denham, Susanne A., Hideko H. Bassett, and Katherine Zinsser. 2012. “Early Childhood Teachers as Socializers of Young Children’s Emotional Competence.” *Early Childhood Education Journal* 40 (3): 137–43.

Denham, Susanne A., Kimberly A. Blair, Elizabeth DeMulder, Jennifer Levitas, Katherine Sawyer, Sharon Auerbach-Major, and Patrick Queenan. 2003. “Preschool Emotional Competence: Pathway to Social Competence.” *Child Development* 74 (1): 238–56.

Denham, Suzanne A., Peter Ji, and Bridget Hamre. 2010. *Compendium of Preschool Through Elementary School: Social-Emotional Learning and Associated Assessment Measures*. Chicago: Department of Psychology, University of Illinois at Chicago, Social and Emotional Learning Research Group.

Devercelli, Amanda E., R. Sayre, and A. Denboba. 2016. *What Do We Know About Early Child Development Policies in Low and Middle-Income Countries?* An initial review from the Systems Approach for Better Educational Results-Early Child Development (SABER-ECD), World Bank, Washington, DC.

Doig, Katrina B., Michelle M. Macias, Conway F. Saylor, Jeffery R. Craver, and Pamela E. Ingram. 1999. “The Child Development Inventory: A Developmental Outcome Measure for Follow-Up of the High-Risk Infant.” *The Journal of Pediatrics* 135 (3): 358–62.

Dubeck, Margaret M. and Amber Gove. 2015. “The Early Grade Reading Assessment (EGRA): Its Theoretical Foundation, Purpose, and Limitations.” *International Journal of Educational Development* 40: 315–22.

Duchowski, Andrew. 2007. *Eye Tracking Methodology: Theory and Practice*. London: Springer-Verlag.

Duckworth, Angela L., Christopher Peterson, Michael D. Matthews, and Dennis R. Kelly. 2007. “Grit: Perseverance and Passion for Long-Term Goals.” *Journal of Personality and Social Psychology* 92 (6): 1087–101.

Duckworth, Angela L., Eli Tsukayama, and Teri A. Kirby. 2013. “Is It Really Self-Control? Examining the Predictive Power of the Delay of Gratification Task.” *Personality and Social Psychology Bulletin* 39 (7): 843–55.

Dugas, Lara R., Pascal Bovet, Terrence E. Forrester, Estelle Victoria Lambert, Jacob Plange-Rhule, Ramón Angel Durazo-Arvizu, David A. Shoham, Jacolene Kroff, Guichan Cao, Richard S. Cooper, Soren Brage, Ulf Ekelund, and Amy Luke. 2014. “Comparisons of Intensity-Duration Patterns of Physical Activity in the US, Jamaica and 3 African Countries.” *BMC Public Health* 14: 882.

Duncan, Greg J., Chantelle J. Dowsett, Amy Claessens, Katherine Magnuson, Aletha C. Huston, Pamela Kato Klebanov, Linda S. Pagani, Leon Feinstein, Mimi Engel, Jeanne Brooks-Gunn, Holly Sexton, Kathryn Duckworth, and Christa Japel. 2007. “School Readiness and Later Achievement.” *Developmental Psychology* 43 (6): 1428–46.

Durkin, Maureen S., Leslie L. Davidson, Patricia Desai, Z. Meher Hasan, Naila Khan, Patrick E. Shrout, Marigold J. Thorburn, Wei Wang, and Sultana S. Zaman. 1994. “Validity of the Ten Questions Screened for Childhood Disability: Results from Population-Based Studies in Bangladesh, Jamaica, and Pakistan.” *Epidemiology* 5 (3): 283–9.

Engle, Patrice L., Maureen M. Black, Jere R. Behrman, Meena Cabral de Mello, Paul J. Gertler, Lydia Kapiriri, Reynaldo Martorell, Mary Eming Young, and the International Child Development Steering Group. 2007. “Strategies to Avoid the Loss of Developmental Potential in More Than 200 Million Children in the Developing World.” *The Lancet* 369 (9557): 229–42.

Engle, Patrice L., Lia C. H. Fernald, Harold Alderman, Jere Behrman, Chloe O’Gara, Aisha Yousafzai, Meena Cabral de Mello, Melissa Hidrobo, Nurper Ulkuer, Ilgi Ertem, Selim Iltus, and the Global Child Development Steering Group. 2011. “Strategies for Reducing Inequalities and Improving Developmental Outcomes for Young Children in Low-Income and Middle-Income Countries.” *The Lancet* 378 (9799): 1339–53.

Ertem, Ilgi O., Derya G. Dogan, Canan G. Gok, Sevim U. Kizilates, Ayliz Caliskan, Gulsum Atay, Nilgun Vatandas, Tugba Karaaslan, Sevgi G. Baskan, and Domenic V. Cicchetti. 2008. “A Guide for Monitoring Child Development in Low- and Middle-Income Countries.” *Pediatrics* 121 (3): e581–9.

Ezeilo, Bernice. 1978. "Validating Panga Munthu Test and Porteus Maze Test (Wooden Form) in Zambia." *International Journal of Psychology* 13 (4): 333–42.

Fagan, Joseph F., Cynthia R. Holland, and Karyn Wheeler. 2007. "The Prediction, From Infancy, of Adult IQ and Achievement." *Intelligence* 35 (3): 225–31.

Fattal, Iris, Naama Friedmann, and Aviva Fattal-Valevski. 2011. "The Crucial Role of Thiamine in the Development of Syntax and Lexical Retrieval: A Study of Infantile Thiamine Deficiency." *Brain* 134 (6): 1720–39.

Faurholt-Jepsen, Daniel, Kristina Beck Hansen, Vincent T. van Hees, Line Brinch Christensen, Tsinuel Girma, Henrik Friis, and Søren Brage. 2014. "Children Treated for Severe Acute Malnutrition Experience a Rapid Increase in Physical Activity a Few Days after Admission." *The Journal of Pediatrics* 164 (6): 1421–4.

Feng, Gary. 2011. "Eye Tracking: A Brief Guide for Developmental Researchers." *Journal of Cognition and Development* 12 (1): 1–11.

Fenson, Larry, Virginia A. Marchman, Donna J. Thal, Phillip S. Dale, J. Steven Reznick, and Elizabeth Bates. 2006. *The MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual, Second Edition*. Baltimore, MD: Paul H. Brookes Publishing.

Fernald, Lia C., Patricia Kariger, Melissa Hidrobo, and Paul Gertler. 2012. "Socioeconomic Gradients in Child Development in Very Young Children: Evidence from India, Indonesia, Peru, and Senegal." *Proceedings of the National Academy of Sciences of the United States of America* 109 (Suppl 2): 17273–80.

Fernald, Lia C., Lynnette M. Neufeld, Lauren R. Barton, Lourdes Schnaas, Juan Rivera, and Paul J. Gertler. 2006. "Parallel Deficits in Linear Growth and Mental Development in Low-Income Mexican Infants in the Second Year of Life." *Public Health Nutrition* 9 (2): 178–86.

Fernald, Lia C. H., Ann Weber, Emanuela Galasso, and Lisy Ratsifandrihamanana. 2011. "Socioeconomic Gradients and Child Development in a Very Low Income Population: Evidence from Madagascar." *Developmental Science* 14 (4): 832–47.

Fernandes, Michelle, Alan Stein, Charles R. Newton, Leila Cheikh-Ismail, Michael Kihara, Katharina Wulff, Enrique de León Quintana, Luis Aranzeta, Aureli Soria-Frisch, Javier Acedo, David Ibanez, Amina Abubakar, and Francesca Giuliani. 2014. "The INTERGROWTH-21st Project Neurodevelopment Package: A Novel Method for the Multi-Dimensional Assessment of Neurodevelopment in Pre-School Age Children." *PloS One* 9 (11): e113360.

Ferrari, Marco and Valentina Quaresima. 2012. "A Brief Review on the History of Human Functional Near-Infrared Spectroscopy (fNIRS) Development and Fields of Application." *NeuroImage* 63 (2): 921–35.

Fischer, Vinicius Jobim, Jodi Morris, and José Martines. 2014. "Developmental Screening Tools: Feasibility of Use at Primary Healthcare Level in Low- and Middle-Income Settings." *Journal of Health, Population, and Nutrition* 32 (2): 314–26.

Forssman, Linda, Per Ashorn, Ulla Ashorn, Kenneth Maleta, Andrew Matchado, Emma Kortekangas, and Jukka M. Leppänen. 2016. "Eye-Tracking-Based Assessment of Cognitive Function in Low-Resource Settings." *Archives of Disease in Childhood* 102 (4): 301–2.

Frank, Michael C., Elise Sugarman, Alexandra C. Horowitz, Molly L. Lewis, and Daniel Yurovsky. 2016. "Using Tablets to Collect Data From Young Children." *Journal of Cognition and Development* 17 (1): 1–17.

Frankenburg, William K. 1985. "The Denver Approach to Early Case Finding." In *Early Identification of Children at Risk: An International Perspective*, edited by William K. Frankenburg, Robert N. Emde, and Joseph W. Sullivan, 135–56. New York: Plenum Press.

Frankenburg, William K. and Cecilia E. Coons. 1986. "Home Screening Questionnaire: Its Validity in Assessing Home Environment." *The Journal of Pediatrics* 108 (4): 624–6.

Frankenburg, William K., Josiah B. Dodds, Philip A. Archer, Howard Shapiro, and Beverly Bresnick. 1992. "The Denver II: A Major Revision and Restandardization of the Denver Developmental Screening Test." *Pediatrics* 89 (1): 91–7.

Franzen, Michael D. 2011. “Test Construction.” In *Encyclopedia of Clinical Neuropsychology*, edited by Jeffrey S. Kreutzer, John DeLuca, and Bruce Caplan, 2489–90. New York: Springer.

Fuligni, Allison Sidle and Jeanne Brooks-Gunn. 2013. “Mother–Child Interactions in Early Head Start: Age and Ethnic Differences in Low-Income Dyads.” *Parenting: Science and Practice* 13 (1): 1–26.

Fyro, K. and G. Bodegard. 1987. “Four-Year Follow-Up of Psychological Reactions to False Positive Screening Tests for Congenital Hypothyroidism.” *Acta Pædiatrica Scandinavica* 76 (1): 107–14.

Gervain, Judit, Jacques Mehler, Janet F. Werker, Charles A. Nelson, Gergely Csibra, Sarah Lloyd-Fox, Mohinish Shukla, and Richard N. Aslin. 2011. “Near-Infrared Spectroscopy: A Report from the McDonnell Infant Methodology Consortium.” *Developmental Cognitive Neuroscience* 1 (1): 22-46.

Gesell, Arnold. 1946. “The Ontogenesis of Infant Behavior.” In *Manual of Child Psychology*, edited by Leonard Carmichael, 295–331. New York: John Wiley & Sons.

Gilkerson, Jill, Yiwen Zhang, Dongxin Xu, Jeffrey A. Richards, Xiaojuan Xu, Fan Jiang, James Harnsberger, and Keith Topping. 2015. “Evaluating Language Environment Analysis System Performance for Chinese: A Pilot Study in Shanghai.” *Journal of Speech, Language, and Hearing Research* 58 (2): 445–52.

Gladstone, Melissa, G. A. Lancaster, A. P. Jones, K. Maleta, E. Mtitimila, P. Ashorn, and R. L. Smyth. 2008. “Can Western Developmental Screening Tools Be Modified for Use in a Rural Malawian Setting?” *Archives of Disease in Childhood* 93 (1): 23–9.

Gladstone, Melissa, Gillian A. Lancaster, Eric Umar, Maggie Nyirenda, Edith Kayira, Nynke R. van den Broek, and Rosalind L. Smyth. 2010. “The Malawi Developmental Assessment Tool (MDAT): The Creation, Validation, and Reliability of a Tool to Assess Child Development in Rural African Settings.” *PLoS Medicine* 7 (5): e1000273.

Glascoe, Frances P. 2001. “Are Overreferrals on Developmental Screening Tests Really a Problem?” *Archives of Pediatrics & Adolescent Medicine* 155 (1): 54–9.

Glascoe, Frances Page. 2005. “Screening for Developmental and Behavioral Problems.” *Mental Retardation and Developmental Disabilities Research Reviews* 11 (3): 173–9.

Glascoe, Frances Page, Kevin P. Marks, and Michelle M. Macias. 2013. “Test Construction and Psychometrics, Quality Improvement and Other Research in Developmental-Behavioral Screening.” In *Identifying and Addressing Developmental Behavioral Problems: A Practical Guide for Medical and Non-Medical Professionals, Trainees, Researchers and Advocates,* ed. by Frances Page Glascoe, Kevin P. Marks, Jennifer K. Poon, and Michelle M. Macias, 424–52. Nolensville, TN: PEDStest.com.

Goodman, Robert. 2001. “Psychometric Properties of the Strengths and Difficulties Questionnaire.” *Journal of the American Academy of Child and Adolescent Psychiatry* 40 (11): 1337–45.

Gottlieb, Gilbert. 1991. “Experiential Canalization of Behavioral Development: Theory.” *Developmental Psychology* 27 (1): 4–13.

Grantham-McGregor, Sally, Yin Bun Cheung, Santiago Cueto, Paul Glewwe, Linda Richter, and Barbara Strupp. 2007. “Developmental Potential in the First 5 Years for Children in Developing Countries.” *The Lancet* 369 (9555): 60–70.

Greenfield, Patricia M., L. Monique Ward, and Jennifer Jacobs. 1997. “You Can’t Take It With You: Why Ability Assessments Don’t Cross Cultures.” *American Psychologist* 52 (10): 1115–24.

Greenwood, Charles R., Kathy Thiemann-Bourque, Dale Walker, Jay Buzhardt, and Jill Gilkerson. 2011. “Assessing Children’s Home Language Environments Using Automatic Speech Recognition Technology.” *Communication Disorders Quarterly* 32 (2): 83–92.

Griffiths, Ruth. 1984. *The Abilities of Young Children: A Comprehensive System of Mental Measurement for the First Eight Years of Life*. Henley-on-Thames, UK: ARICD.

Grigorenko, Elana L. and Robert J. Sternberg. 1999. *Assessing Cognitive Development in Early Childhood.* Department Working Paper 22927, World Bank, Washington, DC.

Guhn, Martin, Anne Gadermann, and Bruno D. Zumbo. 2007. “Does the EDI Measure School Readiness in the Same Way Across Different Groups of Children?” *Early Education and Development* 18 (3): 453–72.

Gwet, Kilem L. 2014. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters (Fourth Edition)*. Gaithersburg, MD: Advanced Analytics.

Hackman, Daniel A., Robert Gallop, Gary W. Evans, and Martha J. Farah. 2015. "Socioeconomic Status and Executive Function: Developmental Trajectories and Mediation." *Developmental Sci*ence 18 (5): 686–702.

Haeussler, Isabel Margarita and Teresa Marchant. 1980. *TEPSI test de desarrollo psicomotor 2-5 años (TEPSI test of psychomotor development 2-5 years). 8 ed.* Santiago, Chile: Ediciones Universidad Católica.

Hagie, Marilyn Urquhart, Peggy L. Gallipo, and Lana Svien. 2003. "Traditional Culture Versus Traditional Assessment for American Indian Students: An Investigation of Potential Test Item Bias." *Assessment for Effective Intervention* 29 (1): 15–25.

Halle, Tamara G., Elizabeth C. Hair, Margaret Burchinal, Rachel Anderson, and Martha Zaslow. 2012. *In the Running for Successful Outcomes: Exploring the Evidence for Thresholds of School Readiness.* Technical report, Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services, Washington DC. http://aspe.hhs.gov/pdf-report/ running-successful-outcomes-exploring-evidence-thresholds-school-readinesstechnical-report.

Hallgren, Kevin A. 2012. "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial." *Tutorials in Quantitative Methods for Psychology* 8 (1): 23–34.

Hamadani, Jena D., Helen Baker-Henningham, Fahmida Tofail, Fardina Mehrin, Syed N. Huda, and Sally M. Grantham-McGregor. 2010."Validity and Reliability of Mothers' Reports of Language Development in 1-Year-Old Children in a Large-Scale Survey in Bangladesh." *Food and Nutrition Bulletin* 31 (2): S198–206.

Hamadani, Jena Derakhshani, Fahmida Tofail, Tim Cole, and Sally Grantham-McGregor. 2013. "The Relation Between Age of Attainment of Motor Milestones and Future Cognitive and Motor Development in Bangladeshi Children." *Maternal & Child Nutrition* 9 (S1): 89–104.

Hamadani, Jena D., Fahmida Tofail, Afroza Hilaly, Syed N. Huda, Patrice L. Engle, and Sally M. Grantham-McGregor. 2010. "Use of Family Care Indicators and Their Relationship with Child Development in Bangladesh." *Journal of Health, Population and Nutrition* 28 (1): 23–33.

Hamadani, Jena D., Fahmida Tofail, Syed N. Huda, Dewan S. Alam, Deborah A. Ridout, Orazio Attanasio, and Sally M. Grantham-McGregor. 2014. "Cognitive Deficit and Poverty in the First 5 Years of Childhood in Bangladesh." *Pediatrics* 134 (4): e1001–8.

Hambleton, Ronald K. and Liane Patsula. 1998. "Adapting Tests for Use in Multiple Languages and Cultures." *Social Indicators Research* 45 (1–3): 153–71.

Hamoudi, Amar and Margaret Sheridan. 2015. *Unpacking the Black Box of Cognitive Ability: A Novel Tool for Assessment in a Population-Based Survey.* Manuscript under review, Young Lives Study.

Hamre, Bridget K., Robert C. Pianta, Jason T. Downer, Jamie DeCoster, Andrew J. Mashburn, Stephanie M. Jones, Joshua L. Brown, Elise Cappella, Marc Atkins, Susan E. Rivers, Marc A. Brackett, and Aki Hamagami. 2013. "Teaching Through Interactions: Testing a Developmental Framework of Teacher Effectiveness in over 4,000 Classrooms." *Elementary School Journal* 113 (4): 461–87.

Handal, Alexis J., Betsy Lozoff, Jaime Breilh, and Siobán D. Harlow. 2007. "Effect of Community of Residence on Neurobehavioral Development in Infants and Young Children in a Flower-Growing Region of Ecuador." *Environmental Health Perspectives* 115 (1): 128–33.

Hanson, Jamie L., Brendon M. Nacewicz, Matthew J. Sutterer, Amelia A. Cayo, Stacey M. Schaefer, Karen D. Rudolph, Elizabeth A. Shirtcliff, Seth D. Pollak, and Richard J. Davidson. 2015. "Behavioral Problems After Early Life Stress: Contributions of the Hippocampus and Amygdala." *Biological Psychiatry* 77 (4): 314–23.

Harkness, Sara and Charles M. Super. 1977. "Why African Children Are So Hard to Test." *Annals of the New York Academy of Sciences* 285 (1): 326–31.

Harms, Thelma, Debby Cryer, and Richard M. Clifford. 2006. *Infant/Toddler Environment Rating Scale, Revised Edition*. Chapel Hill: Teachers College Press.

Harris, Dale B. 1963. *Children's Drawings as Measures of Intellectual Maturity: A Revision and Extension of the Goodenough Draw-a-Man Test*. New York: Harcourt, Brace & World.

Hart, Betty and Todd R. Risley. 1992. "American Parenting of Language-Learning Children: Persisting Differences in Family-Child Interactions Observed in Natural Home Environments." *Developmental Psychology* 28 (6): 1096–105.

Hart, Betty and Todd R. Risley. 1995. "42 American Families." In *Meaningful Differences in the Everyday Experience of Young American Children*, by Betty Hart and Todd R. Risley, 53–74. Baltimore, MD: Paul H. Brookes Publishing.

Hart, Betty and Todd R. Risley. 2003. "The Early Catastrophe: The 30 Million Word Gap By Age 3." *American Educator* 27 (1): 4–9.

Hauglann, Lisbeth, Bjørn Helge Handegaard, Stein Erik Ulvund, Marianne Nordhov, John A. Rønning, and Per Ivar Kaaresen. 2015. "Cognitive Outcome of Early Intervention in Preterms at 7 and 9 Years of Age: A Randomised Controlled Trial." *Archives of Disease in Childhood. Fetal and Neonatal Edition.* 100 (1): F11–6.

Heo, Kay H., Jane Squires, and Paul Yovanoff. 2008. "Cross-Cultural Adaptation of a Pre-School Screening Instrument: Comparison of Korean and US Populations." *Journal of Intellectual Disability Research* 52 (3): 195–206.

Hoff, Erika. 2003. "The Specificity of Environmental Influence: Socioeconomic Status Affects Early Vocabulary Development Via Maternal Speech." *Child Development* 74 (5): 1368–78.

Hohm, Erika, Christine Jennen-Steinmetz, Martin H. Schmidt, and Manfred Laucht. 2007. "Language Development at Ten Months. Predictive of Language Outcome and School Achievement Ten Years Later?" *European Child & Adolescent Psychiatry* 16 (3): 149–56.

Holding, Penny A., H. Gerry Taylor, Sidi D. Kazungu, Thadeaus Mkala, Joseph Karisa Gona, Bernard Mwamuye, Leonard Mbonani, and Jim Stevenson. 2004. "Assessing Cognitive Outcomes in a Rural African Population: Development of a Neuropsychological Battery in Kilifi District, Kenya." *Journal of the International Neuropsychological Societ*y 10 (2): 246–60.

Holmqvist, Kenneth, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost van de Weijer. 2011. *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford: Oxford University Press.

Hongwanishkul, Donaya, Keith R. Happaney, Wendy S. C. Lee, and Philip David Zelazo. 2005. "Assessment of Hot and Cool Executive Function in Young Children: Age-Related Changes and Individual Differences." *Developmental Neuropsychology* 28 (2): 617–44.

Howard, Douglas P. and Marisol N. de Salazar. 1984. "Language and Cultural Differences in the Administration of the Denver Developmental Screening Test." *Child Study Journal* 14 (1): 1–9.

Howes, Carollee, Margaret Burchinal, Robert Pianta, Donna Bryant, Diane Early, Richard Clifford, and Oscar Barbarin. 2008. "Ready to Learn? Children's Pre-Academic Achievement in Pre-Kindergarten Programs." *Early Childhood Research Quarterly* 23 (1): 27–50.

Ispa, Jean M., Annamaria Csizmadia, Duane Rudy, Mark A. Fine, Jennifer L. Krull, Robert H. Bradley, and Natasha Cabrera. 2013. "Patterns of Maternal Directiveness by Ethnicity Among Early Head Start Research Participants." *Parenting: Science and Practice* 13 (1): 58–75.

Jaffari-Bimmel, Nicole, Femmie Juffer, Marinus H. van IJzendoorn, Marian J. Bakermans-Kranenburg, and Ab Mooijaart. 2006. "Social Development from Infancy to Adolescence: Longitudinal and Concurrent Factors in an Adoption Sample." *Developmental Psychology* 42 (6): 1143–53.

Janus, Magdalena and David R. Offord. 2007. "Development and Psychometric Properties of the Early Development Instrument (EDI): A Measure of Children's School Readiness." *Canadian Journal of Behavioural Science* 39 (1): 1–22.

Janus, Magdalena and Caroline Reid-Westoby. 2016. "Monitoring the Development of All Children: The Early Development Instrument." *Early Childhood Matters* 125: 40–5.

Johnson, Samantha and Neil Marlow. 2006. "Developmental Screen or Developmental Testing?" *Early Human Development* 82 (3): 173–83.

Jones, Stephanie M., Martha Zaslow, Kristen E. Darling-Churchill, and Tamara G. Halle. 2016. "Assessing Early Childhood Social and Emotional Development: Key Conceptual and Measurement Issues." *Journal of Applied Developmental Psychology* 45: 42–8.

Jurado, Maria-Beatriz and Monica Rosselli. 2007. "The Elusive Nature of Executive Functions: A Review of Our Current Understanding." *Neuropsychology Review* 17 (3): 213–33.

Kagan, Sharon Lynn and Pia Rebello Britto. 2005. *Going Global with Early Learning and Development Standards: Final Report*. A UNICEF Final Report. New York: Columbia University National Center for Children and Families.

Kagan, Sharon Lynn, Pia Rebello Britto, and Patrice L. Engle. 2005. *Going Global with Early Learning and Development Standards: Final Report*. A UNICEF 87 (3): 205–8. SAME? CANT FIND THIS W/ ENGLE?

Kagan, Sharon Lynn, Evelyn Moore, and Sue Bredekamp, eds. 1995. *Reconsidering Children's Early Learning and Development: Toward Common Views and Vocabulary: Report of the National Education Goals Panel, Goal 1 Technical Planning Group, No. ED 391 576*. Washington, DC: U.S. Government Printing Office.

Kagitcibasi, Cigdem, Diane Sunar, and Sevda Bekman. 2001. "Long-Term Effects of Early Intervention: Turkish Low-Income Mothers and Children." *Journal of Applied Developmental Psychology* 22 (4): 333–61.

Karatekin, Canan. 2007. "Eye Tracking Studies of Normative and Atypical Development." *Developmental Review* 27 (3): 283–348.

Kariger, Patricia, Edward A. Frongillo, Patrice Engle, Pia M. Rebello Britto, Sara M. Sywulka, and Purnima Menon. 2012. "Indicators of Family Care for Development for Use in Multicountry Surveys." *Journal of Health, Population and Nutrition* 30 (4): 472–86.

Kariger, Patricia K., Rebecca J. Stoltzfus, Deanna Olney, Sunil Sazawal, Robert Black, James M. Tielsch, Edward A. Frongillo, Sabra S. Khalfan, and Ernesto Pollitt. 2005. "Iron Deficiency and Physical Growth Predict Attainment of Walking But Not Crawling in Poorly Nourished Zanzibari Infants." *The Journal of Nutrition* 135 (4): 814–9.

Kathuria, Ravinder and Robert Serpell. 1998. "Standardization of the Panga Munthu Test—A Nonverbal Cognitive Test Developed in Zambia." *The Journal of Negro Education* 67 (3): 228–41.

Katzmarzyk, Peter T., Tiago V. Barreira, Stephanie T. Broyles, Catherine M. Champagne, Jean-Phillippe Chaput, Mikael Fogelholm, Gang Hu, William D. Johnson, Rebecca Kuriyan, Anura Kurpad, Estelle Victoria Lambert, Carol Maher, Jose Maia, Victor Matsudo, Tim Olds, Vincent Ochieng Onywera, Olga L. Sarmiento, Martyn Standage, Mark S. Tremblay, Catrine Tudor-Locke, Pei Zhao, and Timothy S. Church. 2015. "Physical Activity, Sedentary Time, and Obesity in an International Sample of Children." "Physical Activity, Sedentary Time, and Obesity in an International Sample of Children." *Medicine & Science in Sports & Exercise* 47 (10): 2062–9.

Kesiktas, Ayse Dolunay, Nimet Bulbin Sucuoglu, Bahar Keceli-Kaysili, Selma Akalin, Gozde Gul, and Binnur Yildirim. 2009. "The Home Environments of Young Children With and Without Disabilities." *Infants & Young Children* 22 (3): 201–10.

Khan, Naila Zaman, Humaira Muslima, Asma Begum Shilpi, Dilara Begum, Monowara Parveen, Nasima Akter, Shamim Ferdous, Kamrun Nahar, Helen McConachie, and Gary L. Darmstadt. 2013. "Validation of Rapid Neurodevelopmental Assessment for 2- to 5-Year-Old Children in Bangladesh." *Pediatrics* 131 (2): e486–94.

Khan, Naila Zaman, Humaira Muslima, Dilara Begum, Asma Begum Shilpi, Selina Akhter, Khaleda Bilkis, Nasreen Begum, Monowara Parveen, Shamim Ferdous, Romella Morshed, Maneesh Batra, and Gary L. Darmstadt. 2010. "Validation of Rapid Neurodevelopmental Assessment Instrument for Under-Two-Year-Old Children in Bangladesh." *Pediatrics* 125 (4): e755–62.

Kihara, Michael, Michelle de Haan, Harrun H. Garrashi, Brian G. R. Neville, and Charles R. J. C. Newton. 2010. "Atypical Brain Response to Novelty in Rural African Children with a History of Severe Falciparum Malaria." *Journal of the Neurological Sciences* 296 (1–2): 88–95.

Kim, Jim Yong. 2016. "Remarks by World Bank Group President Jim Yong Kim at the Early Childhood Development Event." Speech delivered at the World Bank-International Monetary Fund Spring Meetings, Early Childhood Development event, Washington, DC.

Kim, Youngwon, Michael W. Beets, and Gregory J. Welk. 2012. "Everything You Wanted to Know About Selecting the 'Right' Actigraph Accelerometer Cut-Points for Youth, But...: A Systematic Review." *Journal of Science and Medicine in Sport* 15 (4): 311–21.

Kishiyama, Mark M., W. Thomas Boyce, Amy Marie Jimenez, Lee M. Perry, and Robert T. Knight. 2009. "Socioeconomic Disparities Affect Prefrontal Function in Children." *Journal of Cognitive Neuroscience* 21 (6): 1106–15.

Knauer, Heather A., Rose M. C. Kagawa, Armando Garcia-Guerra, Lourdes Schnass, Lynette M. Neufeld, and Lia C. H. Fernald. 2016. "Pathways to Improved Development for Children Living in Poverty: A Randomized Effectiveness Trial in Rural Mexico." *International Journal of Behavioral Development* 40 (6): 492–99.

Kochanska, Grazyna, Kathleen T. Murray, and Elena T. Harlan. 2000. "Effortful Control in Early Childhood: Continuity and Change, Antecedents, and Implications for Social Development." *Developmental Psychology* 36 (2): 220–32.

Kovas, Yulia, Claire M. A. Haworth, Philip S. Dale, and Robert Plomin. 2007. "The Genetic and Environmental Origins of Learning Abilities and Disabilities in the Early School Years." *Monographs of the Society for Research in Child Development* 72 (3): 1–144.

Kramer, Joel H., Dan Mungas, Katherine L. Possin, Katherine P. Rankin, Adam L. Boxer, Howard J. Rosen, Alan Bostrom, Lena Sinha, Ashley Berhel, and Mary Widmeyer. 2014. "NIH EXAMINER: Conceptualization and Development of an Executive Function Battery." *Journal of the International Neuropsychological Society* 20 (1): 11–9.

Kuklina, Elena V., Usha Ramakrishnan, Aryeh D. Stein, Huiman H. Barnhart, and Reynaldo Martorell. 2004. "Growth and Diet Quality Are Associated with the Attainment of Walking in Rural Guatemalan Infants." *The Journal of Nutrition* 134 (12): 3296–300.

La Paro, Karen M. and Robert C. Pianta. 2000. "Predicting Children's Competence in the Early School Years: A Meta-Analytic Review." *Review of Educational Research* 70 (4): 443–84.

Lai, Meng-Lung, Meng-Jung Tsai, Fang-Ying Yang, Chung-Yuan Hsu, Tzu-Chien Liu, Silvia Wen-Yu Lee, Min-Hsien Lee, Guo-Li Chiou, Jyh-Chong Liang, and Chin-Chung Tsai. 2013. "A Review of Using Eye-Tracking Technology in Exploring Learning from 2000 to 2012." *Educational Research Review* 10: 90–115.

Lansdown, Rachael G., Harvey Goldstein, Pankaj Manibhai Shah, John H. Orley, Giovanna Di, Kanwar K. Kaul, Vineet Kumar, Udom Laksanavicharn, and Vangala Reddy. 1996. "Culturally Appropriate Measures for Monitoring Child Development at Family and Community Level: A WHO Collaborative Study." *Bulletin of the World Health Organization* 74 (3): 283–90.

Larson, Leila M., Melisssa F. Young, Usha Ramakrishnan, Amy Webb Girard, Pankaj Verma, Indrajit Chaudhuri, Sridhar Srikantiah, and Reynaldo Martorell. 2017. "A Cross-Sectional Survey in Rural Bihar, India, Indicates That Nutritional Status, Diet, and Stimulation Are Associated with Motor and Mental Development in Young Children." *The Journal of Nutrition* 147 (8): 1578–85.

Lazar, Irving, Richard Darlington, Harry Murray, Jacqueline Royce, Ann Snipper, and Craig T. Ramey. 1982. "Lasting Effects of Early Education: A Report from the Consortium for Longitudinal Studies." *Monographs of the Society for Research in Child Development* 47 (2/3).

Lee, I-Min and Eric J. Shiroma. 2014. "Using Accelerometers to Measure Physical Activity in Large-Scale Epidemiologic Studies: Issues and Challenges." *British Journal of Sports Medicine* 48 (3): 197–201.

Lee, Valerie E., J. Brooks-Gunn, Elizabeth Schnur, and Fong-Ruey Liaw. 1990. "Are Head Start Effects Sustained? A Longitudinal Follow-Up Comparison of Disadvantaged Children Attending Head Start, No Preschool, and Other Preschool Programs." *Child Development* 61 (2): 495–507.

Leerkes, Esther M. and Susan C. Crockenberg. 2003. "The Impact of Maternal Characteristics and Sensitivity on the Concordance Between Maternal Reports and Laboratory Observations of Infant Negative Emotionality." *Infancy* 4 (4): 517–39.

Lewis, Charlie, Masuo Koyasu, Seungmi Oh, Ayako Ogawa, Benjamin Short, and Zhao Huang. 2009. "Culture, Executive Function, and Social Understanding." In *Social Interaction and the Development of Executive Function: New Directions for Child and Adolescent Development, Number 123,* edited by Charlie Lewis and Jeremy I. M. Carpendale, 69–85. New York: John Wiley & Sons.

Leyva, Diana, Christina Weiland, M. Barata, Hirokazu Yoshikawa, Catherine Snow, Ernesto Treviño, and Andrea Rolla. 2015. "Teacher–Child Interactions in Chile and Their Associations With Prekindergarten Outcomes." *Child Development* 86 (3): 781–99.

Liew, Jeffrey. 2011. "Effortful Control, Executive Functions, and Education: Bringing Self-Regulatory and Social-Emotional Competencies to the Table." *Child Development Perspectives* 6 (2): 105–11.

Lloyd-Fox, Sarah, A. Blasi, C. E. Elwell, T. Charman, D. Murphy, and M. H. Johnson. 2013. "Reduced Neural Sensitivity to Social Stimuli in Infants at Risk for Autism. *Proceedings of the Royal Society B: Biological Sciences* 280 (1758): 20123026.

Lloyd-Fox, Sarah, M. Papademetriou, M. K. Darboe, N. L. Everdell, R. Wegmuller, A. M. Prentice, S. E. Moore, and C. E. Elwell. 2014. "Functional Near Infrared Spectroscopy (fNIRS) to Assess Cognitive Function in Infants in Rural Africa." *Scientific Reports* 2014 (4): 4740.

López Boo, Florencia. 2016. "Socio-Economic Status and Early Childhood Cognitive Skills: A Mediation Analysis Using the Young Lives Panel." *International Journal of Behavioral Development* 1–9.

López Boo, Florencia, María Caridad Araujo, and Romina Tomé. 2016. *How is Child Care Quality Measured? A Toolkit*. Washington, DC: Inter-American Development Bank. https://publications.iadb.org/bitstream/handle/11319/7432/How-is-child-care-quality-measured.pdf?sequence=4.

Loprinzi, Paul D. and Bradley J. Cardinal. 2011. "Measuring Children's Physical Activity and Sedentary Behaviors." *Journal of Exercise Science & Fitness* 9 (1): 15–23.

Lukowski, Angela F., Marlene Koss, Matthew J. Burden, John Jonides, Charles A. Nelson, Niko Kaciroti, Elias Jimenez, and Betsy Lozoff. 2010. "Iron Deficiency in Infancy and Neurocognitive Functioning at 19 Years: Evidence of Long-Term Deficits in Executive Function and Recognition Memory." *Nutritional Neuroscience* 13 (2): 54–70.

Lupien, Sonia J., Bruce S. McEwen, Megan R. Gunnar, and Christine Heim. 2009. "Effects of Stress Throughout the Lifespan on the Brain, Behaviour and Cognition." *Nature Reviews Neuroscience* 10 (6): 434–45.

Malda, Maike, Fons J. R. van de Vijver, Krishnamachari Srinivasan, Catherine Transler, Prathima Sukumar, and Kirthi Rao. 2008. "Adapting a Cognitive Test for a Different Culture: An Illustration of Qualitative Procedures." *Psychology Science Quarterly* 50 (4): 451–68.

Malmberg, Lars-Erik, Peter Mwaura, and Kathy Sylva. 2011. "Effects of a Preschool Intervention on Cognitive Development Among East-African Preschool Children: A Flexibly Time-Coded Growth Model." *Early Childhood Research Quarterly* 26 (1): 124–33.

Marks, Kevin, Frances Page Glascoe, Glen P. Aylward, Michael I. Shevell, Paul H. Lipkin, and Jane K. Squires. 2008. "The Thorny Nature of Predictive Validity Studies on Screening Tests for Developmental-Behavioral Problems." *Pediatrics* 122 (4): 866–8.

Matias, Susan L., Malay K. Mridha, Fahmida Tofail, Charles D. Arnold, Showcat A. Khan, Zakia Siddiqui, Barkat Ullah, and Kathryn G. Dewey. 2017. "Home Fortification During the First 1000 Days Improves Child Development in Bangladesh: A Cluster-Randomized Effectiveness Trial." *The American Journal of Clinical Nutrition* 105 (4): 958–69.

McCall, Robert B. 1981. "Nature-Nurture and the Two Realms of Development: A Proposed Integration with Respect to Mental Development." *Child Development* 52 (1): 1–12.

McCormick, Marie C., Jeanne Brooks-Gunn, Stephen L. Buka, Julie Goldman, Jennifer Yu, Mikhail Salganik, David T. Scott, Forrest C. Bennett, Libby L. Kay, Judy C. Bernbaum, Charles R. Bauer, Camilia Martin, Elizabeth R. Woods, Anne Martin, and Patrick H. Casey. 2006. "Early Intervention in Low Birth Weight Premature Infants: Results at 18 Years of Age for the Infant Health and Development Program." *Pediatrics* 117 (3): 771–80.

McCoy, Dana Charles, Evan D. Peet, Majid Ezzati, Goodarz Danaei, Maureen M. Black, Christopher R. Sudfeld, Wafaie Fawzi, and Günther Fink. 2016. "Early Childhood Developmental Status in Low- and Middle-Income Countries: National, Regional, and Global Prevalence Estimates Using Predictive Modeling." *PLoS Medicine* 13 (6): e1002034.

McCoy, Dana Charles, Christopher R. Sudfeld, David C. Bellinger, Alfa Muhihi, Geofrey Ashery, Taylor E. Weary, Wafaie Fawzi, and Günther Fink. 2017. "Development and Validation of an Early Childhood Development Scale for Use in Low-Resourced Settings." *Population Health Metrics* 15 (1): 3.

McLaughlin, Katie A., Margaret A. Sheridan, Florin Tibu, Nathan A. Fox, Charles H. Zeanah, and Charles A. Nelson III. 2015. "Causal Effects of the Early Caregiving Environment on Development of Stress Response Systems in Children." *Proceedings of the National Academy of Sciences of the United States of America* 112 (18): 5637–42.

Mischel, Walter and David Brooks. 2011. "The News From Psychological Science: A Conversation Between David Brooks and Walter Mischel." *Perspectives on Psychological Science* 6 (6): 515–20.

Mischel, Walter, Yuichi Shoda, and Philip K. Peake. 1988. "The Nature of Adolescent Competencies Predicted by Preschool Delay of Gratification." *Journal of Personality and Social Psychology* 54 (4): 687–96.

Moffitt, Terrie E., Louise Arseneault, Daniel Belsky, Nigel Dickson, Robert J. Hancox, HonaLee Harrington, Renate Houts, Richie Poulton, Brent W. Roberts, Stephen Ross, Malcolm R. Sears, W. Murray Thomson, and Avshalom Caspi. 2011. "A Gradient of Childhood Self-Control Predicts Health, Wealth, and Public Safety." *Proceedings of the National Academy of Sciences of the United States of America* 108 (7): 2693–8.

Montie, Jeanne E., Zongping Xiang, and Lawrence J. Schweinhart. 2006. "Preschool Experience in 10 Countries: Cognitive and Language Performance at Age 7." *Early Childhood Research Quarterly* 21 (3): 313–31.

Mott, Frank L. 2004. "The Utility of the HOME-SF Scale for Child Development Research in a Large National Longitudinal Survey: The National Longitudinal Survey of Youth 1979 Cohort." *Parenting: Science and Practice* 4 (2-3): 259–70.

Mulatu, Mesfin Samuel. 1995. "Prevalence and Risk Factors of Psychopathology in Ethiopian Children." *Journal of the American Academy of Child and Adolescent Psychiatry* 34 (1): 100–9.

Mung'ala-Odera, Victor, Rose Meehan, P. Njuguna, Neema Mturi, K. J. Alcock, J. A. Carter, and Charles Newton. 2004. "Validity and Reliability of the 'Ten Questions' Questionnaire for Detecting Moderate to Severe Neurological Impairment in Children Aged 6-9 Years in Rural Kenya." *Neuroepidemiology* 23 (1–2): 67–72.

Mustard, J. Fraser and Mary Eming Young. 2007. "Measuring Child Development to Leverage ECD Policy and Investment." In *From Measurement to Action*, edited by Mary Eming Young and Linda M. Richardson, 253–91. Washington, DC: World Bank.

Mwamwenda, Tuntufye S. and Bernadette B. Mwamwenda. 1990. "Assessing Africans' Cognitive Development: Judgement Versus Judgement Plus Explanation." *The Journal of Genetic Psychology* 151 (2): 245–54.

Nair, M., G. L. Prasanna, Lakshmanan Jeyaseelan, Babu George, V. R. Resmi, and R. M. Sunitha. 2009. "Validation of Home Screening Questionnaire (HSQ) Against Home Observation for the Measurement of Environment (HOME)." *Indian Pediatrics* 46 (Suppl): s55–8.

Nampijja, Margaret, Barbara Apule, Swaib Lule, Hellen Akurut, Lawrence Muhangi, Emily L. Webb, Charlie Lewis, Alison M. Elliott, and Katie J. Alcock. 2012. "Effects of Maternal Worm Infections and Anthelminthic Treatment During Pregnancy on Infant Motor and Neurocognitive Functioning." *Journal of the International Neuropsychological Society* 18 (6): 1019–30.

Nelson, Charles A. 2016. *Brain Imaging as a Measure of Future Cognitive Function in Children*. Boston, MA: Boston Children's Hospital: Research and Innovation.

Nelson, Charles A. and Joseph P. McCleery. 2008. "Use of Event-Related Potentials in the Study of Typical and Atypical Development." *Journal of the American Academy of Child and Adolescent Psychiatry* 47 (11): 1252–61.

Nerlove, Marc. 1974. "Household and Economy: Toward a New Theory of Population and Economic Growth." *Journal of Political Economy* 82 (2): S200–18.

Neville, H. J., et al. 2013. "Family-Based Training Program Improves Brain Function, Cognition, and Behavior in Lower Socioeconomic Status Preschoolers." *Proceedings of the National Academy of Sciences of the United States of America* 110 (29): 12138–43.

NICHD Early Child Care Research Network. 2002. "Child-Care Structure, Process, Outcome: Direct and Indirect Effects of Child-Care Quality on Young Children's Development." *Psychological Science* 13 (3): 199–206.

NICHD Early Child Care Research Network. 2003. "Do Children's Attention Processes Mediate the Link Between Family Predictors and School Readiness?" *Developmental Psychology* 39 (3): 581–93.

Nkhoma, Owen W. W., Maresa E. Duffy, Deborah A. Cory-Slechta, Philip W. Davidson, Emeir M. McSorley, J. J. Strain, and Gerard M. O'Brien. 2013. "Early-Stage Primary School Children Attending a School in the Malawian School Feeding Program (SFP) Have Better Reversal Learning and Lean Muscle Mass Growth Than Those Attending a Non-SFP School." *The Journal of Nutrition* 143 (8): 1324–30.

Noble, Kimberly G., Suzanne M. Houston, Natalie H. Brito, Hauke Bartsch, Eric Kan, Joshua M. Kuperman, Natacha Akshoomoff, David G. Amaral, Cinnamon S. Bloss, Ondrej Libiger, Nicholas J. Schork, Sarah S. Murray, B. J. Casey, Linda Chang, Thomas M. Ernst, Jean A. Frazier, Jeffrey R. Gruen, David N. Kennedy, Peter Van Zijl, Stewart Mostofsky, Walter E. Kaufmann, Tai Kenet, Anders M. Dale, Terry L. Jernigan, and Elizabeth R. Sowell. 2015. "Family Income, Parental Education and Brain Structure in Children and Adolescents." *Nature Neuroscience* 18 (5): 773–8.

Noble, Kimberly G., Bruce D. McCandliss, and Martha J. Farah. 2007. "Socioeconomic Gradients Predict Individual Differences in Neurocognitive Abilities." *Developmental Science* 10 (4): 464–80.

Noble, Kimberly G., Nim Tottenham, and B.J. Casey. 2005. "Neuroscience Perspectives on Disparities in School Readiness and Cognitive Achievement." *The Future of Children* 15 (1): 71–89.

Obradović, Jelena. 2016. "Physiological Responsivity and Executive Functioning: Implications for Adaptation and Resilience in Early Childhood." *Child Development Perspectives* 10 (1): 65–70.

Ogunnaike, Oluyomi A. and Robert F. Houser, Jr. 2002. "Yoruba Toddlers' Engagement in Errands and Cognitive Performance on the Yoruba Mental Subscale." *International Journal of Behavioral Development* 26 (2): 145–53.

Oiberman, Alicia. 2006. "Resiliencia y factores de protección en bebés vulnerables. Aplicación de la Escala Argentina de Inteligencia Sensoriomotriz." *Acta Psiquiátrica y Psicológica de America Latina* 52 (1): 19–25.

Oiberman, Alicia, Orellana L., Mansilla, M. 2005. "Evaluación de la inteligencia en bebés argentinos: Escala Argentina de Inteligencia Sensoriomotriz." *Revista Argentina de Clínica Psicológica* 14 (3): 213–8.

Pavlakis, Alexandra E., Kimberly Noble, Steven G. Pavlakis, Noorjahan Ali, and Yitzchak Frank. 2015. "Brain Imaging and Electrophysiology Biomarkers: Is There a Role in Poverty and Education Outcome Research?" *Pediatric Neurology* 52 (4): 383–8.

Paxson, Christina and Nobert Schady. 2007. "Cognitive Development Among Young Children in Ecuador: The Roles of Wealth, Health, and Parenting." *Journal of Human Resources* 42 (1): 49–84.

Payne, Kay T. and Orlando L. Taylor. 2002. "Multicultural Influences on Human Communication" In *Human Communication Disorders: An Introduction (6th ed)*, edited by George H. Shames and Norma B. Anderson. Boston, MA: Allyn & Bacon.

Peña, Elizabeth D. 2007. "Lost in Translation: Methodological Considerations in Cross-Cultural Research." *Child Development* 78 (4): 1255–64.

Pianta, Robert C., Karen M. La Paro, and Bridge K. Hamre. 2008. *Classroom Assessment Scoring System (CLASS)*. Baltimore, MD: Paul H. Brookes Publishing.

Pisani, Lauren, Ivelina Borisova, and Amy Jo Dowd. 2015. "International Development and Early Learning Assessment Technical Working Paper." Working Paper, Save the Children, London.

Plomin, Robert and Ian J. Deary. 2015. "Genetics and Intelligence Differences: Five Special Findings." *Molecular Psychiatry* 20 (1): 98–108.

Pollitt, E. 2001. “The Developmental and Probabilistic Nature of the Functional Consequences of Iron-Deficiency Anemia in Children.” *Journal of Nutrition* 131 (2): 669S–75S.

Pollitt, Ernesto and Nina Triana. 1999. “Stability, Predictive Validity, and Sensitivity of Mental and Motor Development Scales and Pre-School Cognitive Tests Among Low-Income Children in Developing Countries.” *Food and Nutrition Bulletin* 20 (1): 45–52.

Posner, Michael I., Mary K. Rothbart, Brad E. Sheeshe, and Pascale Voelker. 2014. “Developing Attention: Behavioral and Brain Mechanisms.” *Advances in Neuroscience (Hindawi)* 2014: 405094.

Prado, Elizabeth L., Amina A. Abubakar, Souheila Abbeddou, Elizabeth Y. Jimenez, Jérôme W. Somé, and Jean-Bosco Ouédraogo. 2014. “Extending the Developmental Milestones Checklist for Use in a Different Context in Sub-Saharan Africa.” *Acta Paediatrica* 103 (4): 447–54.

Prado, Elizabeth L., Seth Adu-Afarwuah, Anna Lartey, Maku Ocansey, Per Ashorn, Steve A. Vosti, and Kathryn G. Dewey. 2016. “Effects of Pre- and Post-Natal Lipid-Based Nutrient Supplements on Infant Development in a Randomized Trial in Ghana.” *Early Human Development* 99: 43–51.

Prado, Elizabeth L. and Kathryn G. Dewey. 2014. “Nutrition and Brain Development in Early Life.” *Nutrition Reviews* 72 (4): 267–84.

Prado, Elizabeth L., Sri Hartini, Atik Rahmawati, Elfa Ismayani, Astri Hidayati, Nurul Hikmahym Husni Muadz, Mandri S. Apriatni, Michael T. Ullman, Anuraj H. Shankar, and Katherine J. Alcock. 2010. “Test Selection, Adaptation, and Evaluation: A Systematic Approach to Assess Nutritional Influences on Child Development in Developing Countries.” *The British Journal of Educational Psychology*. 80 (Pt 1): 31–53.

Prado, Elizabeth L., Kenneth Maleta, Per Ashorn, Ulla Ashorn, Steve A. Vosti, John Sadalaki, and Kathryn G. Dewey. 2016. “Effects of Maternal and Child Lipid-Based Nutrient Supplements on Infant Development: A Randomized Trial in Malawi.” *The American Journal of Clinical Nutrition* 103 (3): 784–93.

Prado, Elizabeth L., John C. Phuka, Kenneth Maleta, Per Ashorn, Ulla Ashorn, Steve A. Vosti, and Kathryn G. Dewey. 2016. “Provision of Lipid-Based Nutrient Supplements from Age 6 to 18 Months Does Not Affect Infant Development Scores in a Randomized Trial in Malawi.” *Maternal and Child Health Journal* 20 (10): 2199–208.

Prado, Elizabeth L., Suzy K. Sebayang, Mandri Apriatni, Sita R. Adawiyah, Nina Hidayati, Ayuniarti Islamiyah, Sudirman Siddiq, Benyamin Harefa, Jarrad Lum, Katherine J. Alcock, Michael T. Ullman, Husni Muadz, and Anuraj H. Shankar. 2017. “Maternal Multiple Micronutrient Supplementation and Other Biomedical and Socioenvironmental Influences on Children’s Cognition at Age 9-12 Years in Indonesia: Follow-up of the SUMMIT Randomised Trial.” *The Lancet. Global Health* 5 (2): e 217–28.

Pulakka, Anna, Ulla Ashorn, Y. B. Cheung, Kathryn G. Dewey, Kenneth Maleta, Stephen A. Vosti, and Per Ashorn. 2015. “Effect of 12-Month Intervention with Lipid-Based Nutrient Supplements on Physical Activity of 18-Month-Old Malawian Children: A Randomised, Controlled Trial.” *European Journal of Clinical Nutrition* 69 (2): 173–8.

Pulakka, Anna, Y. B. Cheung, Ulla Ashorn, V. Penpraze, Kenneth Maleta, John C. Phuka, and Per Ashorn. “Feasibility and Validity of the ActiGraph GT3X Accelerometer in Measuring Physical Activity of Malawian Toddlers.” *Acta Paediatrica* 102 (12): 1192–8.

Radach, Ralph, Alan Kennedy, and Keith Rayner. 2004. *Eye Movements and Information Processing During Reading*. New York: Psychology Press.

Rao, Nirmala and Emma Pearson. 2007. *An Evaluation of Early Childhood Care and Education Programmes in Cambodia*. Unpublished manuscript. http://www.unicef.org/evaldatabase/files/CBD_early_childhoodcare_evaluation.pdf.

Rao, Nirmala, Jin Sun, Eva E. Chen, and Patrick Ip. 2017. “Effectiveness of Early Childhood Interventions in Promoting Cognitive Development in Developing Countries: A Systematic Review and Meta-Analysis.” *Hong Kong Journal of Paediatrics (New Series)* 22 (1): 14–25.

Rao, Nirmala, Jin Sun, Marie Ng, Yvonne Becher, Diana Lee, Patrick Ip, and John Bacon-Shone. 2014. *Validation, Finalization and Adoption of the East Asia-Pacific Early Child Development Scales (EAP-ECDS)*. New York: UNICEF. *http://www.arnec.net/wp-content/uploads/2015/07/EAP-ECDS-Final-Report1.pdf*.

Rao, Nirmala, Jin Sun, Veronica Pearson, Emma Pearson, Hongyun Liu, Mark A. Constas, and Patrice L. Engle. 2012. "Is Something Better Than Nothing? An Evaluation of Early Childhood Programs in Cambodia." *Child Development* 83 (3): 864–76.

Rao, Nirmala, et al. 2016. *Final Report. East Asia-Pacific Early Child Development Scales - Short Form*. Unpublished Report, Asia Pacific Regional Network for Early Childhood, Singapore.

Rasheed, Muneera A. and Aisha K. Yousafzai. 2015. "The Development and Reliability of an Observational Tool for Assessing Mother-Child Interactions in Field Studies - Experience from Pakistan." *Child: Care, Health and Development* 41 (6): 1161–71.

Rayner, Keith. 2009. "Eye Movements and Attention in Reading, Scene Perception, and Visual Search." *The Quarterly Journal of Experimental Psychology (Hove)* 62 (8): 1457–506.

Reubens, Andrea. 2009. *Early Grade Mathematics Assessment (EGMA): A Conceptual Framework Based on Mathematics Skills Development in Children*. Research Triangle Park, NC: RTI International.

Richter, Linda M., Bernadette Daelmans, Joan Lombardi, Jody Heymann, Florencia Lopez Boo, Jere R. Behrman, Chunling Lu, Jane E. Lucas, Rafael Perez-Escamilla, Tarun Dua, Zulfiqar A. Bhutta, Karin Stenberg, Paul Gertler, Gary L. Darmstadt. 2017. "Investing in the Foundation of Sustainable Development: Pathways to Scale Up for Early Childhood Development." *The Lancet* 389 (10064): 103–18.

Richter, Linda M. and K. W. Grieve. 1991. "Home Environment and Cognitive Development of Black Infants in Impoverished South African Families. *Infant Mental Health Journal* 12 (2): 88–102.

Richter, Linda, Musawenkosi Mabaso, and Celia Hsiao. 2015. "Predictive Power of Psychometric Assessments to Identify Young Learners in Need of Early Intervention: Data from the Birth to Twenty Plus Cohort, South Africa." *South African Journal of Psychology* 46 (2): 175–90.

Rimm-Kaufman, Sara E. and Robert C. Pianta. 2000. "An Ecological Perspective on the Transition to Kindergarten: A Theoretical Framework to Guide Empirical Research." *Journal of Applied Developmental Psychology* 21 (5): 491–511.

Rimm-Kaufman, Sara E., Robert C. Pianta, and Martha J. Cox. 2000. "Teachers' Judgments of Problems in the Transition to Kindergarten." *Early Childhood Research Quarterly* 15 (2): 147–66.

Rodriguez, Soledad, Violeta Aranciba, and Consuelo Undurraga. 1996. *Escala de evaluación del desarrollo psicomotor: 0 a 24 meses. 12 ed.* Santiago, Chile: Galdoc.

Roggman, Lori A., Gina A. Cook, Mark S. Innocenti, Vonda Jump Norman, and Katie Christiansen. 2013. "Parenting Interactions with Children: Checklist of Observations Linked to Outcomes (PICCOLO) in Diverse Ethnic Groups." *Infant Mental Health Journal* 34 (4): 290–306.

Roncagliolo, Manuel, Marcelo Garrido, T. Walter, Patricio Peirano, and B. Lozoff. 1998. "Evidence of Altered Central Nervous System Development in Infants with Iron Deficiency Anemia at 6 Mo: Delayed Maturation of Auditory Brainstem Responses." *The American Journal of Clinical Nutrition* 68 (3): 683–90.

Roque, Daniela Tsubota, Rosani Aparecida Antunes Teixeira, Elaine C. Zachi, and Dora F. Ventura. 2011. "The Use of the Cambridge Neuropsychological Test Automated Battery (CANTAB) in Neuropsychological Assessment: Application in Brazilian Research with Control Children and Adults with Neurological Disorders." *Psychology & Neuroscience* 4 (2): 255–65.

Rose, Susan A., Judith F. Feldman, Jeffery J. Jankowski, and Ronan Van Rossem. 2012. "Information Processing from Infancy to 11 Years: Continuities and Prediction of IQ." *Intelligence* 40 (5): 445–57.

Rosselli, Monica and Alfredo Ardila. 2003. "The Impact of Culture and Education on Non-Verbal Neuropsychological Measurements: A Critical Review." *Brain and Cognition* 52 (3): 326–33.

Rossiter, John R. 2011. "Validity and Reliability." In *Measurement for the Social Sciences: The C-OAR-SE Method and Why It Must Replace Psychometrics,* by John R. Rossiter, 13–28. New York: Springer.

Rothbart, Mary K., Stephan A. Ahadi, Karen L. Hershey, and Phillip Fisher. 2001. "Investigations of Temperament at Three to Seven Years: The Children's Behavior Questionnaire." *Child Development* 72 (5): 1394–1408.

Rothney, Megan P., Emily V. Schaefer, Megan M. Neumann, Leena Choi, and Kong Y. Chen. 2008. "Validity of Physical Activity Intensity Predictions by ActiGraph, Actical, and RT3 Accelerometers." *Obesity* 16 (8): 1946–52.

Rubin, Kenneth H., Sheryl A. Hemphill, Xinyin Chen, Paul Hastings, Ann Sanson, Alida LoCoco, Ock Boon Chung, Sung-Yun Park, Carla Zappulla, Suman Verma, Chong-Hee Yoon, and Hyun Sim Doh. 2006. "Parenting Beliefs and Behaviors: Initial Findings From the International Consortium for the Study of Social and Emotional Development (ICSSED)." In *Parenting Beliefs, Behaviors, and Parent-Child Relations: A Cross-Cultural Perspective*, edited by Kenneth H. Rubin and Ock Boon Chung, 81–103. New York: Psychology Press.

Rubin, Kenneth H., Sheryl A. Hemphill, Xinyin Chen, Paul Hastings, Ann Sanson, Alida LoCoco, Carla Zappulla, Ock Boon Chung, Sung-Yun Park, Hyun Sim Doh, Huichang Chen, Ling Sun, Chong-Hee Yoon, and Liyin Cui. 2006. "A Cross-Cultural Study of Behavioral Inhibition in Toddlers: East-West-North-South." *International Journal of Behavioral Development* 30 (3): 219–26.

Rubio-Codina, Marta, Orazio Attanasio, and Sally Grantham-McGregor. 2016. "Mediating Pathways in the Socio-Economic Gradient of Child Development: Evidence from Children 6-42 Months in Bogota." *International Journal of Behavioral Development* 40 (6): 483–91.

Rubio-Codina, Marta, Orazio Attanasio, Costas Meghir, Natalia Varela, and Sally Grantham-McGregor. 2015. "The Socioeconomic Gradient of Child Development: Cross-Sectional Evidence from Children 6–42 Months in Bogota." *The Journal of Human Resources* 50 (2): 464–83.

Rubio-Codina, Marta, Maria Caridad Araujo, Orazio Attanasio, Pablo Muñoz, and Sally Grantham-McGregor. 2016. "Concurrent Validity and Feasibility of Short Tests Currently Used to Measure Early Childhood Development in Large Scale Studies." *PLoS One* 11 (8): e0160962.

Rutter, Michael. 1979. "Protective Factors in Children's Responses to Stress and Disadvantage, in Primary Prevention of Psychopathology." In *Primary Prevention of Psychopathology: Social Competence in Children*, edited by Martha Whalen Kent and Jon E. Rolf, 49–74. Hanover, NH: University Press of New England.

Rydz, David, Michael I. Shevell, Annette Majnemer, and Maryam Oskoui. 2005. "Topical Review: Developmental Screening." *Journal of Child Neurology* 20 (4): 4–21.

Saarni, Carolyn, Joseph J. Campos, Linda A. Camras, and David Witherington. 1998. "Emotional Development: Action, Communication, and Understanding." In *Handbook of Child Psychology, Fifth Edition. Volume Three: Social, Emotional, and Personality Development*, edited by William Damon and Nancy Eisenberg, 237–310. New York: John Wiley & Sons

Sabanathan, Saraswathy, Bridget Wills, and Melissa Gladstone. 2015. "Child Development Assessment Tools in Low-Income and Middle-Income Countries: How Can We Use Them More Appropriately?" *Arch Dis Child* 100 (5): 482–8.

Sameroff, Arnold J., Ronald Seifer, Alfred Baldwin, and Clara Baldwin. 1993. "Stability of Intelligence from Preschool to Adolescence: The Influence of Social and Family Risk Factors." *Child Development* 64 (1): 80–97.

Scarborough, Anita A., Kathleen M. Hebbeler, Rune J. Simeonsson, and Donna Spiker. 2007. "Caregiver Descriptions of the Developmental Skills of Infants and Toddlers Entering Early Intervention Services." *Journal of Early Intervention* 29 (3): 207–27.

Schady, Nobert, Jere Behrman, Maria Caridad Araujo, Rodrigo Azuero, Raquel Bernal, David Bravo, Florencia Lopez Boo, Karen Macours, Daniela Marshall, Christina Paxson, and Renos Vakis. 2015. "Wealth Gradients in Early Childhood Cognitive Development in Five Latin American Countries." *The Journal of Human Resources* 50 (2): 446–63.

Schofield, Thomas J., Monica J. Martin, Katherine J. Conger, Tricia M. Neppl, M. Brent Donnellan, and Rand D. Conger. 2011. "Intergenerational Transmission of Adaptive Functioning: A Test of the Interactionist Model of SES and Human Development." *Child Development* 82 (1): 33–47.

Schofield, Thomas J., Monica J. Martin, Katherine J. Conger, Tricia M. Neppl, M. Brent Donnellan, and Rand D. Conger. 2012. "Parent Personality and Positive Parenting as Predictors of Positive Adolescent Personality Development over Time." *Merrill-Palmer Quarterly* 58 (2): 255–83.

Schweizer, Karl and Christine DiStefano, eds. 2016. *Principles and Methods of Test Construction: Standards and Recent Advances*. Boston, MA: Hogrefe Publishing.

Sebate, K. M. 2000. *Report on the Standardisation of the Grover-Counter Scale of Cognitive Development*. Pretoria, South Africa: Human Sciences Research Council.

Semrud-Clikeman, Margaret, Regilda Anne A. Romero, Elizabeth L. Prado, Elsa G. Shapiro, Paul Bangirana, and Chandy John. 2016. "Selecting Measures for the Neurodevelopmental Assessment of Children in Low- and Middle-Income Countries." *Child Neuropsychology* 1–42.

Sen, Amartya. 1999. *Development as Freedom*. New York: Knopf.

Serpell, Robert. 2015. "Selecting Indicators of Healthy Early Childhood Development." *Human Development Intervention Network* (blog), September 11. http://hdin.org/selecting-indicators-of-healthy-early-childhood-development/.

Serpell, Robert and Jacqueline Jere-Folotiya. 2008. "Developmental Assessment, Cultural Context, Gender and Schooling in Zambia." *International Journal of Psychology* 43 (2): 1–9.

Shankar, Anita V., Zaitu Asrilla, Josephine K. Kadha, Susy Sebayang, Mandri Apriatni, Ari Sulastri, Euis Sunarsih, and Anuraj H. Shankar. 2009. "Programmatic Effects of a Large-Scale Multiple-Micronutrient Supplementation Trial in Indonesia: Using Community Facilitators as Intermediaries for Behavior Change." *Food and Nutrition Bulletin* 30 (2 Suppl): S207–14.

Shaw, Arthur, Lena Nguyen, Ulrike Nischan, and Herschel Sy. 2011. *Comparative Assessment of Software Programs for the Development of Computer-Assisted Personal Interview (CAPI) Applications*. College Park, MD: IRIS Center, University of Maryland.

Shoda, Yuichi, Walter Mischel, and Philip K. Peake. 1990. "Predicting Adolescent Cognitive and Self-Regulatory Competencies From Preschool Delay of Gratification: Identifying Diagnostic Conditions." *Developmental Psychology* 26 (6): 978–86.

Shonkoff, Jack P. and Paul C. Marshall. 2000. "The Biology of Developmental Vulnerability." In *Handbook of Early Childhood Intervention*, edited by Jack P. Shonkoff and Samuel J. Meisels, 35-53. Cambridge: Cambridge University Press.

Shonkoff, Jack P. and Deborah A. Phillips, eds. 2000. *From Neurons to Neighborhoods: The Science of Early Childhood Development*. Washington, DC: National Academies Press.

Sizonenko, Stephane V., Claudio Babiloni, Eveline A. de Bruin, Elizabeth B. Isaacs, L. S. Jönsson, David O. Kennedy, Marie E. Latulippe, M. Hasan Mohajeri, Judith Moreines, Pietro Pietrini, Kristine B. Walhovd, Robert J. Winwood, and John W. Sijben. 2013. "Brain Imaging and Human Nutrition: Which Measures to Use in Intervention Studies?" *The British Journal of Nutrition* 110 (Suppl 1): S1–30.

Snow, Catherine E. and Susan B. Van Hemel, eds. 2008. *Early Childhood Assessment: Why, What, and How*. Washington, DC: National Academies Press.

Solarsh, Barbara and Erna Alant. 2006. "The Challenge of Cross-Cultural Assessment—The Test of Ability To Explain for Zulu-Speaking Children." *Journal of Communication Disorders* 39 (2): 109–38.

Squires, Jane K., LaWanda Potter, Diane D. Bricker, and Suzanne Lamorey. 1998. "Parent-Completed Developmental Questionnaires: Effectiveness with Low and Middle Income Parents." *Early Childhood Research Quarterly* 13 (2): 345–54.

Stoltzfus, Rebecca J., Jane D. Kvalsvig, Hababu M. Chwaya, Antonio Montresor, Marco Albonico, James M. Tielsch, Lorenzo Savioli, and Ernesto Pollitt. 2001. "Effects of Iron Supplementation and Anthelmintic Treatment on Motor and Language Development of Preschool Children in Zanzibar: Double Blind, Placebo Controlled Study." *The BMJ* 323 (7326): 1–8.

Stright, Anne Dopkins, Kathleen Cranley Gallagher, and Ken Kelley. 2008. "Infant Temperament Moderates Relations Between Maternal Parenting in Early Childhood and Children's Adjustment in First Grade." *Child Development* 79 (1): 186–200.

Sudfeld, Christopher R., Dana Charles McCoy, Goodarz Danaei, Günther Fink, Majid Ezzati, Kathryn G. Andrews, and Wafaie F. Fawzi. 2015. "Linear Growth and Child Development in Low- and Middle-Income Countries: A Meta-Analysis." *Pediatrics* 135 (5): e1266–75.

Sun, Feng-Tso, Cynthia Kuo, Heng-Tze Cheng, Senaka Buthpitiya, Patricia Collins, and Martin Griss. 2012. "Activity-Aware Mental Stress Detection Using Physiological Sensors." In *Mobile Computing, Applications, and Services: Second International ICST Conference, MOBICASE 2010, Santa Clara, CA, USA, October 25–28, 2010, Revised Selected Papers*, edited by Martin Griss and Guang Yang, 211–30. Berlin, Heidelberg: Springer.

Super, Charles M. 1981. "Behavioral Development in Infancy." In *Handbook of Cross-Cultural Human Development*, edited by Ruth H. Munroe, Robert L. Munroe, and Beatrice B. Whiting, 181–271. New York: Garland Press.

Suskind, Dana L., Kristin R. Leffel, Eileen Graf, Marc W. Hernandez, Elizabeth A. Gunderson, Shannon G. Sapolich, Elizabeth Suskind, Lindsey Leininger, Susan Goldin-Meadow, and Susan C. Levine. 2016. "A Parent-Directed Language Intervention for Children of Low Socioeconomic Status: A Randomized Controlled Pilot Study." *Journal of Child Language* 43 (2): 366–406.

Sylva, Kathy, Iram Siraj-Blatchford, Brenda Taggart, Pam Sammons, Edward Melhuish, Karen Elliott, and Vasiliki Totsika. 2006. "Capturing Quality in Early Childhood Through Environmental Rating Scales." *Early Childhood Research Quarterly* 21 (1): 76–92.

Tamayo, Jose M. 1987. "Frequency of Use as a Measure of Word Difficulty in Bilingual Vocabulary Test Construction and Translation." *Educational and Psychological Measurement* 47 (4): 893–902.

Tamis-LeMonda, Catherine S., Marc H. Bornstein, and Lisa Baumwell. 2001. "Maternal Responsiveness and Children's Achievement of Language Milestones." *Child Development* 72 (3): 748–67.

Tasbihsazan, Reza, Ted Nettelbeck, and Neil Kirby. 2003. "Predictive Validity of the Fagan Test of Infant Intelligence." *British Journal of Developmental Psychology* 21 (4): 585–97.

Taylor, Margot J. and Torsten Baldeweg. 2002. "Application of EEG, ERP and Intracranial Recordings to the Investigation of Cognitive Functions in Children." *Developmental Science* (3): 318–34.

Teixeira, Rosani Aparecida Antunes, Elaine Cristina Zachi, Daniela Tsubota Roque, Anita Taub, and Dora Fix Ventura. 2011. "Memory Span Measured by the Spatial Span Tests of the Cambridge Neuropsychological Test Automated Battery in a Group of Brazilian Children and Adolescents." *Dementia & Neuropsychologia* 5 (2): 129–34.

Thelen, Esther. 2000. "Grounded in the World: Developmental Origins of the Embodied Mind." *Infancy* 1 (1): 3–28.

Thompson, Ross A. and Charles A. Nelson. 2001. "Developmental Science and the Media: Early Brain Development." *American Psychologist* 56 (1): 5.

Thompson, Ross A. and H. Abigail Raikes. 2006. "The Social and Emotional Foundations of School Readiness." In *Early Childhood Mental Health*, edited by Jane Knitzer, Roxane Kaufmann, and Deborah Perry, 13-35. Baltimore, MD: Paul H. Brookes Publishing.

Tluczek, Audrey, Elaine H. Mischler, Philip M. Farrell, Norman Fost, Nanette M. Peterson, Patrick Carey, W. Theodore Bruns, and Catherine McCarthy. 1992. "Parents' Knowledge of Neonatal Screening and Response to False-Positive Cystic Fibrosis Testing." *Journal of Development and Behavioral Pediatrics* 13 (3): 181–6.

tobiipro. 2016. "Tobii Pro Glasses 2." http://www.tobiipro.com/product-listing/tobii-pro-glasses-2/.

Tsetlin, Marina M., S. I. Novikova, Elena Orekhova, Natalya Pushina, E. V. Malakhovskaia, A. I. Filatov, and Tatiana Stroganova. 2012. "Developmental Continuity in the Capacity of Working Memory from Infancy to Preschool Age." *Neuroscience and Behavioral Physiology* 42 (7): 692–9.

Ullman, Michael T. 2014. "Language and the Brain." In *An Introduction to Language and Linguistics*, edited by Ralph W. Fasold and Jeff Connor-Linton, 249–86. Cambridge: Cambridge University Press.

Unay, Bulent, S. Sarici, U. Ulas, R. Akin, F. Alpay, and E. Gokcay. 2004. "Nutritional Effects on Auditory Brainstem Maturation in Healthy Term Infants." *Archives of Disease in Childhood. Fetal and Neonatal Edition* 89 (2): F177–9.

United Nations. 2016. *Goal 4: Ensure Inclusive and Equitable Quality Education and Promote Lifelong Learning Opportunities for All.* Sustainable Development Knowledge Platform, New York.

United Nations News Centre. 2016. "UNICEF, World Bank Urge Greater Investment in Early Childhood Development." Press Release, April 14.

United States, National Education Goals Panel, Goal 1 Technical Planning Group. 1995. *Reconsidering Children's Early Development and Learning: Toward Common Views and Vocabulary,* by Sharon Lynn Kagan, Evelyn Moore, and Sue Bredekamp, eds. Report, Washington, DC.

van de Vijver, Fons J. R. and Ronald K. Hambleton. 1996. "Translating Tests: Some Practical Guidelines." *European Psychologist* 1 (2): 89–99.

van de Vijver, Fons J. R. and Ype H. Poortinga. 2005. "Conceptual and Methodological Issues in Adapting Tests." In *Adapting Educational and Psychological Tests for Cross-Cultural Assessment,* edited by Ronald K. Hambleton, Peter F. Merenda, and Charles D. Spielberger, 39–63. Mahwah, NJ: Lawrence Erlbaum Associates.

van Widenfelt, Brigit M., Philip D. Treffers, Edwin de Beurs, Bart M. Siebelink, and Els Koudijs. 2005. "Translation and Cross-Cultural Adaptation of Assessment Instruments Used in Psychological Research with Children and Families." *Clinical Child and Family Psychology Review* 8 (2): 135–47.

VanDam, Mark, D. Kimbrough Oller, Sophie E. Ambrose, Sharmistha Gray, Jeffrey A. Richards, Dongxin Xu, Jill Gilkerson, Noah H. Silbert, and Mary Pat Moeller. 2015. "Automated Vocal Analysis of Children with Hearing Loss and Their Typical and Atypical Peers." *Ear and Hearing* 36 (4): e146–52.

Vanderwert, Ross E. and Charles A. Nelson. 2014. "The Use of Near-Infrared Spectroscopy in the Study of Typical and Atypical Development." *NeuroImage* 85 (0 1): 264–71.

Vazir, S. and K. Kashinath. 1999. "Influence of the ICDS on Psychosocial Development of Rural Children in Southern India." *Journal of the Indian Academy of Applied Psychology* 25 (1–2): 11–24.

Vierhaus, Marc, Arnold Lohaus, Thorsten Kolling, Manuel Teubert, Heidi Keller, Ina Fassbender, Claudia Freitag, Claudia Goertz, Frauke Graf, Bettina Lamm, Sibylle M. Spangler, Monika Knopf, and Gudrun Schwarzer. 2011. "The Development of 3- to 9-Month-Old Infants in Two Cultural Contexts: Bayley Longitudinal Results for Cameroonian and German Infants." *European Journal of Developmental Psychology* 8 (3): 349–66.

Wachs, Theodore. 1993. "Family Environmental Influences and Development: Illustrations from the Study of Undernourished Children." In *Families, Risk, and Competence,* edited by Michael Lewis and Candice Feiring, 245–68. Mahway, NJ: Lawrence Erlbaum Associates Publishers.

Wachs, Theodore D., Santiago Cueto, and Haogen Yao. 2016. "More Than Poverty: Pathways from Economic Inequality to Reduced Developmental Potential." *International Journal of Behavioral Development* 40 (6): 536–43.

Wachs, Theodore and Smita Desai. 1993. "Parent-Report Measures of Toddler Temperament and Attachment: Their Relation to Each Other and to the Social Microenvironment." *Infant Behavior and Development* 16 (3): 391–6.

Wachs, Theodore, Marian Sigman, Zeinab Bishry, Wafaa Moussa, Norge Jerome, Charlotte Neumann, Nimrod Bwibo, and Mary Alice McDonald. 1992. "Caregiver Child Interaction Patterns in Two Cultures in Relation to Nutritional Intake." *International Journal of Behavioral Development* 15 (1): 1–18.

Wagstaff, Adam, Flavia Bustreo, Jennifer Bryce, Mariam Claeson, and the WHO-World Bank Child Health and Poverty Working Group. 2004. "Child Health: Reaching the Poor." *American Journal of Public Health* 94 (5): 726–36.

Walker, Susan P., Theodore D. Wachs, Sally Grantham-McGregor, Maureen M. Black, Charles A. Nelson, Sandra L. Huffman, Helen Baker-Henningham, Susan M. Chang, Jena D. Hamadani, Betsy Lozoff, Julie M. Meeks Gardner, Christine A. Powell, Atif Rahman, and Linda Richter. 2011. "Inequality in Early Childhood: Risk and Protective Factors for Early Child Development." *The Lancet* 378 (9799): 1325–38.

Walker, Susan P., Theodore D. Wachs, Julie Meeks Gardner, Betsy Lozoff, Gail A. Wasserman, Ernesto Pollitt, Julie A. Carter, and the International Child Development Steering Group. 2007. "Child Development: Risk Factors for Adverse Outcomes in Developing Countries." *The Lancet* 369 (9556): 145–57.

Warlaumont, Anne S., Jeffrey A. Richards, Jill Gilkerson, and D. Kimbrough Oller. 2014. "A Social Feedback Loop for Speech Development and Its Reduction in Autism." *Psychological Science* 25 (7): 1314–24.

Weber, A., A. Fernald, and Y. Diop. 2017. "When Cultural Norms Discourage Talking to Babies: Effectiveness of a Parenting Program in Rural Senegal." *Child Development* 88 (5): 1513-26.

Weisleder, Adriana and Anne Fernald. 2013. "Talking to Children Matters: Early Language Experience Strengthens Processing and Builds Vocabulary." *Psychological Science* 24 (11): 2143–52.

Welsh, Marilyn C., Sarah L. Friedman, and Susan J. Spieker. 2006. "Executive Functions in Developing Children: Current Conceptualizations and Questions for the Future." In *Blackwell Handbook of Early Childhood Development*, edited by Kathleen McCartney and Deborah Phillips, 167–87. London: Wiley-Blackwell.

Whiting, Beatrice and Carolyn P. Edwards. 1973. *A Cross-Cultural Analysis of Sex Differences in the Behavior of Children Aged Three Through 11.* Lincoln, NE: Faculty Publications, Department of Psychology, University of Nebraska - Lincoln.

WHO (World Health Organization). 2013. "WHO Director-General Addresses Health Promotion Conference." Press Release, June 10.

WHO Multicentre Growth Reference Study Group. 2006. *WHO Child Growth Standards: Length/Height-for-Age, Weight-for-Age, Weight-for-Length, Weight-for-Height and Body Mass Index-for-Age: Methods and Development*. Geneva: World Health Organization.

WHO Multicentre Growth Reference Study Group. 2006. "WHO Motor Development Study: Windows of Achievement for Six Gross Motor Development Milestones." *Acta Paediatrica Supplement* 450: 86–95.

Woerner, Wolfgang, Bacy Fleitlich-Bilyk, Rhonda Martinussen, Janet Fletcher, Giuletta Cucchiaro, Paulo Dalgalarrondo, Mariko Lui, and Rosemary Tannock. 2004. "The Strengths and Difficulties Questionnaire Overseas: Evaluations and Applications of the SDQ Beyond Europe." *European Child & Adolescent Psychiatry* 13 (Suppl 2): 47–54.

Wong, Suzy L., Rachel Colley, Sarah Connor Gorber, and Mark Tremblay. 2011. "Actical Accelerometer Sedentary Activity Thresholds for Adults." *Journal of Physical Activity & Health* 8 (4): 587–91.

Woodward, Amanda L. and Ellen M. Markman. 1998. "Early Word Learning." In *Handbook of Child Psychology, Fifth Edition. Volume 2: Cognition, Perception and Language*, edited by Damon, William, Deanna Kuhn, and Robert S. Siegler, 371–420. New York: John Wiley & Sons.

Wuermli, Alice. J., Hirokazu Yoshikawa, J. Lawrence Aber, Carly T. Dolan, Christine Campo, Stephanie Odiase, Ilana Sigal, Maya Nauphal, Leslie Williams, Rachel A. Rosenfeld, Joost de Laat (2016) *An Inventory of Child and Adolescent Measures and Assessments used in Low- and Middle-Income Countries. Global TIES for Children.* New York: New York University; Washington, DC: Strategic Impact Evaluation Fund, World Bank.

Yoshikawa, Hirokazu, Christina Weiland, Jeanne Brooks-Gunn, Margaret R. Burchinal, Linda M. Espinoza, William T. Gormley, Jens Ludwig, Katherine A. Magnuson, Deborah Phillips, and Martha J. Zaslow. 2013. *Investing in Our Future: The Evidence Base on Preschool Education*. Washington, DC: Society for Research in Child Development.

Yousafzai, Aisha K., Jelena Obradović, Muneera A. Rasheed, Arjumand Rizvi, Ximena A. Portilla, Nicole Tirado-Strayer, Saima Siyal, and Uzma Memon. 2016. "Effects of Responsive Stimulation and Nutrition Interventions on Children's Development and Growth at Age 4 Years in a Disadvantaged Population in Pakistan: A Longitudinal Follow-Up of a Cluster-Randomised Factorial Effectiveness Trial." *The Lancet Global Health* 4 (8): e548–58.

Zuilkowski, Stephanie Simmons, Dana Charles McCoy, Robert Serpell, Beatrice Matafwali, and Günther Fink. 2016. "Dimensionality and the Development of Cognitive Assessments for Children in Sub-Saharan Africa." *Journal of Cross-Cultural Psychology* 43 (3): 341–54.

Glossary

Ability test: An assessment that provides a range of scores representing children's development levels across the range of typical development, in contrast to a screening test, which indicates risk of delay but does not provide a range of scores in a group of typically developing children.

Accelerometer: A small device that can be worn on the body and provides a continuous, objective measure of physical activity.

Adaptation: For assessments, modification of items, materials, and procedures to fit the local context.

Adaptive behavior: The ability to perform daily-life skills, such as self-feeding, dressing, toilet training, interaction with others, and to adjust to new situations.

Adoption: For assessments, direct translation of a test to a new context without modification.

Assembly: For assessments, creation of a new test by bringing together items and methods from various existing sources.

Baseline: The situation prior to an intervention, against which progress can be assessed. Baseline data allow to establish if groups are comparable before the intervention.

Bias: In terms of assessment, the presence of systematic differences among results of the test-takers, whether due to culture, gender, race or other factors.

Canalized abilities: Skills that all normal human beings eventually acquire, such as walking and talking.

Cognitive skills: The processes or faculties by which knowledge is acquired and manipulated, including abilities such as memory, problem solving and analytical skills.

Concurrent validity: Extent to which the results of a particular test or measurement tool correlate with results of a previously established measurement of the same construct measured at the same time.

Construct bias: Bias due to the failure of the instrument to measure the same underlying construct across different contexts or groups.

Cronbach's alpha: A measure of internal consistency, or how closely related a set of items is as a group; used to measure a scale's internal reliability.

Developmental delay: The condition in which a child's development lags behind established normal ranges for his or her age. Delay is determined relative to normative development within a given population.

Developmental milestone: A behavior, skill or ability that is demonstrated by a specified age during infancy and early childhood in typical development.

Developmental quotient: Developmental quotient refers to a numerical measure of a child's performance on a developmental test relative to the performance of other children of the same age.

Developmental trajectory: A curve of repeated observations of an aspect of development throughout childhood. Individuals may differ in the starting point, the degree of acceleration or deceleration, the timing of acceleration or deceleration, or overall shape of the curve.

Effortful control: The ability to inhibit a dominant response to perform a subdominant response.

Endline: In terms of surveys, data gathered after a program to measure how much has changed from the baseline. It allows the measurement of the impact of an intervention comparing treatment and control groups.

Equivalence: In assessments, whether the language translation, functional meaning, cultural relevance or level of difficulty is equivalent for children in different contexts or groups. When tests are equivalent, they are by definition unbiased.

Ethical review board: An institution that reviews and approves all measurement protocols involving human subjects.

Event-related potential: An electrophysiological brain response that is the direct result of a specific sensory, cognitive, or motor event.

Executive function: A set of cognitive processes that are necessary to control behavior and cognition, including abilities such as inhibitory control, attention, and working memory.

Factor analysis: A statistical procedure that extracts one or more latent or unmeasurable variables from a set of observed variables, based on the shared variance between the observed variables.

Fine motor skills: The ability to coordinate precise movements, such as picking up writing or holding a spoon, that use the small muscles of the hands, feet, wrists, lips, and tongue.

Fixation: The act of looking at a specific point for a period of time.

Fluid ability: In psychology, the ability to solve new problems, use logic in new situations, and identify patterns.

Galvanic skin response: A measurable change in the electrical resistance of the skin caused by emotional arousal. Also called skin conductance or electro-dermal activity.

Grit: The tendency to sustain interest in, and effort toward, long-term goals.

Gross motor skills: Child's ability to control and coordinate his or her body gross movements, such as walking, running, jumping, or throwing.

Head Start: A national program to improve school readiness among low-income children under age five in the United States.

Height-for-age z-score: A standardized measure of a child's height in comparison to children his or her age based on standard norms. In global research, the most commonly used norms are from the World Health Organization Multi-Centre Growth Reference Study.

Impact evaluation: An assessment of changes, both intended and unintended, that can be attributed to a particular intervention, such as a project, program, or policy. An impact evaluation measures the causal effect of the intervention on a set of outcomes.

Intelligence quotient: Also known as IQ, a score derived from standardized testing to assess human intelligence. A numerical measure of a child's performance on a intelligence test relative to the performance of other children of the same age.

Intervention: In the context of impact evaluation, this is the project, program, design innovation, or policy to be evaluated. Also known as the treatment.

Intraclass correlation coefficient: A statistic used to calculate the reliability of measurements or ratings.

Item bias: Bias that occurs when individual test items do not measure the same way across groups; sources of such bias include poor translation and culturally inappropriate content.

Item response theory: A theory for the design, analysis, and scoring of tests, questionnaires, and similar instruments measuring complex abilities, attitudes, or other variables. It is a theory of testing based on the relationship between individuals' probability of passing an item and levels of performance on an overall measure of the ability that item was designed to measure.

Language skills: The ability to understand and express verbal communication.

Latency: In the case of a response to a stimulus, the delay before a response takes place after a stimulus occurs.

Latent factor: A score derived from factor analysis that represents an underlying ability that cannot be measured directly, but is derived from observed indicators.

Likert scale: A five- or seven-point scale used to allow individuals to express how much they agree or disagree with a particular statement.

Longitudinal study: A research design that involves repeated observations of the same variables, such as people, over a long period of time, often many decades.

Macronutrient: An energy-providing substance consumed by organisms in large quantities. The three macronutrients in nutrition are carbohydrates, lipids, and proteins.

Magnetoencephalography: A neuroimaging technique for mapping brain activity by recording magnetic fields caused by electrical currents occurring naturally in the brain.

Method bias: Bias that occurs when the administration or procedures of the test – such as the use of unfamiliar stimuli – differentially affect the scores of the groups being tested.

Measurement invariance: A statistical property of measurement that shows the same construct is being measured across specific groups. If measurement invariance is established, then unbiased comparisons can be made between groups.

Micronutrient: Nutrients required in the diet in small amounts that enable the body to produce enzymes, hormones, and other substances essential for proper growth and development. Examples include iodine, vitamin A, and iron.

Naturalistic observation: A research method in which a subject is observed in his or her natural environment without any manipulation by the observer.

Neuroimaging: Also called brain imaging, the use of various techniques to either directly or indirectly image the structure, function, or pharmacology of the nervous system.

Neuronal system: System related to neurons.

Novelty preference: The tendency for infants to pay more attention to new objects or people than those they've seen before.

Pearson's correlation: In statistics, a measure of the linear correlation between variables X and Y, and also known as the Pearson correlation coefficient (PCC), the Pearson's r or bivariate correlation.

Plasticity: In psychology, neuroplasticity, or brain plasticity, refers to the brain's ability to change throughout life.

Positron emission tomography (Petscan): An imaging test that helps reveal how tissues and organs are functioning.

Predictive validity: In psychometrics, the extent to which a score on a test predicts the scores of some criterion measured at a later time point.

Psychometrics: The field of study concerned with the theory and technique of psychological measurement.

Psychometric evaluation: The evaluation of the properties of a psychological measurement tool.

Pre-academic skills: Skills needed to learn reading and math, such as counting and letters.

Representative sample: A subset of a statistical population that accurately reflects the members of the entire population

Saccade: Rapid eye movement aimed at bringing an object into focus.

Scale score: Conversion of the raw score onto a scale that is common to all test forms for that assessment.

School readiness: Broadly defined, readiness of the individual child, the school's readiness for children, and the ability of the family and community to support a child's performance in school. In terms of the child specifically, social, emotional, cognitive, and physical development that prepares him or her for a successful learning trajectory.

Screening test: A brief measure used to identify children who are at risk of developmental problems in one or more domains,

Selective attention: The capacity to respond to certain stimuli selectively when presented with multiple stimuli simultaneously.

Sequential processing: Ability to solve problems by ordering items or placing them in sequence.

Simultaneous processing: Ability to solve problems by integrating diverse pieces of information simultaneously.

Sleeper effect:. An effect that is not detected at an early time point but is detected at a later time point.

Social-emotional skills: The regulation of emotional responses and social interactions, which is a function of both temperament and self-regulation, including behavior problems, social competency, and emotional competency

Standard: An expectation or norm of typical development.

Stimulus: In psychology, any object or event that elicits a sensory or behavioral response in an organism.

Stunting: Impaired growth in height compared to a healthy population. Children are defined as stunted if their height-for-age is more than two standard deviations below the WHO Child Growth Standards median.

Sulcus: A groove in the fold of the cerebral cortex.

Synaptogenesis: The formation of synapses between neurons in the nervous system, which occurs throughout a healthy person's lifespan, but which occurs most rapidly during early brain development.

Temperament: Biological influences on the experience and expression of emotion, including extraversion/surgency (positive affect, activity level, impulsivity, risk-taking), negative affectivity (fear, anger, sadness, discomfort), and effortful control (attention shifting and focusing, perceptual sensitivity, inhibitory and activation control)

Wasting: Below minus two standard deviations from median weight for height of a reference population, indicates in most cases a recent and severe process of weight loss, which is often associated with acute starvation or severe disease.

Working memory: Also called short-term memory, a system for temporarily storing and managing the information required to carry out complex cognitive tasks such as learning, reasoning, and comprehension.

Z-score: A score with a mean of zero and standard deviation of one. When transforming a raw score to a z-score, the z-score indicates how far above or below the sample mean the raw score is, in units of standard deviation.