Measuring Young Children's Development and School Readiness: A Scan of Population-Level Measures

Katherine Paschall, Catherine Schaefer, Kathryn Tout, Tamara Halle

Overview

States and communities are making investments in cross-sector initiatives to improve outcomes for young children and their families. These initiatives typically set increased school readiness or children's healthy development as the ultimate goal for the work; however, the capacity to define and measure school readiness and/or child well-being is limited, especially at scale, for very young children (0-3). Current efforts include tools that different reporters (parents, teachers, or practitioners) complete in different settings (kindergarten, pediatric settings, or homes) with varying constructs and measurement purposes (screening, developmental assessment). Progress has been made both internationally and in the United States to develop and use new measures that can provide population-level estimates and inform early childhood initiatives. We have an opportunity to learn from these successful efforts and to reflect on the implications for next steps for the field.

This project aims to develop a set of recommendations for the field and for funders seeking guidance about the best way to direct resources to support measures development and the implementation of population measurement strategies in communities and states. In order to achieve these goals, the project first examines what is available and appropriate for use at a population level to understand children's development from ages 0-5 with a focus on measures for infants and toddlers. Following the dissemination of this measures scan, we will convene a meeting of experts and stakeholders to develop recommendations for advancing the validation and use of population-level measures of young children.

Why measure child development at the population-level?

Many efforts around the world are focused on improving children's well-being, with increasing focus on the well-being of young children, ages 0–5. Stakeholders, including philanthropists, educators, politicians, program administrators, and advocacy groups want to know the status of young children's development, their overall well-being, and if investments aimed at improving children's well-being are yielding the desired outcomes. Global and national efforts are emerging to develop common measures of children's development during the first years of life, which can help answer such questions as:

- How well are particular or collective initiatives working to improve children's development?
- Within a geographic area (e.g., community, state, nation), how are child development indicators changing compared to other social and economic conditions, such as rates of homelessness, lead exposure, and child welfare involvement?
- Which subgroups of children face disparate child development or school readiness outcomes?

- How do children in one community or state compare to children in another community or state in terms of their development and potential readiness for school?

Measures either designed for monitoring development at the population-level or that can be feasibly administered periodically to whole or representative populations of children can address these questions.

This document has two parts. First, we conduct a scan of available measures for the purpose of describing young children's development in the years prior to kindergarten and present key considerations for the selection and future validation of such measures. Second, we detail considerations for the development and implementation of a measurement strategy, specifically one focused on monitoring the progress of and informing early childhood initiatives.

Part 1: The Landscape of Population-Level Early Childhood Measurement

Ideal traits of a population measurement tool

There are many factors to consider when choosing a tool or set of tools to capture child development at the population level. Fernald and colleagues[1] outlined ideal traits of an assessment and acknowledged that no tool meets all ideal traits; a modified list of ideal traits specific to population-level measurement of child development is presented in Table 1. Ideals 1–8 concern the validity of a tool, which affects how the score is interpreted (e.g., does the score accurately reflect the child's ability at every age? Does the score predict future success?) and its feasibility (e.g., is the tool easy to administer and low cost?).

Ideal traits 9 and 10 represent a tool's ability to explain a child or group of children's scores. These traits are not essential elements of a population-level measurement tool, and there is no tool designed for population-level assessment that simultaneously captures the process by which physiological or environmental factors drive or impact a given child's score. The ability to link children's scores to outcomes that can be affected by policy and program-level changes is a key condition of the success of a measurement strategy. In Part 2, we dive deeper into a discussion on the elements of a successful measurement strategy, including one that is able to describe the context of the data yielded from a given measure of child outcomes at the population level.

Table 1. Ideal traits of a population-level measurement tool

Ideal	Trait	Type of trait
Ideal 1	The score represents the child's true ability.	Validity
Ideal 2	The tool is appropriate, interpretable, and has high reliability and validity in all contexts and cultures.	Validity and reliability
Ideal 3	The tool shows variance in scores at all ages and ability levels.	Validity
Ideal 4	The tool is easy to administer.	Feasibility

[1] Fernald, L. C. H., Prado, E., Kariger, P., & Raikes, A. (2017). A Toolkit for Measuring Early Childhood Development in Low- and Middle-Income Countries Washington, DC: The World Bank.

Ideal 5	The tool can be administered quickly and at low cost.	Feasibility
Ideal 6	The tool can easily be administered with regularly (at least biennially) and at low cost.	Feasibility
Ideal 7	The score is relevant to a child's practical function in daily life and therefore relevant to policy and program design.	Validity and utility
Ideal 8	The tool and its score are a good indicator of future success.	Validity
Ideal 9	The physiological, neurological, and health mechanisms underlying performance are well-understood.	Explanatory
Ideal 10	The impact of health, nutrition, and environmental factors on the score is well-understood.	Explanatory

Note. Modified from Fernald, L. C. H., Prado, E., Kariger, P., & Raikes, A. (2017). A Toolkit for Measuring Early Childhood Development in Low- and Middle-Income Countries. Washington, DC: The World Bank.

We also recognize that not every early childhood initiative has the same measurement needs. When reviewing the measures profiled in the scan and crafting recommendations for adopting or further validating particular measurement tools, we consider both how the measurement tool aligns with the list of ideal traits and how these traits intersect with the various goals of early childhood initiatives.

Inclusion criteria for the measurement scan

This measures scan includes descriptions of tools that either were designed for use at the population level or show promise for being used as a population-level measurements of children's developmental skills, knowledge, and behavior related to school readiness. The scan includes measures of both whole-child development and specific domains or skills, given that the decision to focus on development broadly or specifically will vary by initiative and community. The scan provides an overview of each measure's current use, purpose, and psychometric properties, including validation and evidence of reliability. It also summarizes each measure's challenges and limitations, particularly related to implementation. The scan of measures is presented in Appendix A. Our scan was informed by several other compendia of measures, including:

- The International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st)[2]
- The World Bank's toolkit for measurement of early childhood development[3]
- Two developmental screener compendia prepared for the Office of Planning, Research, and Evaluation at the Administration of Children and Families[4,5]

[2] Fernandes M, Stein A, Newton CR, Cheikh-Ismail L, Kihara M, et al. (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.
[3] Fernald, L. C. H., Prado, E., Kariger, P., & Raikes, A. (2017). A Toolkit for Measuring Early Childhood Development in Low- and Middle-Income Countries Washington, DC: The World Bank.
[4] Halle, T., Zaslow, M., Wessel, J., Moodie, S., and Darling-Churchill, K. (2011). Understanding and Choosing Assessments and Developmental Screeners for Young Children: Profiles of Selected Measures. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.
[5] Moodie, S., Daneri, P., Goldhagen, S., Halle, T., Green, K., & LaMonte, L. (2014). Early childhood developmental screening: A compendium of measures for children ages birth to five (OPRE Report 201411). Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families.

Our scan does not include all measures used in clinical, research, and community settings to assess children's development[6]; rather, it includes those that are the strongest candidates for measuring children's development at a population level at regular intervals. The main considerations for inclusion in the scan were ***purpose*** and ***feasibility***. We explain why these factors were so important for identifying strong candidate measures below.

Purpose

Measures of early childhood development vary in purpose, with some designed to measure child outcomes and others designed to measure the impact of interventions, inform instruction, or identify children needing additional evaluation for development delays. In general, measures should be used only for their intended and demonstrated purpose.[7] Given this project's focus on describing population-level child development to inform early childhood initiatives, our scan highlights measures that were designed for measuring child outcomes. More specifically, we focus on:

1. Global or national population monitoring tools
2. Measures designed to capture change or outcomes as a result of a program/initiative
3. Screening tools (select measures)

Population monitoring tools are designed to detect trends in child development and are typically used by policymakers. Population monitoring tools can be used to track trends over time and compare outcomes for subpopulations of children. Tools designed for this purpose clearly align with the current project's goals, as they provide population-level information on child development, including differences in groups and differences across time. Two examples of population monitoring assessment included in this scan are the Caregiver Reported Early Development Inventory (CREDI) short form and the Global Scale for Early Development (GSED) short form.

Program evaluation tools are designed to assess child outcomes and change in outcomes as a result of a program or intervention. They are typically designed to capture developmental skills more thoroughly than population monitoring tools. Information from program evaluation tools can be useful for informing early childhood initiatives, making them relevant to this project's purpose. Examples include the MacArthur Bates Communicative Development Inventories (CDI), the Minnesota Executive Function Scale (MEFS), and the GSED long form.

This scan does not include most **developmental screening tools**, which are designed to identify children at risk for developmental delay and are more sensitive to capturing non-normative development. Screening tools are not designed to show variability in normative development and many typically developing children score at the ceiling. In addition, not all screening tools provide a standardized score, and most are designed for assessing individual children rather than providing information at a population level. Data from these tools cannot be used to provide information about school readiness or child development at the population level. Since the vast majority of measures used to assess young children's social-emotional development are designed as screening or diagnostic tools, the scan does not include most social-emotional assessments. If you are interested in learning more about these tools, please see the following compendia: Characteristics of Existing Measures of Social and Emotional Development in Early Childhood and a Review of Measures of Social and Emotional Development.

Two developmental screening tools are included in our scan because they are widely used and demonstrate utility beyond their purpose as a screening tool: The Ages and Stages Questionnaires (ASQ) and the Survey

[6] A list of selected excluded measures and reasons for exclusion is included in Appendix B.
[7] National Research Council. (2008). *Early Childhood Assessment: Why, What, and How*. Washington, DC: The National Academies Press.

of Well-Being of Young Children (SWYC). While the ASQ was designed specifically to identify children who need further evaluation for developmental delays, emerging research suggests that it can also be used as an outcome measure to track change in child development. Given the widespread use of the ASQ, this promising research suggests a unique opportunity to explore the ASQ as a population-level measure of development. The SWYC provides a unique data collection opportunity for states and communities as part of a coordinated measurement strategy for understanding child well-being. While it is designed to identify children in need of more developmental support, it is also integrated in many electronic medical records. Consequently, it has the potential to identify groups of children who need support and to provide information on the link between social determinants of health, such as family risk factors, and children's risk for developmental delay.

Other tools excluded because of their purpose:

- **Tools used exclusively for hypothesis testing and research on child development theories.** These types of tools are designed to explain the process of development (e.g., how early adversity impact social-emotional skills), rather than describing child development. These tools also do not meet our feasibility criteria outlined below.
- **Diagnostic tools.** Diagnostic tools are not designed to capture normative development and often require advanced training to administer.
- **Formative assessment or formative evaluation tools.** These tools are designed to be used by educators to individualize instruction for young children.
- **Tools that do not render scores.**

Feasibility

While many measures of child development exist, few are feasible for implementation across entire populations of children or even across representative samples of children. When selecting measures to include in the scan, we considered cost, time, and other factors that affect the feasibility of large-scale data collection. In addition, we limited the measures included in our scan to those that seemed feasible to collect at least every other year without significant training or administration costs, and whose administration and scoring procedures were relatively quick due to the importance of collecting measurement data regularly to effectively measure progress toward child outcomes.

Measures that take more than 30 minutes to administer, have complex scoring systems, require a highly trained assessor, and/or require observations are likely not feasible to implement at the population level in the United States. The scan includes assessments measured by parents, teachers, or professional reports, as well as direct assessments, because of their low costs and ease of administration. Assessments that are completed by parents, teachers, or professionals who commonly interact with children (for example, medical professionals in primary care settings) are the most cost-effective population-level measures. These types of measurement tools do not necessarily require that the child be present. Direct assessments can pose feasibility issues, including training and administration costs, but some are included in the scan given the availability of technology to support administration. For instance, direct assessments conducted via iPad at doctors' offices could be just as easy and potentially more cost-effective to administer than surveying parents through mail or electronic surveys. Observational assessments have been excluded from the scan because of the prohibitive costs and effort associated with collecting observational data on an entire population or representative sample with any level of useful regularity (annually or biennially).

We recognize that all measures have limitations and feasibility challenges, and that some feasibility challenges are easier to overcome than others (e.g., creating a short form of a measure from a long-form measure; creating a digital/electronic version of an assessment). Individual initiatives, communities, and states will need to consider the feasibility challenges of a given tool or data collection strategy when considering how best to assess young children's development.

Scan of Measurement Tools

In the following section we provide brief overviews of each measure, organized by measurement purpose (population monitoring, program evaluation, screening tool). We then highlight important factors to consider when comparing across the measures including:

1. The status of a measure's development and the strength of psychometric evidence (e.g., reliability, validity)
2. The target age for a measure (i.e., measures that are appropriate for use with infants and toddlers versus preschoolers)

The analysis of the scan concludes with recommendations for choosing a measure, given differences in established validity and the other ideal traits outlined in Table 1.

Overview of measures

Detailed information regarding the characteristics, use, feasibility, and psychometric properties of each measure are presented in Appendix A. In the following section, we provide brief overviews of each measurement tool by purpose: population monitoring, program evaluation, and screening.

Population monitoring tools

Within the United States and globally, momentum is growing for tools to monitor young children's development at the population level. These tools include:

- The **Caregiver Reported Early Development Index (CREDI)** was developed to help address the need for a population monitoring tool of development of children under age 3 in low- and middle-income countries; it has also been tested in the United States and other high-income countries. The CREDI is very similar to the GSED in terms of its design, purpose, and use, and its items are drawn from a similar item bank as those used in the GSED. The CREDI has a long form for program evaluation (see next section) and a short form for population monitoring. The short form was designed to be integrated into household surveys to provide a snapshot of development. The CREDI has been tested in 21 sites in 17 countries and has been tested in the United States in two studies, including one in Boston and one of over 1,000 children via internet data collection. This tool is free to use and score.
- The **Global Scale for Early Development (GSED)** is a promising set of new tools for tracking population-level early development for children up to age 3. The GSED has a long form for program evaluation (see next section) and a short form for population monitoring. The GSED short form is a caregiver-reported instrument intended to monitor child development, track trends, and identify populations in need of support. The pilot measure estimates whether children, as a group, are 'on track' or 'not on track' developmentally. The tool is designed for use in low-, middle- and high-income countries around the world. Backed by the World Health Organization, the tool is currently undergoing pilot testing and validation in several countries around the world. The short form is currently being pilot tested in Nebraska; the goal is to develop population-level estimates of children's development by age 3 in the Omaha metropolitan area. Although the CREDI and the GSED drew items from a similar bank of items, they differ in terms of their methodology; the CREDI was created through an iterative process of item refinement and revision, while the GSED relied on a statistical method that harmonizes hundreds of

items across multiple scales; the final sets of items for the short and long forms were chosen based on duplication and feasibility.[8] This tool is free to use and score.

- The **Intergrowth-21st Neurodevelopmental Assessment (INTER-NDA)** was developed to assess neurodevelopment in young children over time, specifically, to monitor neurodevelopment and identify rates of neurodisabilities. The full package includes modules on vision, hearing, sleep, and circadian rhythms. The INTER-NDA module itself is short, easy to administer, and focuses on cognition, language, fine and gross motor skills, behavior, and social-emotional skills. This tool is free to use and score.
- The **National Outcome Measure of Healthy and Ready to Learn (NOM HRTL)** is a promising new tool that measures whole-child development of children ages 3-5. This parent report measure is included in the National Survey of Children's Health, a nationally-representative annual survey that can generate both national and state-level estimates of whether children are 'on track' or 'need support' for school readiness across four domains. The pilot measure is currently being validated and further refined; additional validity testing, including concurrent and predictive validity (e.g., how does it compare to other tools of school readiness and well-being; does it predict school readiness or school performance?) is needed. This tool is currently not publicly-available, but when it is, it will be free to use and score.

Program evaluation or outcomes tools

Tools used to assess child outcomes and the impact of specific programs or interventions can be useful for characterizing children's development in specific domains, such as language or social-emotional development. Although data from these tools can be used to characterize individual child development, data can also be aggregated across groups of children to generate population-level estimates of the status of children's development in particular domains. Many program evaluation tools did not meet our criteria for feasibility, as many are observational tools, require extensive training, are difficult to administer or are lengthy. The following tools are included in our scan:

- The **CREDI long form** is also a caregiver report assessment that provides domain-specific information regarding children's development up to age 3. The CREDI long form shares many similarities with the GSED long form, in that it is designed for population-level assessment but is intended for use by researchers wanting greater detail about specific developmental domains. This tool is free to use and score.
- The **GSED long form** includes both caregiver report and direct child assessment to capture more detailed information regarding children's development up to age 3. The data from this tool is meant to be used to understand program impacts or to address specific research or evaluation questions; it is not meant for individual child assessment, screening, or diagnosis, but as a population-level assessment. The long form differs from the short form in that it provides greater detail on child development; however, like the short form, it still renders a single score for child development. This tool is free to use and score.
- The **MacArthur Bates Communication Development Inventory (CDI)** is a language assessment for children ages 8–37 months. It has been used as a research and clinical tool to describe children's language and identify children with language delays. The tool is a direct assessment of young children completed by parents. Scores generated from this tool indicate where children fall relative to developmental norms. The tool can take up to 40 minutes to administer. This tool has minimal training and administration costs, and each additional paper assessment costs less than $1.00.

[8] Black, M. (2018). *Development & validation of the D-Score for measurement of Early Childhood Development*. Innovations in Early Childhood Development Assessment [PowerPoint slides]. Available from: https://www.rti.org/event/innovations-early-childhood-development-assessment

- The **Minnesota Executive Function Scale (MEFS)** is a direct assessment of young children's executive function skills for ages 2 through adulthood. Findings from the assessment have been used in research and program evaluation to understand the impacts of policies, programs, or initiatives. The tool can also be used to screen children for executive function difficulties or delays. This direct assessment tool is quick to administer and requires little training. This tool has high training and administrative costs, due to annual re-training of assessors and costs of up to $10 per child.

Select developmental screening tools

Tools designed to identify children needing further evaluation are generally not strong candidates for measuring population-level child development or school readiness, given the limited information they provide; however, we included two screening tools whose purpose has been extended beyond screening. These include:

- The **Ages and Stages Questionnaires** are designed to screen for atypical child development beginning at two months of age. Parents or child care providers generally complete the tool in home, child care, or clinical settings (e.g., pediatricians' offices). The tool can describe child development in five domains, much like the population-monitoring tools. New evidence suggests that the ASQ may function not only as a screening tool but also as a brief developmental assessment that is sensitive to change; preliminary evidence suggests that these change-over-time estimates are best suited to measure changes at a group level rather than to compare developmental outcomes of individual children.[9] The new scoring system that facilitates descriptions of changes to child development is promising because it could be a reliable, quick tool for assessing developmental trajectories of all children ages 2 to 54 months. This work is still under development. This assessment has minimal training and administration costs, and each additional paper assessment costs less than $1.00.
- The **Survey of Well-Being of Young Children (SWYC)** is designed as a developmental screening tool that maximizes the amount of information that can be reliably elicited from parents before they meet with their child's pediatric primary care provider. The SWYC contains several modules or questionnaires that allow pediatric providers to screen for developmental delays, social and emotional problems, Autism Spectrum Disorders, and family risk factors (i.e., social determinants of health). There are questionnaires for children ages 2 months to 5.5 years. The SWYC is unique for its inclusion of family risk factors, which can contextualize screening and developmental findings. The SWYC is also embedded in many electronic medical records, which allow for longitudinal data on large populations. The SWYC is currently untested as an outcome measure, however, there is ongoing work to understand the its policy relevance and predictive validity.[9] Thus, there is potential that the SWYC could be informative beyond its original purpose as a screening tool for a variety of delays and diagnoses. This tool is free to use and score.

Measures are in various stages of development, validation, and use

With the exception of the NOM HRTL and the GSED, all tools profiled in the scan (see Appendix A) demonstrate adequate reliability and at least moderate concurrent validity. The psychometric evidence suggests that the majority of the tools in the scan are ready to be piloted as population-level measures, though it is important to conduct ongoing validation, as construct and concurrent validity may vary by particular subgroups in a community.

The GSED is currently being piloted in Nebraska as a population-level measure of child development in the Omaha metropolitan area. While the findings will reveal more about its validity for use on a sample of children in the United States, findings will also indicate areas that need further validity (e.g., cultural

[9] Bard, D. E. & Hunter, M.D. (2017). *Squeezing Developmental Change out of ASQ-3™ Scores: A Report on Child Development Outcomes for the Home Visiting Data for Performance Initiative*. A report prepared for the Pew Charitable Foundation.

validity). The NOM HRTL is currently undergoing the first stages of validation, including construct validity and invariance across ages and subgroups of children. More work is needed to establish concurrent and criterion validity (i.e., ability to predict or correlate with other similar measures of child development).

The **predictive validity** of the majority of the tools profiled in the scan has not been examined yet. None of the population monitoring tools nor either of the screening tools have been tested for the ability to predict school readiness/later development. Research is underway to discover if the SWYC, administered during early childhood, can predict third grade reading.[10] Predictive validity is a key piece of evidence for policymakers; in order for population-level data on child development to be useful and actionable, it must be predictive of later child development and it must be clear how the data is associated with later school readiness and/or school success. Unsurprisingly, the tools designed for program evaluation and research – the CDI and the MEFS – have demonstrated evidence for predicting school readiness skills. Specifically, the CDI is predictive of later expressive language skills,[11] and the MEFS is predictive of later reading and math abilities.[12,13]

Different measures are available for use with children of different ages

Note that not all measures included in the scan cover the full 0–5 age range. Table 2 indicates which measures can be used for each age group (infants and toddlers; preschoolers), and which offer versions that cover all ages.

Table 2. Measures included in scan by age group designed for measurement

Age group	Tools
Infants and toddlers	GSED; CREDI; INTER-NDA; CDI
Preschoolers	NOM HRTL; MEFS
Versions for both age groups	ASQ; SWYC

Different measures, and potentially different strategies, may be warranted for infants and toddlers versus preschoolers. We explore this idea in more depth on p.13.

[10] Sheldrick, R. C. (2019). *Measuring Young Children's Well-Being and School Readiness: The Survey of Well-Being of Young Children* [Webinar presentation].
[11] Can, D., Ginsbug-Block, M., Golinkoff, R., & Hirsh-Pasek, K. (2013). A long-term predictive validity study: Can the CDI Short Form be used to predict language and early literacy skills four years later? *Journal of Child Language*, 40(4), 821-835.
[12] Reflection Sciences. (2017). *Minnesota Executive Function Scale Technical Report.* Retrieved from https://reflectionsciences.com/wp-content/uploads/2018/12/MEFS-Tech-Report-December-2018.pdf
[13] Hassinger-Das, B., Jordan, N. C., Glutting, J., Irwin, C., & Dyson, N., (2014). Domain-general mediators of the relation between kindergarten number sense and first-grade mathematics achievement. *Journal of Experimental Child Psychology*, 118, 78-92.

Recommendations for choosing a measure

Matching goals to tools

If the goal is to broadly describe child development, then measures such as the GSED and CREDI for children 0-3 and the NOM HRTL for children 3-5 are tools worth considering. However, it is important to note that the GSED and NOM HRTL require additional validation. If the goal includes assessing not only cognition, language, gross and fine motor, and social-emotional development, but other neurodevelopmental constructs, then the INTER-NDA full package is worth considering.

If the goal is to understand specific skills related to school readiness with precision, then research tools such as the MEFS or CDI are worth considering. In addition, the GSED and CREDI long forms may be useful for measuring specific domains of development for children under age 3. Given growing interest in measuring social-emotional development at a population level, the GSED and CREDI long forms' social-emotional domains are worth considering.

If the goal is to maximize the information from developmental screening tools, the ASQ and/or the SWYC may be worth considering as one tool is a larger measurement strategy of young children's development.

The tools included in the scan that are designed for population monitoring, such as the GSED, CREDI, and NOM HRTL, are designed to provide group-based estimates and trajectories of child development that can be compared across subgroups of children. While these measures lack specificity in terms of pinpointing developmental status on exact skills or for individual children, they can provide useful estimates of change over time at the population level. These measures are also designed to be feasible, quick, and easy to administer. They are designed for use at scale, meaning they collect data on representative samples or entire populations of children. They do not provide information on the underlying mechanisms or contextual reasons behind developmental outcomes. The INTER-NDA is the only population-monitoring tool that includes a more thorough assessment of children's neurodevelopmental functioning; however, the modules of sleep, vision, auditory processing may not be of interest to all communities or initiatives.

The tools included in the scan that are designed for evaluating programs or assessing outcomes provide a more thorough assessment of children's skill in a particular domain. These types of assessments may be the strongest measures of indicators of future success in school, particularly among preschool-aged children. The measures included in the scan have strong evidence for reliability and validity for populations of children in the United States. However, given that these tools require direct assessment of young children, they may not ultimately be feasible for implementation at a population level given the cost and time involved.

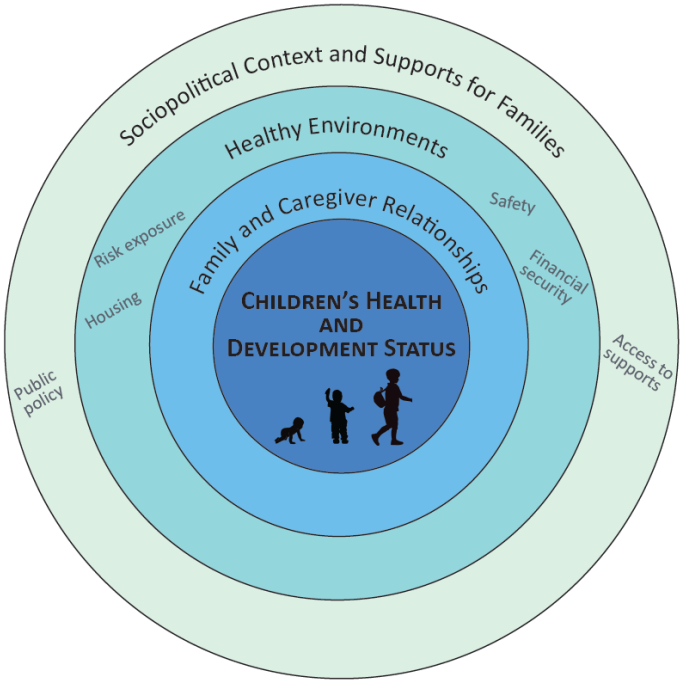
The developmental screeners included in the scan, the ASQ and the SWYC, are low-cost or free, easy to administer, and already commonly used in pediatric settings. Emerging evidence suggests that the ASQ can be scored in a way that provides highly reliable population-level information that is relevant for policy and practice. The SWYC, although not useful as an outcome measure, includes information on family risk factors that may be useful for contextualizing estimates of the need for developmental supports. Both the ASQ and SWYC require additional research on their validity for predicting later development and school success. We emphasize that screening tools are not the ideal, single tool for

population monitoring. They can, however, provide useful complementary information as part of a larger, coordinated data collected strategy for an early childhood initiative.

Part 2: Designing a measurement strategy for evaluating the impacts of early childhood initiatives

Data on child development outcomes alone may not provide enough information to indicate where strategic investments in early childhood should be made. For instance, if community leaders were interested in answering the question, "What supports do children in this community need to promote their development and readiness for school?" or "What family-level risk factors are linked to the child development outcomes we see in our community?", the answers would require a broader understanding of child and family well-being in that community. For this reason, a measurement strategy designed to inform early childhood initiatives would need to capture several aspects of the child's environment, including their parent-child relationship, family dynamics and composition, household and neighborhood safety, quality of non-parental care environments, and the sociopolitical context in their area, which drive access to supports (see Figure 1).[14] Research indicates that each of these factors undergirds development and shapes children's skills, knowledge, and behaviors.

Figure 1. An ecological model of children's development

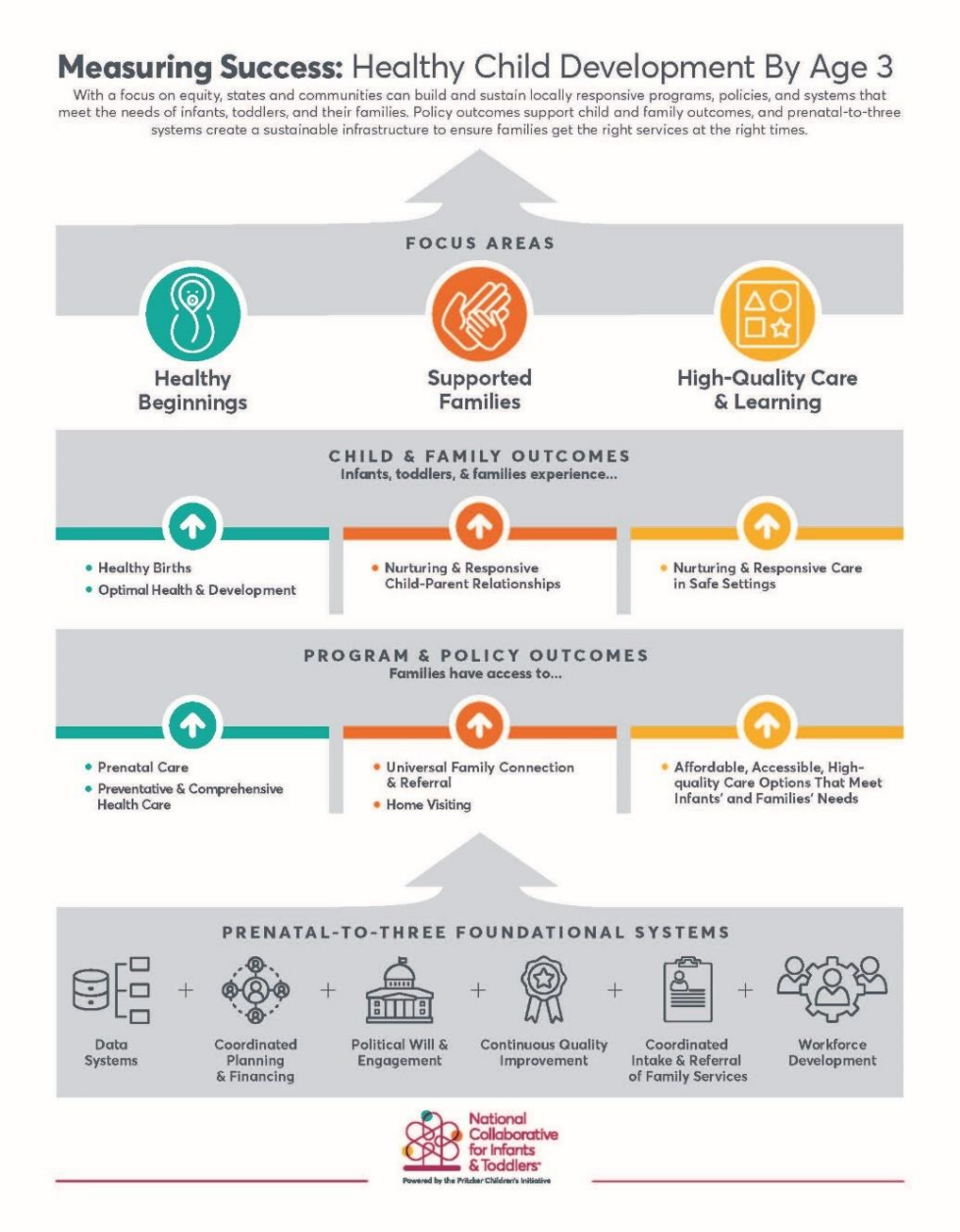


Note. Figure adapted from National Research Council. (2019). *Vibrant and Healthy Kids: Aligning Science, Practice, and Policy to Advance Health Equity*. Washington, DC: The National Academies Press.

[14] National Research Council. (2019). *Vibrant and Healthy Kids: Aligning Science, Practice, and Policy to Advance Health Equity*. Washington, DC: The National Academies Press

In addition, promoting child well-being and school readiness at the policy level requires not only involvement from, but coordination across multiple sectors, as shown in the National Collaborative for Infant and Toddlers' Prenatal-to-Three Outcomes Framework, which identifies healthy child development by age 3 as the key outcome promoted through the promotion of systems, policies, and programs (see Figure 2). A measurement strategy to profile young children's well-being and school readiness in a community should include multiple aspects of child and family ecology not only to characterize well-being and readiness at a population level at multiple points in time, but also to provide sufficient information to inform early childhood initiatives.

Figure 2. National Collaborative for Infants and Toddlers Prenatal-to-Three Outcomes Framework[15]



[15] Child Trends. (2018). Available from: https://www.thencit.org/resources/prenatal-to-three-outcomes-framework

Measuring young children's development at a population-level: Further considerations

Measuring young children's developmental status and school readiness across ages 0 to 5 will require substantial coordination across sectors and coordinated data collection efforts. A successful measurement strategy must consider not only the landscape of current measurement tools, but also the quality of such tools, the purpose and use of the data yielded from such tools, and the infrastructure necessary to collect and analyze the data. Considerations for measurement include:

- **The selected measure(s) should demonstrate adequate reliability, validity (including predictive and cultural), and demonstrate sensitivity to change.** For the data to be appropriately interpreted, it is essential that the measurement strategy rely on measures with the most robust psychometric properties. As populations and contexts shift, however, the validity of certain tools can change. Testing for validity should be an ongoing process. The NOM HRTL and GSED require additional testing to establish validity across contexts, ages, and cultures.
- **Consider the data collection strategy necessary for each type of measure**. Direct observation and parent survey measures will require very different infrastructures, financing, and plans for sustained data collection at regular intervals. In addition, there is no system that reaches all children ages 0 to 5 equally. Consequently, the best data collection strategy for a given community or state may differ, with some relying on data collection in primary care contexts, others relying on early education or K-12 educational contexts, and others relying on household surveys. Differences in strategy may also be tied to differences in the focus of early childhood initiatives (see Figure 2). When considering a measurement strategy, the feasibility of data collection should inform the choice of tool(s).
    - Specific to household surveys, consider the current landscape of state and federal household surveys of children, youth, and families. For instance, the NOM HRTL is administered via the National Survey of Children's Health, an annually administered nationally-representative survey. State-wide and sub-state oversampling can be purchased, which increases the number of completed surveys in a particular geographic region. With the purchase of additional samples, states and sub-state areas (e.g., counties) may be able to report on the NOM HRTL for specific subpopulations in their area (e.g., American Indian/Alaska Native).[16]
- **The precursors for school readiness vary with age**. The greatest predictors of school readiness at kindergarten entry for infants and toddlers is not necessarily their own development, but the quality of the home environment and parent-child relationships. For preschool-aged children, measures of skill, knowledge, and behavior are strong predictors of school readiness. The younger the age of the child, the more that it is necessary to measure the conditions of the child's ecology rather than focusing on direct child outcomes. When selecting a measure of child development or school readiness, consider the outcome of interest and the information it yields for decision-makers.
- **Consider how the chosen measure will yield information that can be used for decision-making**. A dashboard approach, which monitors the status of several indicators of well-being, including those outlined in the National Collaborative for Infants and Toddlers' Outcomes Framework (see Figure 2), may make sense for understanding the contextual or systemic characteristics that contribute to the outcome of child development and/or school readiness.

---

[16] Ghandour, R. & Moore, K. A. (2019. *Measuring Young Children's Well-Being and School Readiness: The Survey of Well-Being of Young Children* [Webinar presentation].

Additional considerations regarding validity, bias, parental involvement, and other issues relevant to implementing a population-level measure designed to capture the development of all children are presented in depth in a National Research Council report on early childhood measurement.[17]

[17] National Research Council. (2008). *Early Childhood Assessment: Why, What, and How*. Washington, DC: The National Academies Press.

Acknowledgments

The preparation of this measurement scan was supported by the Pritzker Children's Initiative, a project of the J.B. and M.K. Pritzker Family Foundation, the David and Lucile Packard Foundation, and the Heising-Simons Foundation.

Appendix A. Measures Scan Tables

Table A1. Purpose and use of each measure

Measure Name	Purpose	Intended Use	Widespread Use	Used/Developed internationally?	Audience
The Caregiver Reported Early Development Index (CREDI)[18]	Serves as a population-level measure of early childhood development.	Short form—Population monitoring: • Assesses development at the population level in a global context. Long form—Program evaluation. • Measures each domain of development in detail; designed to understand change or differences in domains as a result of an intervention or initiative.	21 sites in 17 countries	Created by researchers at Harvard University in the US. Measure development focused on multiple international sites; measure was designed as "intentionally culturally neutral".[18] In use in several sites in the US.	International stakeholders, health administrators, and policymakers
Global Scale for Early Development (GSED)[19]	Measures population-level child development prior to age 3.	Short form—Population monitoring: • Monitors healthy development within and between countries; tracks change over time; compares subpopulations Long form—Program evaluation:	New tools (short form and long form) currently being tested and validated in multiple countries.	Developed and piloted specifically for use in low- and middle-income countries. Constructed to be culturally neutral. Current work includes validation sites in the US.	International stakeholders (e.g., WHO), health administrators, and policymakers

[18] Harvard T. H. Chan School of Public Health. (2019). *Caregiver Reported Early Childhood Development Instruments (CREDI).* Retrieved from https://sites.sph.harvard.edu/credi/

[19] Cavallera, V., Dua, T., Black, M., Bromley, K., Cuartas, J., Eekhout, I., Fink, G., Gladstone, M., Hepworth, K., Janus, M., Kariger, P., Lancaster, G., McCoy D., McCray, G., Raikes, A., Rubio-Codina, M., van Buuren, S., Waldman, M., Walker, S., & Weber, A. (2019). The Global Scale for Early Development (GSED). *Early Childhood Matters*, 80 – 84.

Measure Name	Purpose	Intended Use	Widespread Use	Used/Developed internationally?	Audience
		• Measures each domain of development in detail; designed to understand change or differences in domains as a result of an intervention or initiative.			
Intergrowth-21st Neurodevelopmental Assessment (INTER-NDA)[20,21]	Assesses neurodevelopment in children; designed for use across the world. Assesses cognitive, motor, language and behavioral outcomes. The INTER-NDA is part of a larger suite that measures vision, hearing, and activity.	Developmental screening: • Screens multiple dimensions of early childhood development. Characterizes outcomes across a spectrum in order to screen for neuro-disability in population-based settings.	Population monitoring, large scale screening	Developed internationally; used in Brazil, India, Italy, Kenya and the UK.	Health administrators, policymakers
National Outcome Measure of Healthy and Ready to Learn (NOM HRTL) [22]	Assesses children's health and preparedness for school at population level.	Population monitoring: • Describes children's readiness for school in the years leading up to kindergarten entry • Identifies subgroups of children who are less likely to be on track for school readiness • Can be compared year after year in order to assess progress toward the goal of ensuring that	Part of National Survey of Children's Health, but not yet developed enough for widespread use.	Developed in US; validation in a US community ongoing.	Researchers, communities, states, and federal agencies after the items, scales, and index are assessed and validated.

[20] Intergrowth 21st Project. (2019). *Inter-NDA*. Retrieved from https://www.inter-nda.com/inter-nda.html
[21] Murray, E., Fernandes, M., Newton, C.R.J., Abubakar, A., Kennedy, S.H., Villar, J., & Stein, A. (2018). Evaluation of the INTERGROWTH-21st Neurodevelopment Assessment (INTER-NDA) in 2-year-old children. *PLOS ONE* 13(2): e0193406.
[22] Ghandour, R. M., Moore, K. A., Murphy, K., Bethell, C., Jones, J. R., Harwood, R., Buerlein, J., Kogan, M. & Lu, M. (2019). School readiness among US children: Development of a pilot measure. *Child Indicators Research*, *12*(4), 1389-1411.

Measure Name	Purpose	Intended Use	Widespread Use	Used/Developed internationally?	Audience
		all children are ready for school			
MacArthur-Bates Communicative Development Inventories (CDI)[23,24]	Assesses normative and nonnormative development of language skills. Also screens for receptive and expressive language delays.	Program evaluation: • Tests change over time and group differences in language development. Developmental screening	Public health, research	Developed in the US; adaptations have been used internationally.	Researchers, public health administrators
Minnesota Executive Function Scale (MEFS)[25]	Measures executive function skills in young children.	Developmental screening, program evaluation, research	Research and kindergarten screening	Developed in the US; translated and used in primarily in mid- and high-income countries.	Researchers, school administrators, health administrators
Ages & Stages Questionnaires (ASQ)[26,27]	Identifies children as on track developmentally, or in need of further assessment or evaluation.	Developmental screening: • Comprehensive measure of developmental skills for use home settings, clinics, education and child care facilities, and community settings.	Developmental screening, research, policy, population monitoring; used in primary care/pediatric clinics.	Developed and widely used in the US; used internationally but predominantly in high-income countries;.	Researchers, policy makers, pediatricians, educators, community workers
The Survey of Well-being of Young Children (SWYC)[28]	A publicly available, comprehensive screening instrument.	Developmental screening	State-level program quality monitoring	Developed and used in the US. Teams from Brazil and Nigeria are in the process of developing translations.	Parents, pediatricians, preschool teachers, nurses, other professionals involved in child care and education.

[23] CDI Advisory Board. (2015). *The MacArthur-Bates Communicative Development Inventories (MB-CDIs)*. Retrieved from https://mb-cdi.stanford.edu/
[24] Paul H. Brookes Publishing Co., Inc. (2019). *CDI*. https://brookespublishing.com/product/cdi/
[25] Reflection Sciences, Inc. (2019). *MEFS App - Minnesota Executive Function Scale - Reflection Sciences.* Retrieved from https://reflectionsciences.com/mefs-app/
[26] Moodie, S., Daneri, P., Goldhagen, S., Halle, T., Green, K., & LaMonte, L. (2014). *Early childhood developmental screening: A compendium of measures for children ages birth to five (OPRE Report 2014¬ 11)*. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.
[27] Paul H. Brookes Publishing Co., Inc. (2019). *Ages and Stages*. Retrieved from https://agesandstages.com/
[28] Tufts Medical Center (2010). *The Survey of Well-being of Young Children (SWYC)*. https://www.floatinghospital.org/The-Survey-of-Wellbeing-of-Young-Children/Overview

Table A2. Content and administration of each measure

Measure Name	Developmental Domains	Age Range	Cost	Languages Available	Must be administered by someone with technical background?	Administration Format (e.g. electronic, paper, online)	Reporter	Administration Time
The Caregiver Reported Early Development Index (CREDI)	• Motor development • Cognition • Language • Social-emotional development • Mental health	0-3 years	Free	English, Armenian, Cebuano, Chinese, Filipino, French, Hindi, Llonggo, Japanese, Khmer, Korean, Mandarin, Nepali, Portuguese, Spanish, Swahili	No	Paper, but can be adapted.	Caregiver	<5 minutes for short form, 15 minutes for long form
Global Scale for Early Development (GSED)	• Cognition • Language • Social-emotional development • Motor development	0-3 years	Free	English, Urdu, Kiswahili. Spanish, Portuguese, Dutch, and French translations to be completed in 2019.	No	Electronic and paper	Caregiver (Long form also has direct assessment)	15 minutes for short form, 30 minutes for long form
Intergrowth-21st Neurodevelopmental Assessment (INTER-NDA)	• Cognition • Gross motor • Fine motor • Language • Child behavior • Attention and emotional reactivity (subscales of the Child Behavior Checklist)	18-36 months	Free for non-commercial use. Contact developers for equipment costs.	English, Spanish, French, Italian	Yes, healthcare workers and midwives	Tablet	Direct assessment and caregiver	Full package is 35-45 minutes
National Outcome Measure of Healthy and Ready to Learn (NOM HRTL)	• Early learning skills • Self-regulation • Social-emotional development • Physical well-being	3-5 years	Free	English and Spanish	No	Electronic or paper	Caregiver	10-15 minutes

	• Motor development							
MacArthur-Bates Communicative Development Inventories (CDI)	• Language	8–37 months	Starter kit with forms, manual, technical manual, and CDI-III costs $123.95 from Brookes Publishing. Additional assessments are approximately $0.85	English, Spanish, other adaptations on request	Yes, for scoring	Paper	Caregiver	20–40 minutes
Minnesota Executive Function Scale (MEFS)	• Executive function	2 years and up	Pricing is determined by the number of children to be tested and the number of testers to be trained. Starting costs are $10 per child annually for the app license. Each examiner must take the Examiner Training Course, which costs $250/educator or $175/researcher or clinician. Examiners are required to re-certify each year at a cost of $100/educator ($50/researcher or clinician).	English, Spanish, Dutch, German (Swiss), Swedish, Mandarin Chinese, Somali, Hmong, French, Arabic	No	Electronic	Child	3-6 minutes
Ages & Stages Questionnaires (ASQ)	• Gross motor • Fine motor • Communication • Problem solving • Personal-social	1 month – 7 years	A starter kit containing 21 paper forms and scoring sheets, a PDF file for printing additional questionnaires, a	English, Spanish, French, Arabic, Chinese, Vietnamese; other languages depending on version	No	Electronic and paper, online systems	Caregiver	10-15 minutes

	• General parental concerns		manual and a quick start guide can be purchased through Brookes Publishing for $295. Annual subscription costs are $0.50 per screen.					
The Survey of Well-being of Young Children (SWYC)	• Cognitive development • Language • Motor development • Social-emotional development • Family risk factors • Behaviors suggestive of autism spectrum disorder	1-65 months	Free	English, Spanish, Burmese, Nepali, Portuguese, Haitian-Creole, Yoruba Khmer, Arabic, Somali, Vietnamese	No	Available in electronic, paper, and online formats; can be integrated with medical record systems	Caregiver	15 minutes or less

Strength of the evidence regarding reliability and validity

Criteria for evaluating reliability and validity are derived from Early Childhood Developmental Screening: A Compendium of Measures for Children Ages Birth to Five and summarized in the Table A3. Table A4 is an evaluation of the psychometric evidence for each measure.

Table A3. Description and sources of evidence for reliability and validity criteria

Type of Reliability or Validity	Description and Source of Evidence Used to Establish Criteria	Criterion and Terminology Used
Interrater Reliability	Measured by the level of agreement between two raters when assessing the same children. No established standard in the field.	0.80 or higher=acceptable 0.79 or below=low/weak
Test-Retest Reliability	Measured by correlating the scores on two administrations of the same assessment/developmental screener given to the same child within a short period of time to determine consistency. No established standard in the field.	0.70 or higher=acceptable (across a period of three months or less) 0.69 or below=low/weak
Internal Consistency Reliability	Measured by correlating items within a construct to determine the interrelatedness of the items. No established standard in the field.	0.70 or higher=acceptable 0.69 or below=low/weak
Construct Validity	Measured by examining associations between subscales within the developmental screener. Also measured by examining associations between subscale scores and child characteristics, such as age. No established standard in the field.	0.50 or higher=strong/high 0.30–0.49=moderate 0.29 or below=weak/low
Convergent/Concurrent Validity	Measured by correlating the scores of the developmental screener with scores on other developmental screeners of similar content to determine the strength of relationships between the two.[29]	0.50 or higher=strong/high 0.30–0.49=moderate 0.29 or below=weak/low

[29] Administration for Children and Families (2003). *Resources for measuring services and outcomes in Head Start Programs Serving Infants and Toddlers.* E. Kisker, K. Boller, C. Nagatashi, C. Sciarrino, V. Jethwani, T. Zavitsky, M. Ford, J. Love, &Mathematica Policy Research, Inc. Washington, DC: U.S. Department of Health and Human Services. Retrieved from http://www.acf.hhs.gov/programs/opre/ehs/perf_measures/reports/resources_measuring/res_meas_cdi.html

Table A4. Strength of the evidence regarding validity and reliability

Measure Name	Inter-Rater Reliability	Test-Retest Reliability	Internal Consistency Reliability	Construct Validity	Concurrent Validity	Predictive Validity	Tested in the US
The Caregiver Reported Early Development Index (CREDI) Short Form [Long form is in progress]	Not examined	Acceptable	Acceptable	Moderate	Moderate	Not examined	Testing in progress in the US. Validated with an international sample that includes 17 countries in South America, Southeast Asia, and Africa.
Global Scale for Early Development (GSED)	Validation in progress	Validation in progress	Validation in progress; preliminary evidence suggests internal reliability is likely acceptable	Validation in progress; preliminary evidence suggests problems with socioemotional domain	Validation in progress; preliminary evidence suggests weak evidence	Not examined	Pilot testing underway
Intergrowth-21st Neurodevelopmental Assessment (INTER-NDA)	Weak	Acceptable	Acceptable for all domains except negative behavior	Strong	Strong	Not examined	Not tested in the US; validated with an international sample that includes participants from the UK, Brazil, Kenya, India, and Italy.
National Outcome Measure of Healthy and Ready to Learn (NOM HRTL)	Validation in progress	Validation in progress	Validation in progress	Validation in progress	Validation in progress	Not examined	Yes, tested and validated with a sample of US children
MacArthur-Bates Communicative Development Inventories (CDI)	Not examined	Not examined	Acceptable	Strong, compared to long form	Strong, compared to long form	Early expressive vocabulary skills measured by parental reports on the CDI-SF significantly predicted expressive vocabulary, syntax and semantics, as	Yes, tested and validated with a sample of US children

Measure Name	Inter-Rater Reliability	Test-Retest Reliability	Internal Consistency Reliability	Construct Validity	Concurrent Validity	Predictive Validity	Tested in the US
						measured by standardized direct assessment of these skills four years later; explaining 17%, 11% and 7% of the variance in those skills, respectively.[30]	
Minnesota Executive Function Scale (MEFS)	Not examined; the child completes measure	Acceptable	Acceptable	Strong	Strong	A kindergarten entry MEFS score was shown to be predictive of end of kindergarten reading scores[31] Kindergarten scores predict first grade math scores[32]	Yes, tested and validated with a sample of US children
Ages & Stages Questionnaires (ASQ)	Acceptable	Acceptable	Not examined	Not examined	Strong	Not examined	Yes, tested and validated with a sample of US children
The Survey of Well-being of Young Children (SWYC)	Not examined	Acceptable for the Baby Pediatric Symptom Checklist (BPSC) and Preschool Pediatric Symptom Checklist (PPSC)	Acceptable	Not examined	Moderate for the Developmental Milestones Checklist, BPSC and PPSC	Not examined	Yes, tested and validated with a sample of US children

[30] Can, D., Ginsbug-Block, M., Golinkoff, R., & Hirsh-Pasek, K. (2013). A long-term predictive validity study: Can the CDI Short Form be used to predict language and early literacy skills four years later? *Journal of Child Language*, 40(4), 821-835.
[31] Reflection Sciences. (2017). *Minnesota Executive Function Scale Technical Report.* Retrieved from https://reflectionsciences.com/wp-content/uploads/2018/12/MEFS-Tech-Report-December-2018.pdf
[32] Hassinger-Das, B., Jordan, N. C., Glutting, J., Irwin, C., & Dyson, N., (2014). Domain-general mediators of the relation between kindergarten number sense and first-grade mathematics achievement. *Journal of Experimental Child Psychology*, 118, 78-92.

Appendix B. Selected Excluded Measures

A number of measures were initially considered because they are commonly used, referenced in the literature, or recommended by experts. This scan does not represent a fully exhaustive list of early childhood measures. Table A5 lists excluded measures with short explanations of rationale. For greater detail on each of these measures, please see The World Bank's Toolkit for Measuring Early Childhood Development in Low and Middle Income Countries, Birth to Five's Compendium of Screening Measures for Young Children, and the Early Child Developmental Screening Compendium from Child Trends and the US Office of Planning, Research and Evaluation. In addition, for a list of social-emotional screening instruments and assessment, please see the following compendia: Characteristics of Existing Measures of Social and Emotional Development in Early Childhood and Review of Measures of Social and Emotional Development.

Table A5. Excluded measures

Measure	Exclusion Rationale
Ages and Stages Questionnaires: Social-Emotional (ASQ:SE)	Exclusively a screening tool; no evidence that it can be used to track change in development over time
Parent's Assessment of Developmental Status: Developmental Milestones (PEDS:DM)	Exclusively a screening tool; does not provide information on school readiness
The Early Development Instrument (EDI)	Exclusively for use in kindergarteners; narrow age range
BITSEA (or subscales of the ITSEA)	Exclusively a screening tool