EARLY LEARNING PARTNERSHIP

Guiding Questions for Choosing the Right Tools to Measure Early Childhood Outcomes: Why, What, Who, and How

Adelle Pushparatnam, Jonathan Seiden, and Diego Luna-Bazaldua

INTRODUCTION

A growing understanding of the importance of children's earliest years has led to an increasing desire to measure early childhood development (ECD) outcomes. There are now nearly 150 tools for measuring ECD outcomes internationally,[1] which can make it challenging to choose an appropriate measurement tool for a given measurement effort. These tools vary widely in terms of the:

- purpose for which they were designed (**why**),
- relevant populations and age ranges with whom they are appropriate to use (**who**),
- information about child development they produce including skills, developmental domains, and behaviors they assess (**what**),
- manner in which they are administered (**how**).

This document guides the user through the **why**, **who**, **what**, and **how** questions that must be considered prior to selecting tools for measuring ECD outcomes. Users should document their responses at each step to collate the information needed to identify and select an appropriate ECD measurement tool.

[1] For an inventory of ECD measurement tools, please see the ECD Measurement Inventory that accompanies the Toolkit for Measuring Early Childhood Development produced by the Strategic Impact Evaluation Fund. The toolkit also contains detailed information on how to measure child development for children aged 0-8. This guidance note draws from these comprehensive resources.

STEP 1. Clarify the purpose of measurement: the "why"

Clearly identifying the rationale for collecting information (in other words, the **why** of data collection) is the primary basis of selecting an appropriate tool for measuring ECD outcomes. Unfortunately, this step is often skipped, with users deciding on a tool without first ensuring that the tool is well-aligned with the goals of their measurement efforts.

Different tools are designed for different purposes, such as monitoring ECD outcomes at the population or system level, screening children who are at risk for delay, or evaluating the impact of interventions. A misalignment between the purpose of measurement and the design of the measurement tool can limit the utility of data collected and the conclusions about child development that can be drawn from the data. Further, this misalignment could result in inefficient expenditures of financial resources and time, with users potentially spending too much or too little on measurement efforts given their goals.

Five common purposes of ECD outcome measurement are presented below.

- **Population monitoring** consists of measuring ECD outcomes of a large representative sample of a given population of young children. Data collection usually produces cross-sectional data over different points in time (for instance, a yearly survey of children aged 3 to 5 in a given context). In population monitoring studies, the focus is on aggregated information to describe trends at the population level rather than the individual scores on ECD outcomes of every child. ECD outcomes tend to be just one of many aspects being measured as part of the data collection effort; thus, measurement efforts with this purpose typically require brief, holistic ECD measurement tools.
- **Program/impact evaluation** tests how a policy or intervention affects ECD outcomes. Typically, data is collected at multiple timepoints (with at least baseline and endline measures) and may attempt to follow a sample of children in treatment and control groups longitudinally. Implementers of impact evaluations typically desire ECD measurement tools with high reliability in order to ensure accurate measurement with minimal measurement bias.[2] The domain coverage of the tools will depend on the focus of the intervention and can encompass a holistic set of developmental outcomes or focus on specific skills targeted by the intervention (e.g., foundational reading skills or social emotional development).
- **Formative assessments** are most often used *within classrooms* by teachers/caregivers or school leaders in order to adjust teaching practice, to provide constructive feedback to children, and to offer additional opportunities to promote development and learning. Usually, the results of formative assessments do not leave a classroom/school setting and are used by teachers to offer tailored support to individual children and their class as a whole. Formative assessments are usually repeated frequently and are not used for high-stakes decision making.

[2] Reliability and measurement error are interrelated concepts in psychological measurement. Reliability is defined as the extent to which assessment scores are free of random measurement error; reliability is attained when there is consistency in scores across different administrations of the same tool with the same child, or when different enumerators yield similar test scores for the same child. In the selection process of a tool, ensure that it has documented evidence of its reliability and validity. Experts in psychometrics can provide specialized guidance on these technical issues during the tool selection process.

On the other side, measurement error refers to nonsystematic variability in scores caused by factors unrelated to the developmental domain measured. These factors may include guessing, ambiguity in the assessment administration, or irregularities in the scores assigned by enumerators.

- **Screening for further evaluation or diagnosis** is conducted to identify *individual children* who may be at risk of developmental delays and to help children access further needed services. The results of screening tools alone are usually insufficient to diagnose children, and instead are used to refer children to professionals for further evaluation and support.
- **Research to explore relationships or test hypotheses** is most often conducted by academics and research centers studying how children develop and what factors influence their development. Researchers often require much more intense measurement, often using multiple assessments with the same children on multiple occasions, but typically with the use of smaller sample sizes than in impact evaluation or population monitoring.

An additional consideration that spans the purposes listed above is the need for data that is comparable across different groups and contexts. Data comparability is important when there is the intention to compare ECD outcomes across different populations (e.g., comparing data across countries, regions, or cultural contexts), but is also relevant when there is an intention to compare outcomes across sub-groups within a given populations (e.g., comparing data across gender, ages groups, ethnicities, urban vs rural groups, among others).

If there are multiple purposes that need to be achieved due to varied stakeholders' information needs, multiple approaches and measurement tools are likely required to serve each purpose, as a single tool is highly unlikely to yield optimal data for multiple purposes. The use of multiple tools and approaches will have implications for the time, human and financial resources needed to achieve these multiple purposes. Table 1 at the end of this document illustrates how identifying the purpose of measurement can impact the answers to the other questions when selecting a tool.

STEP 2. Identify the population of interest: the "who"

After identifying **why** data is being collected, the next consideration when selecting an ECD measurement tool is to consider **who** the target population of interest is.

- **Age** is typically the most important factor to consider—many measurement tools are only relevant for a narrow age range (for instance, 0 to 3 and 4 to 6 years old are common age ranges). When multiple age ranges are of interest, or data collection hopes to follow children over time, finding a measurement tool, or multiple tools, with appropriate age coverage is critical.
- **Regional, linguistic, or cultural aspects** of the population should also be considered. Some tools are designed to be globally relevant and available in dozens of languages, whereas others are tailored for use in a particular country, context, or region. Most global ECD measurement tools attempt to capture universal aspects of development but may miss important context-specific aspects of development. In contrast, highly contextual tools offer additional depth but may come at the expense of limited comparability and generalizability of results to other contexts. One approach to balance contextual relevance and global comparability is to embed a standard core set of global items across measurement efforts, and to supplement this core with contextually specific items that address local information needs. Regardless of the tool selected, translation and adaptation activities are often required when using a tool in a new cultural and linguistic context.
- **Developmental status/ability** of the population of interest influences the type of assessment being used. ECD measurement tools designed for use with typically developing children may not be appropriate when assessing children with developmental delays or disabilities.

STEP 3. Map the relevant ECD domains or outcomes: the "what"

The **what** of measurement requires a clear articulation of what kind of scores the user intends to collect and use. ECD measurement tools can generate a holistic **overall score** of development that captures information ECD outcomes across multiple developmental domains, such as psychomotor, language, or socioemotional development. Measurement tools can also generate **domain-specific scores** that focus on a narrow range of specific skills or domains of development.

Some of the domains most commonly covered by ECD measurement tools include:[3]

- **Cognitive skills** including children's memory and problem-solving skills.
- **Language skills** required to express and understand language.
- **Numeracy skills** commonly used to compare quantities, identify and use numbers, and perform basic arithmetic operations.
- **Executive function** and children's ability to control inhibitions, to focus their attention, and to regulate their behavior.
- **Motor skills** including both fine and gross motor skills.
- **Social-Emotional skills** including children's emotional knowledge and conflict resolution.

Tools that generate **overall scores** of development attempt to measure a variety of developmental domains in a single tool. Each measurement tool has slightly different domain coverage, but typically attempt to cover three or more domains. Shorter tools that are often used in population monitoring typically only generate an overall score of development.

Domain-specific scores are generated based on a child's ability on a specific set of skills. Some ECD outcome measurement tools focus on even more specific subdomains within a broader developmental domain, including fine motor skills, expressive language skills, emotion self-recognition, or short-term memory. Users should examine in detail the domains and subdomains that each tool covers; usually this detailed information is included in the assessment framework, reports, and manuals for enumerators. More complex data collections may require multiple tools to ensure adequate domain coverage or to capture developmental trends over time, especially for research projects and some program/impact evaluations. Domain-specific tools are more commonly used in impact evaluations of policies evaluating the effects of specific program/intervention on specific skills, tailored formative assessments, or research projects attempting to deeply understand development in a particular domain.

STEP 4. Consider logistical realities of data collection: the "how"

After clarifying the **why**, **who**, and **what** of data collection, the logistical realities of **how** data will be collected frame important questions about which ECD outcome measurement tool to choose. The factors described below can help to determine which measurement tool is feasible for use in a data collection effort, particularly whether to use a tool that involves the **direct** or **indirect** assessment of the child.

[3] There is no single comprehensive list of all developmental domains identified in the early childhood literature. Table 3.1 of A Toolkit for Measuring Early Childhood Development in Low- and Middle-Income Countries includes nine domains, which can be further subdivided into subdomains. Figure 2.3 in the Toolkit also demonstrates how the relevance of various developmental domains varies across different ages.

Direct assessment tools utilize trained enumerators to engage children in a series of games, tasks or activities following a defined protocol. **Indirect assessment** tools rely on parents, caregivers, teachers, or other stakeholders to answer questions about individual children's development. When possible, the joint use of indirect and direct assessment tools provide a valuable opportunity to triangulate data collected from different sources, which enhances the credibility of the results of a given measurement effort.

When doing both direct and indirect assessment is not an option, some factors that have a role in the deciding whether to use a direct or indirect assessment tool include:

- **Data collection context** can define which assessment modalities are feasible. If data collection will be conducted at ECD centers or preschool classrooms, it may be easier to assess children directly or rely on teacher-reported measures than to survey parents or caregivers. Home-based data collection efforts provide the most flexibility. Electronic or phone-based surveys make direct assessment challenging and typically use parent-, caregiver- or teacher reported assessments.
- **Training** intensity varies depending on the assessment modality of a given tool. While all tools require training to ensure reliable administration, direct assessments typically require longer and more involved training to assure proper standardization, understanding of assessment administration protocols, and quality assurance. Direct assessment tools also often require administrators with more qualifications and/or experience related to measuring child development and interaction with young children. The capacity of data collectors and available time for training influences how complex the administration of the measurement tool can be.
- **Timing and frequency** of planned data collection can influence the choice of measurement tool. Typically, tools involving direct assessment of children taking longer to administer are more useful for less frequent but more in-depth data collections. Data collection efforts that need to be carried out regularly are generally better served by shorter and less resource-intensive tools relying on indirect assessments.
- **Costs of implementing measurement tools** also vary. The resources needed to train enumerators and the time per child needed to implement a given tool vary depending on the complexity of the tool and assessment modality. Direct assessments tend to be more complex than indirect assessments, and thus require more time and resources to train enumerators and to implement the tools in the field. These costs matter particularly within measurement efforts that are meant to be repeated frequently at scale, such as for monitoring efforts and formative assessments. On the other hand, smaller-scale, less-frequent data collection, it may be feasible and even desirable to use a potentially more complex tool.

Finally, when thinking through the **"how⚙"** of data collection, it is worth considering **if the ECD measurement effort could be built into ongoing data collection initiatives** (e.g., existing household surveys, Education Management Information Systems (EMIS), etc.), as this could create efficiencies in terms of the resources needed to collect the data. There might also be **previous ECD measurement efforts** by other stakeholders that the user could draw from. For example, there might be existing data that the user could use, tools that have already been adapted to a given context, or lessons learned from past data collection experiences.

STEP 5. Consolidate information and select an assessment

After documenting the considerations prompted by the above steps, the next step is to identify potential measurement tools that may be fit for purpose. For this step, it may be helpful to refer to the ECD Measurement Inventory for a comprehensive list of 147 tools for children aged 0-8. While there is not a generalizable approach for deciding the exact tool for each situation, the steps above relate to columns in this spreadsheet. By filtering by each criterion, users can identify a subset of measurement tools that meet the requirements for a given ECD measurement effort.

After identifying potential ECD measurement tools for use, review each and complete the following checklist to verify that the tool meets the specific needs of the ECD measurement effort.

✓	The **purpose of measurement** is clearly defined, and the tool was designed for this identified purpose.
✓	The **desired domain(s) of child development** are covered and include a sufficient level of depth to allow for domain-specific reporting if desired.
✓	The **tool covers the relevant age(s).** For longitudinal studies, or repeated measurement of a given sample, it is important to ensure that the selected tool covers the relevant ages across all data collection time points.[4]
✓	The **direct/indirect nature of the assessment** is aligned with the access points of data collection.
✓	The **assessment is relevant** in the cultural/linguistic context of interest, or adequate time and resources exist to translate and adapt the assessment.
✓	There are **resources available** to cover the licensing fees associated with implementing the tool (if applicable; some tools are free and publicly available for use)
✓	**Training requirements** are reasonable given enumerator capacity, time available for training and resources available.
✓	**Data collection cost and time requirements** are reasonable considering the desired frequency of data collection.

CONCLUSION

There has been a proliferation of ECD measurement tools in recent years. The key questions of **why**, **who**, **what**, and **how** laid out above (also summarized in Table 1 below) provide a road map for selecting tools that are fit-for-purpose within a given ECD measurement effort. Choosing an appropriate tool is crucial to the success of a given measurement effort, as it determines whether the measurement effort will yield relevant and credible data that satisfies the information needs of the target stakeholders.

[4] In some longitudinal studies, there might not be a single tool that covers the whole age range of the target population of children across timepoints. In those instances, users may need two or more tools that cover the entire age span of children for the study. At each data collection round, use the specific tool(s) that align in age coverage with the aging of children assessed. To increase the comparability of scores of two or more tools, ensure that these tools measure the same developmental domains and share a few common items. Consult with a psychometrician for technical procedures to produce equivalent scores for your selected tools.

ANNEX

Table 1. Compendium of considerations for ECD assessment selection

Identifying the intended purpose of measurement helps to answer other questions related to assessment selection. Below are five of the most common purposes of measurement, along with a description of how each purpose affects the answers to other questions.

		Population monitoring	Impact Evaluation	Research	Formative	Screening
Purpose of assessment (Why?)		Usually to understand the developmental status of a general population, often to track changes over time	Measure the effect of a program or policy on child development	Generate deeper knowledge of child development and its determinants/ correlates	Help an individual teacher or ECD Facilitator understand the class/group ability to inform and improve practice	Identify children at risk of disability or developmental delay for intervention
Population of interest (Who?)	**Focus of sample?**	Usually a broad representative sample of a given population	Can be either representative sample of a broad population or a specific sub-population	Can be either representative sample of a broad population or a specific sub-population	Usually a small group of children in a childcare or preschool setting	Can be either broad population or a specific sub-population, often at risk of developmental delay or disability
	Age?	Determined by the focus of the measurement effort	Determined by the focus of the measurement effort	Determined by the focus of the measurement effort	More often used for preschool aged children	Determined by the focus of the measurement effort
Desired ECD information (What?)		Usually holistic scores of early childhood development; in some cases specific policy-relevant domains scores may be needed	Can be holistic and/or domain-focused depending on nature of program/policy under evaluation; many evaluations use multiple measures	Can be holistic and/or domain-focused depending on nature of program/policy under evaluation; many research projects use multiple measures	Can be holistic or domain- or skills-focused depending on objectives of Teacher/ ECD Facilitator	Usually focused on a specific disability or developmental delay; some screenings are more general
Logistical realities of data collection (How?)	**Collection setting?**	Often in conjunction with large household surveys	Center-based and home-based data collection both common; center-based data collection more common for older children	Center-based and home-based data collection both common; center-based data collection more common for older children	Center-based data collection more common	Center- and clinic-based collection common; sometimes included in home-based data collection
	Direct/ Indirect?	Indirect more common	Both are common	Both are common	Direct more common	Indirect more common
	Timing?	Annually/semi-annually/regular intervals	Usually at least twice in accordance with program/policy implementation	Usually at least twice in accordance with research questions	Usually conducted multiple times per year	Often conducted at key stages of child development
	Who collects?	Usually trained enumerator teams	Usually trained enumerator teams; some tools require higher capacity enumerators and/or intensive training.	Usually trained enumerator teams; some tools require higher capacity enumerators and/or intensive training.	Usually teachers/ECD Facilitators	Often conducted by trained paraprofessionals or trained enumerators teams

ADDITIONAL RESOURCES

Fernald, L. C., Prado, E., Kariger, P., & Raikes, A. (2017). *A toolkit for measuring early childhood development in low and middle-income countries*. Washington, D.C.: World Bank. Website: https://openknowledge.worldbank.org/handle/10986/29000

Early Learning Partnership (2016). *Measuring child development and early learning.* Washington, D.C.: World Bank. Website: https://documents1.worldbank.org/curated/en/659701473955877219/pdf/108286-REVISED-PUBLIC-ELP-IB4-MeasuringCD-v7-CEP.pdf

Graphics and layout design: Danielle Willis

EARLY LEARNING PARTNERSHIP

For more information on the World Bank's Early Childhood Development work:
http://www.worldbank.org/en/topic/earlychildhooddevelopment#3

For more information on the World Bank's Early Learning Partnership:
http://www.worldbank.org/en/topic/education/brief/early-learning-partnership

This guidance note was funded by the Early Learning Partnership (ELP). We would like to thank Marie-Helene Cloutier, Alisa Katharina Arif Currimjee, Amanda Devercelli, Elaine Ding, and Melissa Kelly for providing feedback on earlier versions of the note. We would also like to acknowledge the support of the broader ELP team, led by Amanda Devercelli under the management of Omar Arias, Practice Manager of the Global Knowledge and Innovation Team of the Education Global Practice at the World Bank.

WORLD BANK GROUP