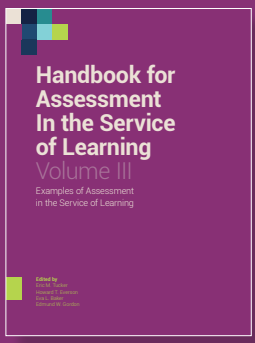
Centering the Voices of Assessment Users in the Advancement of Early Learning Measures

Emily C. Hanno, Elizabeth Mokyr Horner, Ximena A. Portilla, and JoAnn Hsueh

UMassAmherst
University Libraries



Series Editors:
Edmund W. Gordon, Stephen G. Sireci, Eleanor Armour-Thomas, Eva L. Baker, Howard T. Everson, & Eric M. Tucker

ISBN: 978-1-945764-33-2

Suggested Citation:
Hanno, E. C., Horner, E. M., Portilla, X. A., & Hsueh, J. (2025). Centering the voices of assessment users in the advancement of early learning measures. In E. M. Tucker, E. L. Baker, H. T. Everson, & E. W. Gordon (Eds.), *Handbook for assessment in the service of learning, Volume III: Examples of assessment in the service of learning.* University of Massachusetts Amherst Libraries.

Centering the Voices of Assessment Users in the Advancement of Early Learning Measures

Emily C. Hanno, Elizabeth Mokyr Horner, Ximena A. Portilla, and JoAnn Hsueh

Acknowledgments: This chapter is based on research funded by the Gates Foundation.

Abstract

Increasingly young children spend time in formalized learning settings before they enter kindergarten. Although this period can be an impressive time of growth and development for young learners, early education practitioners and leaders often lack easy-to-use, reliable, and valid tools to inform their work. This chapter describes the Measures for Early Success Initiative aimed at developing novel child assessments that accurately capture what all young learners know and can. This chapter begins by introducing the ambitious vision motivating the Measures for Early Success Initiative, describing the goals and features of child assessment tools that are likely to be usable and useful across today's early childhood education landscape. Then it describes the Measures for Early Success Initiative's approach to working towards this vision through inclusive, iterative research and development cycles involving interdisciplinary assessment developer teams working in collaboration with communities across the United States. Initial learnings from this approach underscore the value of integrating user perspectives in the assessment design and development process to ensure tools can be used in the service of learning. In line with principles underlying this Handbook, the chapter highlights promising approaches to support engagement in assessment activities and allow respondents to draw upon their background knowledge and experiences.

Introduction

Early childhood education programs intended to care for and educate children before they enter kindergarten are a promising approach for fostering healthy development and supporting working families. These programs can offer young children complex, dynamic environments in which they can interact, explore, and develop new skills and abilities that prepare them for success in elementary school and beyond (Yoshikawa et al., 2013). The evidence base on early childhood education programs underscores their potential to positively impact children, families, and communities (McCoy et al., 2017), yet it also illuminates challenges of scaling high-quality early learning systems. Not all children have access to the sorts of high-quality early childhood programs thought to confer a developmental boost (Jones et al., 2020), and pre-K-related benefits to children's skills at kindergarten entry tend to disappear quickly during early elementary school (Abenavoli, 2019). Understanding and addressing these unsolved challenges relies on having comprehensive, accurate information on how children's development progresses over time.

Data from assessments that capture children's skills can inform the work of early learning systems, educators, and families in supporting young children's development, as well as identify programs, policies, and practices that allow children to reach their full potential (deMonsabert et al., 2021; Im, 2017). Yet, several key limitations of most existing tools can make it challenging to regularly gather reliable insights into young children's skills at scale. First, most child assessments for early learners focus on narrow sets of skills that are not consistently linked with longer-term indicators of success (McCormick & Mattera, 2022). Second, most tools have been developed and validated with homogenous study samples that are not representative of the children enrolled in public pre-K, which means they may not yield accurate insights about all children (Hsueh, 2021). Third, child assessment data are often burdensome to collect, analyze, and act on in real-world settings. These limitations mean that families, educators, and systems are unlikely to have accurate insights into the strengths and needs of all children, as well as early learning programs. Responding to data from these tools may ultimately exacerbate false narratives about specific subpopulations of children and widen gaps in children's early learning experiences.

This chapter describes recent efforts aimed at addressing these limitations by improving the measurement of young children's outcomes to better meet the needs of assessment users, defined broadly as children, families, educators, administrators, systems, and researchers. Specifically, it outlines the progression and approach of the Measures for Early Success Initiative (or Measures Initiative), a large-scale research and development (R&D) initiative involving collaboration between researchers, practitioners, product developers, and technologists to develop innovative, evidence-based direct child assessments that are usable in and useful for public pre-K settings across the United States.[1]

The Measures Initiative focuses on identifying new practitioner-friendly direct assessment approaches, using methods that collect information from children through standardized tasks or activities as opposed to from observations or work sampling approaches. Tools coming out of this initiative are intended to be used by practitioners to inform instructional decisions but also yield insights that can speak to broader questions about programs and policies. For example, a center-based educator might use the tools to understand children's progress toward early math standards, while program leaders may also use them to consider whether additional math supports are needed program-wide.

The first section of this chapter outlines opportunities for reimagined direct assessments in the areas of content, psychometrics, experience, usefulness, and scalability that serve as the foundation for this work. The second section introduces several novel direct child assessment concepts emerging from the Measures Initiative. The final section describes the iterative, user-centered R&D approach the initiative is taking to develop these concepts into functional assessment products that capture the strengths and skills of all learners and can inform efforts to ensure all children have high-quality early educational experiences that foster meaningful learning. Throughout, the chapter highlights ways that research approaches and design principles of the Measures Initiative can be leveraged in assessment development in alignment with the *Principles for Assessment in the Service of Learning*, particularly in regard to assessment transparency, fairness, and design.

1 Public pre-K settings vary across states and localities. They may include public schools, child care, Head Start, and home-based child care.

A Target Product Profile as the Foundation for the Development of New Direct Child Assessment Tools

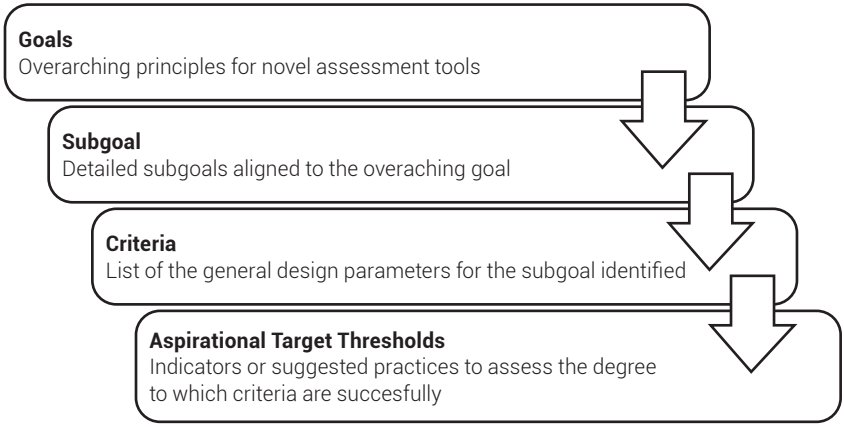
The Measures Initiative is driven by the early learning field's need for easy-to-use child assessment tools that reliably capture the widespread competencies of young learners and provide actionable insights. As an initial step, the initiative convened a wide array of users of early learning assessments to imagine the specific design features of measures meeting this vision. This process involved interviews with parents/caregivers, educators, and program leaders, as well as working sessions with academic experts in assessment, early childhood education, and developmental psychology to understand specific needs and desires for improved child assessment tools. These insights were then synthesized and organized into a target product profile (TPP). TPPs are commonly used in the health sector to outline goals, requirements, and specifications to inform the development of health care solutions such as medications and vaccines. TPPs for pharmaceutical products often include specifications for aspects like delivery mode, dosage, risks/side effects, and cost (Tebbey & Rink, 2009). Adapting this TPP approach for educational products has the potential to similarly encourage innovative solutions driven by user and market needs. Moreover, it offers a framework for reflecting on existing tools to identify their strengths, gaps, and areas for future improvement.

Developing a TPP for early learning assessments, as compared to for health products, posed unique challenges and opportunities. Whereas TPPs for pharmaceuticals typically outline parameters in set categories (e.g., dosage), the categories for key features of early learning assessments were not predefined. Similarly, whereas there are often agreed upon standards for successful medical tools (e.g., shelf stability; minimal counterindications), there were few agreed upon metrics of success for early learning assessments. The initiative's emphasis on assessment user perspectives and experiences also meant that the input received on the ideal features of child assessment tools prioritized by different engaged users was at times conflicting. For example, some preferred assessments that children could complete entirely independently to minimize teacher burden, whereas others desired tools that involve teachers to ensure they actively observe children demonstrating skills and behaviors.

Establishing a clear organizing taxonomy for early learning assessment products from thousands of user insights and comprehensive review of existing early learning assessment products and literature relied on thematic analysis. From

this approach, five areas of focus–or goals–for the next generation of early learning assessments emerged: (1) content, (2) psychometrics, (3) experience, (4) usefulness, and (5) scalability. Specific criteria elevated during interviews, focus groups, feedback sessions, and literature reviews were then organized into these five categories. Example aspirational target thresholds aligned to each criterion were generated as potential metrics to signal whether a product was progressing toward that criterion. Thresholds were designed as aspirational because, in some cases, it is unclear whether they are possible to achieve (e.g., fully offline capabilities may not be technologically feasible; it may not be possible to have brief assessments that also comprehensively cover content). Exhibit 1 describes the structure of the child assessment TPP, entitled the User-Informed Principles (MDRC & Substantial, 2022), and Exhibit 2 provides examples of subgoals, criteria, and aspirational target thresholds within each goal.

Exhibit 1. Taxonomy of the User-Informed Principles



The first goal of the User-Informed Principles describes aspirations for assessment **content** or the skills and competencies measured by early learning assessments. This section takes an expansive view on the content that child assessments should capture, emphasizing the importance of content breadth (skills across developmental domains captured) *and* depth (skills within developmental domains comprehensively reflected). Developmental domains reflect those typically included in whole-child frameworks for early learning such as the Head Start Early Learning

Outcomes Framework (Office of Head Start, 2015). This expansive perspective on assessment content stands in contrast to the common practice of focusing narrowly on assessing a subset of skills within foundational academic domains like math, language, and literacy. For example, most measures of young children's language abilities tend to center on receptive vocabulary. Although vocabulary is a foundational skill, typical vocabulary prompts testing which words children know do not reflect the full range of language skills young children are developing and ultimately need to navigate through the world. This narrow focus also advantages children who have received formal vocabulary instruction or who regularly hear and use commonly tested vocabulary words in daily life, but disadvantages those whose language strengths lie in other areas (e.g., oral storytelling). Assessments should give all children an equal opportunity to demonstrate their full range of skills and not rely on specific experiences or knowledge that are unlikely to be universal.

The remaining four goals of the TPP are relevant to assessments covering any content area. The second goal of the User-Informed Principles outlines **psychometric** properties of child assessment tools intended to ensure that they generate reliable estimates of children's skills and reflect minimal statistical bias. All assessment scores inherently include measurement error that does not reflect children's skills in the constructs intended to be measured by the tools. Therefore, this goal includes criteria for acceptable levels of measurement error in line with field standards for internal properties of educational assessments (AERA et al., 2014), but also emphasizes properties related to the fairness of the tools or their ability to generate comparable information across time, target constructs, and communities of children. A particular challenge with many existing early learning assessment tools used for both formative and summative purposes is that they rely on ratings of children's skills by an adult caregiver, typically an educator or parent. Scores from these assessments are likely to be subjective, reflecting the perspectives of those assigning the ratings, including their knowledge of early childhood development and any implicit biases about specific subpopulations of children (Cameron, McClelland et al., 2023; Gardner-Neblett et al., 2023; Russo et al., 2019). Parameters in the psychometrics goal describe features of tools that minimize the influence of rater bias on assessment scores. As a flexible framework, the TPP offers developers the opportunity to prioritize psychometric features most aligned with the intended uses of the tools they are designing.

The third goal describes parameters for the **experience** of children taking the assessment and the educators who are often responsible for using the tools to collect data. All too often, assessments are a source of stress and burden for the students and educators using them. Traditionally, direct assessments with young learners have required one-on-one educator-child sessions. For children, the repeated questioning format of these traditional one-on-one direct assessments (e.g., "What is this called?" "What color is this?") can be uncomfortable, particularly for those from communities or cultures in which these types of interactions with adults are uncommon for young children (Peña & Halle, 2011). For educators, collecting child assessment data in this way with young children can detract from their ability to engage in instructional activities in their classrooms. This is also the case when using observation-based tools, which requires extensive educator time to document and rate anecdotes on children's behaviors (Cameron, Kenny et al., 2023). This section therefore outlines parameters for direct assessment tools that are fun, engaging experiences for children to partake in and that are intuitive for educators, requiring minimal time for training and use. It also underscores the importance of these tools' integration into normal classroom routines and practices, such as free choice time and small group instruction.

The fourth goal documents properties of the data outputs from early learning assessment tools reflecting their **usefulness**. Collecting assessment data is only as meaningful as the actions the data can inform. For educators and families, assessment data can inform decisions about how to best support individual children to be successful. For early education systems, these data can inform decisions about the most effective approaches to improve early learning experiences in ways that ultimately foster positive child outcomes. Parameters in this goal underscore the importance of making data outputs timely, understandable, and actionable for a variety of purposes. It particularly emphasizes the need to make data accessible for families who typically receive limited information on their children's skills. It also highlights the potential for child assessments to serve as a conduit for collaborative communication between educators and families about children's development.

Exhibit 2.

Example subgoal, criteria, and aspirational target thresholds for each User-Informed Principles goal

Goal	Content	Psychometrics	Experience
Subgoal	1.4 Assessments capture children's skills in objective, strengths-based ways.	2.1 Assessments generate valid, psychometrically sound, and useful information for multiple purposes.	3.2 Assessments can be integrated into everyday classroom activities seamlessly.
Criteria	1.4.1 Assessments capture measures of children's learning across target domains that minimize reporter bias.	2.1.1 Assessments generate comparable construct-specific scores—with high levels of content validity as described in prior goals—across groups of 3-, 4-, and 5-year-olds.	3.2.1 Administration of the assessments can be embedded within typical pre-K routines and does not take away from other activities.
Aspirational Target Thresholds	• Assessments primarily rely on direct assessments to capture children's learning, development, and competencies. • Assessments can provide opportunities for educators to report on children's development as a supplement to direct assessment information.	• Assessments capture growth relative to a criterion (i.e., what children know and are able to do) developed specifically for priority groups with a representative sample of 3-, 4-, and 5-year-old children from diverse settings and geographic regions of the United States. • Criterion-referenced standards are available for each domain of learning and competency within-age for children ages 3, 4, and 5. • Domain scores can be compared across ages to examine growth relative to criterion-referenced standards. • Assessments yield reliable and valid scores within each age group (3, 4, 5).	• Time spent in typical instructional activities is largely unchanged (or potentially increased) before and after the take-up and implementation of the assessments in diverse pre-K settings. • Assessments are designed to be used in a variety of activities throughout the day (e.g., individual choice time or project-based time).

Exhibit 2. (continued)

Goal	Usefulness	Scalability
Subgoal	4.1 Assessments regularly generate meaningful and actionable information about children's learning, development, and competencies in separate early learning domains for multiple purposes.	5.1 Assessments are affordable for publicly funded pre-K systems and centers to administer. (Feasible price and time burden target points are currently being determined through discussions with pre-K system leaders, program administrators, and educators.)
Criteria	4.1.1 Assessments produce results that can be used to identify how children are learning and tailor instruction to support children's development.	5.1.1 There are low costs and burdens to adopt the assessments for pre-K systems and programs.

Aspirational Target Thresholds	• Assessments produce results for each child at least 6 times—or as frequently as needed by the educator to support an individual child's development—during the program year that:
	  • Can produce point-in-time holistic profiles for child development across multiple domains.
	  • Can produce reports on individual children's growth and areas for supported learning in domain-specific areas from one assessment period to the next, from the beginning of the year to the most recent assessment, and from the beginning to the end of the program year.
	  • Can produce reports on individual children's performance relative to overall classroom/group performance.
	  • Can suggest groupings of children with like abilities or mixed abilities in small groups.
	  • Can produce reports on overall classroom/group performance across multiple domains.
		• Cost of initial take-up is reasonable and feasible as agreed on by a panel of program administrators, center directors, and policymakers (costs here include the hardware and software costs to start up, and staff time to learn a new system of assessments, such as training time for educators and administrators, educators' and administrators' time to review and interpret data, and costs for IT support staff to launch, divided by the total number of children in a program or system).
		• Families, educators, and administrators in diverse early learning settings perceive the benefits of taking up and collecting the assessments to outweigh the costs of doing so after having used the assessments for at least 6 months.
		• Families, educators, and administrators are able to understand information from reports quickly and efficiently. Panel of families, educators, and administrators agree (> 80%) that implementing assessments does not detract from time spent with children or typical learning activities.

The fifth and final goal outlines aspects of tools that would ensure their **scalability** across the diverse landscape of publicly funded pre-K programs. Today most states have what is known as "a mixed delivery system" of publicly-funded programs that ranges from formal group-based settings like those within traditional public schools to informal settings like those in family child care programs typically in someone's house (Jones et al., 2020). The resources, strengths, and needs of children, educators, and administrators across these various settings greatly varies (Hanno et al., 2021). Recognizing that technology-based solutions are likely to facilitate meeting other aspects of the User-Informed Principles (e.g., covering more expansive content, improving child and educator experience), this section imagines the features of technology-enabled assessments that are accessible and implementable across contexts. It includes parameters for required infrastructure and hardware, internet connectivity, data privacy, and interoperability with other commonly used education technologies.

Together, these goals set an ambitious vision for novel child assessment products informed by the experiences and perspectives of assessment users. In contrast to some TPPs that outline minimum product specifications, the User-Informed Principles are collectively envisioned to be an aspirational target that is intentionally not prescriptive about how to accomplish individual criteria outlined within them. Assessment developers grappling with this document therefore have enormous opportunities to innovate and tackle different aspects of the User-Informed Principles with far-ranging solutions. They also are confronted by inherent tensions across and within goals: How can assessments have content that is relevant to children from different communities while also generating scores that are comparable across groups of children? How can assessments be expansive in the content they cover while also minimizing child and educator burden of collecting assessments? The User-Informed Principles are intended to empower developers to address these challenges head on with the end goal of spurring breakthrough innovations in the child assessment space.

Innovative Concepts to Advance Towards the User-Informed Principles

After developing the User-Informed Principles, the Measures Initiative moved into the next phase of work to identify and develop innovative solutions that move towards this ambitious vision for child assessments. The Measures Initiative first identified organizations with relevant expertise in early childhood, child assessment, and technology. Aligned organizations then had the opportunity to interact during in-person working sessions intended to introduce the User-Informed Principles and encourage ideation around how to further the goals outlined in the document. With these sessions as a shared foundation, organizations with complementary capacities worked collaboratively to develop initial concepts for novel direct assessments for use in public pre-K classrooms to inform instruction and decision-making. This section introduces three of the early-stage assessment concepts that emerged through this process, describing how they prioritize specific aspects of the User-Informed Principles and challenge existing assessment paradigms. All three assessments are focused initially on content areas that have traditionally been assessed through direct approaches: language, literacy, math, and executive function.

One team comprised of individuals from the Universities of Minnesota and Oregon, Aviellah Curriculum and Consulting, and FableVision Studios, an educational media and technology company, envisioned a tablet-based digital storybook-based assessment approach wherein children navigate through interactive narratives reflective of varied experiences and, while doing so, respond to prompts that capture their abilities across a broad range of skills in English and Spanish. As part of this work, the team hypothesized that individual prompts could simultaneously capture multiple skill domains. That is, the same item could shed light on children's math and receptive language at the same time. This sort of multi-dimensional assessment approach may more accurately reflect the integrated nature of children's development and skills across domains, as well as allow for measuring more content domains with minimal burden. Early-stage research with elementary schools has demonstrated the promise of engaging, multi-dimensional tools to capture children's creativity and cognitive skills (Rosen et al., 2023). This concept extends this approach by considering how it might work with younger preliterate populations, new skill domains, and within engaging storybooks.

An alternative hypothesis proposed by another Measures Initiative team is that short tablet-based games can expand content measured while also improving children's experiences taking assessments. This hypothesis is posed by Khan Academy Kids, an education technology non-profit that currently has a free, widely used learning application containing thousands of interactive learning activities for young children ages 2 to 8 based in the Head Start Early Learning Outcomes Framework and Common Core Standards. Khan Academy Kids' proposed approach of creating short engaging assessment tests styled after their existing learning activities seeks to break down the silos between play, learning, and assessment. As part of their work, this group is also considering whether their short, embedded assessment tasks can yield both formative and summative insights. That is, can these brief assessment modules be flexibly administered, adapted, and scored to both inform teacher practices and instruction, as well as monitor children's learning and generate large-scale summative assessment scores to support continuous program improvement?

A third assessment concept explored through the Measures Initiative is whether children's skills can be accurately captured as they navigate through physical books and respond to items about book content. Picture books are commonplace in early childhood classrooms and are used throughout the day during whole group circle time, small groups, and independent learning centers. This team, led by learning technology and educational experts at Kibeam and Mighty Play, has developed a book-based assessment learning and reporting system. Children navigate through books using a small handheld device or "wand" developed by Kibeam. The wand is equipped with sensors that read images and text on a page and a speaker that can read aloud text and interactive assessment prompts on each page. This physical technology not only allows preliterate readers to independently engage with written text in new ways, including kinetic interaction, it also offers a potential approach for asking and recording children's responses to prompts related to content on pages. The team proposed exploring the feasibility of continuously collecting data to enable embedded, ongoing assessment through diverse picture books. This novel tool represents a potentially joyful, engaging way of collecting data naturalistically as children engage with activities (i.e., book reading) that they would already typically do in pre-K classrooms.

Each of these assessment concepts represents a unique hypothesis about how to improve data collection on young children's skills and abilities. They advance different assessment methods (i.e., digital storybooks, technology-enabled short learning activities, and a handheld book reading device) intended to engage children and reduce teacher burden when conducting assessments, while also developing content for emergent bilingual preschoolers learning Spanish and English. Across the initiative, each developer team has also considered whether recent technological developments might buoy their efforts to progress toward the User-Informed Principles. Rapid advancements in artificial intelligence (AI), in particular, offer unique opportunities to address various limitations of existing early learning assessments. These technologies could help efforts to broaden and deepen assessment content. For example, generative AI models could quickly develop expansive assessment content or vignettes within which to embed assessment items. Improvements in automated speech recognition (ASR) capabilities may mean that young children's expressive language can be accurately captured and analyzed, further expanding the types of content direct assessments are able to measure. These new technologies may also have implications for educators using the tools, providing them with in-the-moment guidance on how to more efficiently and objectively capture, analyze, and react to data.

Despite the promise of AI for improving various aspects of early learning assessments, these new capabilities remain largely untested and could unintentionally exacerbate existing assessment challenges without careful consideration and monitoring (Ho, 2024). AI models rely on training datasets to learn how to process and generate information. The perspectives and experiences reflected in the training datasets will therefore filter into what is produced by AI models. This could create challenges for the quality of insights generated from ASR-based assessment prompts if ASR models are generated with a limited set of voices. For example, models built with training data comprised of adult speech samples are unlikely to reliably capture the unique speech patterns of young children (Patel & Scharenborg, 2024). Relatedly, models built with datasets that prioritize specific dialects may not accurately capture the speech of those who speak excluded dialects (Wassink et al., 2022), likely resulting in incorrect estimates of certain communities' expressive language abilities. These risks underscore the importance of examining the consequences of integrating AI technologies into assessment products through research conducted in partnership with diverse communities.

Active Engagement with Assessment Users in Iterative R&D

The Measures Initiative employs an iterative research and development approach to ensure early-stage tools and technologies building from these assessment concepts are progressing in line with the priorities and needs of assessment users. Turning an assessment concept into a functional product requires frequent, ongoing decision-making on the part of assessment developers. These countless decisions range from the macro (e.g., the content domain(s) to assess) to the micro (e.g., the color schemes to use in data dashboards; the words to use to direct children to the next assessment item). In traditional assessment development paradigms, assessment developers largely make these decisions based on their personal experiences and expertise. In the absence of external voices, the developers' perspectives are then naturally reflected in the assessment product and the data that come from it. For example, as noted above, measures of young children's language skills often focus on receptive vocabulary tasks that examine children's ability to understand a list of words spoken out loud, meaning that these tools reflect a narrow conceptualization of language skills (excluding expressive, syntactical, and social components; Portilla & Iruka, 2024). Despite the narrow perspectives drawn on during initial development and validation work, these measures are now commonly used across the United States in socio-demographically diverse samples for research and monitoring purposes. Given how the tools were developed, differences in scores across subgroups on these vocabulary measures may indicate differences in the relevance of the tools across populations rather than underlying skill differences. That is, these tools afford some but not all children the opportunity to show what they are learning and doing in their homes and communities, undermining the fairness of assessments.

In contrast to this dominant approach, the Measures Initiative has sought to integrate user perspectives by drawing on the expertise of communities served by today's public pre-K programs at every step of the tool development process. Over time, tools developed through the initiative have evolved from concept ideas to prototypes to functional assessment products with comprehensive item banks that cover multiple content domains. This means that the assessment components and functionalities that the developers are co-designing and pressure testing with assessment users over time exponentially grow. The quantity and variation of assessment features being developed also means that no one research activity is likely to be sufficient to build evidence on the extent to which they meet user

needs, accomplish criteria outlined in the User-Informed Principles, or align with field standards for educational assessments. That is, the same research activity is unlikely to yield meaningful insights on a set of 100 math items as on a new data dashboard prototype.

With these considerations in mind, the Measures Initiative has implemented an iterative, cyclical, and phased R&D approach involving assessment users representing a range of communities in a variety of research activities intended to generate insights and evidence on different aspects of early-stage assessment tools. Assessment user groups represented across different research activities include pre-K students, educators, parents/caregivers, and program administrators. These proximal assessment users–those expected to take assessments or collect and use data from the tools–are recruited to Measures Initiative activities through the initiative's close partnerships with local agencies that have longstanding relationships with early education programs (e.g., recruitment and referral organizations; technical assistance providers).

By working with organizations with system-wide purview, the initiative engages individuals from a broad range of program types across different geographic contexts. The pre-K landscape is notoriously fragmented in the United States with families and children relying on a variety of publicly funded and subsidized education and care types (e.g., Head Start, public school-based pre-K, community-based centers, and family child care programs). Yet, contemporary early childhood research rarely reflects this diverse, patchwork landscape, often constrained to a narrow set of classroom-based programs predominantly in cities (Jones et al., 2020). In the Measures Initiative, having a broader perspective on the early childhood landscape is particularly important for exploring how the tools might work across early learning programs with educators who have different professional experiences (e.g., education levels, certification), instructional supports (e.g., coaching, curricular materials), and technological resources (e.g., high speed internet).

The initiative has also built a network of field leaders who bring practice, policy, and academic expertise to engage in its R&D process. Although these individuals may be less likely to use the tools directly than those recruited from programs, they are often consumers of assessment data from these tools and make or influence decisions about the assessment tools used in publicly funded pre-K programs. Moreover, they each bring valuable expertise relevant to different

aspects of tool development. For example, the academic experts typically bring extensive knowledge on children's skill progressions in early childhood, with many contributing to state and federal early learning standards that guide what early learning programs are hoping their students will learn. In contrast, many of the practice and policy experts have experience acquiring and rolling out new assessments at the systems level.

Research activities with this broad range of assessment users are then meant to provide opportunities for authentic co-design between developer and user, as well as yield insights about tools' progress towards the User-Informed Principles' goals (i.e., content, psychometrics, experience, usefulness, and scalability). Importantly, not all users engage in all activities or are asked to weigh in on all aspects of the tool. For example, pre-K students are important partners for designing the experience of using the tool in real world classroom conditions but are appropriately not tapped to evaluate an assessment's scoring procedures. Similarly, not all research activities occur at every phase of tool development or are expected to provide insights on every aspect of the tools. As tools become more developed with increased functionalities, additional research activities are layered on to more comprehensively evaluate tool features. Below is a list of the types of research activities that the Measures Initiative has integrated into its R&D process, describing the user groups engaged and the assessment areas interrogated through each activity:

Focus groups and feedback sessions: Starting with their concept designs, developers have met with small groups of users to share materials they are developing and iterate on those materials with users. Most often these groups are comprised of educators and administrators or parents/caregivers, but teams also occasionally meet with practice, policy, and content experts either in small groups or individually. These conversations often focus on getting users' thoughts on assessment content, assessment interface or plans for how the tools might be used in classrooms (experience), or data dashboards (usefulness).

Content vetting: Once developers have initial assessment content (e.g., construct maps, item banks), teams of academic experts with deep knowledge of early childhood education, child development, and measurement holistically evaluate the assessments, providing thoughts on each tool's general approach to content, experience, and usefulness. They also review each individual assessment item,

noting whether they believe it is capturing the intended domain/subdomain and is developmentally appropriate for pre-K students.

Cognitive interviews: Once developers have early-stage prototypes of their assessments, they can conduct one-on-one cognitive interviews with children and educators to observe and learn how they navigate the tools (experience) and interpret the item prompts (content). In the case of educators, these interviews also often include having educators review data outputs to evaluate whether the outputs allow educators to accurately interpret scores and make well-supported instructional decisions from the data (usefulness).

User testing: Once developers have functional early-stage assessment tools, they can have small samples of educators use their tools in real world pre-K classrooms over an extended period of time (e.g., several weeks or months). This gives developers the chance to iterate on the training, implementation materials, and ongoing supports they provide to educators using their tools. It also provides insights into what it is like for children and educators using the tools (experience) and whether these experiences differ across settings (scalability).

Psychometric analysis: Assessment data collected through user testing is used to evaluate the quality of the information and, ultimately, scores coming from the tools. At the earliest stage when user testing is constrained to small samples comprised of a handful of classrooms, item-level data are examined to ensure items have varying levels of difficulty and are not likely to produce scores that suffer from ceiling or floor effects. Over time, as user testing involves larger samples, psychometric analyses can become more complex, examining psychometric properties like scale reliability, differential item functioning, and correlations with established measures. They can also be used to build evidence on scoring procedures like stopping rules or computer adaptive testing approaches that can help reduce the number of items children must complete.

Over time, developer teams compile insights from these various co-design activities repeated with many assessment users. In some cases, suggestions and solutions raised in these activities are quickly implementable. For example, when children tested out several of the tools during cognitive interviews, they often forgot to click the on-screen or physical button required to advance to the next item. Based on that observation, developers were able to quickly program consistent

prompts reminding children how to move forward in the assessment and then were able to determine whether those prompts helped keep children advancing through items while observing them in their next round of interviews. Other times, teams are required to synthesize and digest multiple and at times contradictory inputs on the same assessment features. This sometimes occurs within research activities: focus group participants disagree with each other or content vetters rate the same item differently. Other times, teams get contrasting feedback from different research activities. User testing might illuminate the need for assessments to be shorter to more seamlessly integrate them into classroom schedules and sustain child engagement, while psychometric analyses might suggest additional items are needed to yield more reliable estimates of children's skills. Teams are encouraged to directly grapple with these contradictions, recognizing that although there is unlikely to be a single correct path forward, this iterative co-design and development process offers opportunities to rigorously test innovative solutions and understand how assessment users experience and perceive them. Teams are also encouraged to critically reflect on how the solutions they initially propose might reflect their own experiences, expertise, and preferences rather than those of the assessment users they have partnered with. This is in service of continually seeking to prioritize the experiences of those who will ultimately use and be affected by the tools such that they can be usable, useful, and generate accurate insights on all children's abilities.

So far, this approach has brought particularly valuable insights into assessment design features that support young learners' ability to successfully demonstrate what they know and can do. Traditional direct assessment approaches for older students—like pen and paper or computer-based assessments—are not developmentally appropriate for pre-K-aged children who are often preliterate and lack computer skills to navigate through a computer-based assessment (e.g., mouse handling, typing on a keyboard). Consequently, the data collection approaches (i.e., tablet-based or handheld device) tested through the Measures Initiative are relatively new.

Even with strong grounding in developmental science, it can be challenging to predict all the ways young children might interact with new technology-based direct assessment interfaces that might affect data quality. For example, we did not anticipate that young children, when given the chance, would complete the same assessment game or story multiple times without explicit controls preventing

them from doing so or that some children would figure out how to complete assessments under other children's profiles. By observing these behaviors through user testing, developers are able to design new approaches to ensure children can focus on the assessment task at hand rather than being distracted by unintended ways to engage with the tool.

Similarly, developers have begun to identify assessment features that foster children's motivation and continued engagement during administration. Children appreciate the opportunity to have choice, such as being able to select which assessment tasks to complete first or the character that provides instructions during the assessment. Varied, child-friendly design elements—including high-contrast colors and familiar assets—appear to encourage sustained focus during tasks. During user testing, children remarked about familiar or favorite items used in assessment prompts (like manipulatives in patterning tasks). Varying item response modalities, such as those that allow children to verbally or kinesthetically respond to prompts, can also keep children engaged and prevent mindless tapping through items. Although, importantly, these novel response approaches must come with clear and simple instructions about what to do. Children not only need explicit guidance on how to respond to assessment prompts, but also how to navigate the assessment application interface. For example, digital assessments often include a button to click to advance to the next item, which although intuitive for older children and adults, is not familiar for most young children. All assessments coming out of the Measures Initiative include brief training modules to acclimate children to the maneuvers they will need to know to navigate through the assessments.

User-testing has also provided insights into how to make assessments more usable for early educators. Given that pre-K classrooms are dynamic environments with lots going on, educators have requested the option to pause assessments midway to allow children the opportunity to start back where they left off rather than having to repeat completed items. This could accommodate common short interruptions like bathroom breaks. Future testing through the Measures Initiative will explore whether these types of short pauses affect student performance on assessments. Teachers also requested that initial training materials include more concrete guidance on how to set up and use the assessments in their classrooms. This includes how to store, charge, and turn on technology; how to connect devices to the internet; and how to use the tool during different instructional formats (e.g., small groups, centers). These early insights illuminate the importance of

considering user perspectives in assessment design to align features and supports with what will work in real world settings and give children the best opportunity to demonstrate what they know and can do.

Conclusions: Towards a New Paradigm for Assessment Development

Young children have an incredible capacity to build skills and learn new things. The ability to foster this development in early learning programs rests on understanding where children are in their development to best tailor supports and how they grow over time to unearth the best ways to help them advance. This chapter described the ambitious goals of the Measures for Early Success Initiative to build better tools for early education programs that can support high quality early learning experiences for all young children. Beyond the goal of this work to produce new assessment tools, the initiative also advances a vision for assessment development that centers the experiences and perspectives of assessment users rather than assessment developers. This co-design and co-development process can result in tools that are appropriate for use in a broader range of communities and settings. It can also encourage greater transparency about assessment tools by generating clear evidence on tools' strengths and limitations from R&D activities with assessment users. No one assessment tool is likely to meet every user's needs or every aspect of the User-Informed Principles introduced in this chapter, but being clear about what tools can and cannot do for different users can help ensure tools are not used in the wrong ways. All children deserve the opportunity to be able to show what they know and can do and have those gifts recognized by the adults in their lives. Assessments that capture, recognize, and connect to resources that foster these gifts are important starting points.

References

Abenavoli, R. M. (2019). The mechanisms and moderators of "fade-out": Towards understanding why the skills of early childhood program participants converge over time with the skills of other children. *Psychological Bulletin, 145*(12), 1103–1127. https://doi.org/10.1037/bul0000212

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.aera.net/Publications/Books/Standards-for-Educational-Psychological-Testing-2014-Edition

Cameron, C. E., Kenny, S., & Chen, Q. H. (2023). How Head Start professionals use and perceive Teaching Strategies Gold: Associations with individual characteristics including assessment conceptions. *Teaching and Teacher Education*. https://doi.org/10.1016/j.tate.2022.103931

Cameron, C. E., McClelland, M. M., Kwan, T., & Starke, K. (2023). HTKS-Kids: A tablet-based measure of self-regulation to equitably assess preschoolers' school readiness. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1202239

deMonsabert, J., Brookes, S., Coffey, M. M., & Thornburg, K. (2021). Data use for continuous instructional improvement in early childhood education settings. *Early Childhood Education Journal, 50*(3), 493–502. https://doi.org/10.1007/s10643-021-01168-3

Gardner-Neblett, N., De Marco, A., & Ebright, B. D. (2023). Do Katie and Connor tell better stories than Aaliyah and Jamaal? Teachers' perceptions of children's oral narratives as a function of race and narrative quality. *Early Childhood Research Quarterly, 62*, 115–128. https://doi.org/10.1016/j.ecresq.2022.07.014

Hanno, E. C., Gonzalez, K. E., Jones, S. M., & Lesaux, N. K. (2021). Linking features of structural and process quality across the landscape of early education and care. *AERA Open, 7*. https://doi.org/10.1177/23328584211044519

Ho, A. D. (2024). Artificial intelligence and educational measurement: Opportunities and threats. *Journal of Educational and Behavioral Statistics*, 10769986241248771. https://doi.org/10.3102/10769986241248771

Hsueh, J. (2021). *Challenge and opportunity: Equitable pre-k measures for early learning*. https://www.mdrc.org/work/publications/challenge-and-opportunity

Im, H. (2017). Kindergarten standardized testing and reading achievement in the U.S.: Evidence from the early childhood longitudinal study. *Studies in Educational Evaluation*, *55*, 9–18. https://doi.org/10.1016/j.stueduc.2017.05.001

Jones, S. M., Lesaux, N. K., Gonzalez, K. E., Hanno, E. C., & Guzman, R. (2020). Exploring the role of quality in a population study of early education and care. *Early Childhood Research Quarterly*, *53*, 551–570. https://doi.org/10.1016/j.ecresq.2020.06.005

McCormick, M., & Mattera, S. (2022). *Learning more by measuring more: Building better evidence on pre-k programs by assessing the full range of children's skills*. https://www.mdrc.org/work/publications/learning-more-measuring-more

McCoy, D. C., Yoshikawa, H., Ziol-Guest, K. M., Duncan, G. J., Schindler, H. S., Magnuson, K., Yang, R., Koepp, A., & Shonkoff, J. P. (2017). Impacts of Early Childhood Education on Medium- and Long-Term Educational Outcomes. *Educational Researcher*, *46*(8), 474–487. https://doi.org/10.3102/0013189X17737739

MDRC & Substantial. (2022). *User-Informed Principles: Developing Assessments for All Early Learners*. https://www.mdrc.org/work/publications/user-informed-principles

Office of Head Start (2015). *Head Start Early Learning Outcomes Framework*. Administration for Children and Families, U.S. Department of Health and Human Services. https://headstart.gov/interactive-head-start-early-learning-outcomes-framework-ages-birth-five

Patel, T., & Scharenborg, O. (2024). Improving End-to-End Models for Children's Speech Recognition. *Applied Sciences*, *14*(6), 2353. https://doi.org/10.3390/app14062353

Peña, E. D., & Halle, T. G. (2011). Assessing preschool dual language learners: Traveling a multiforked road. *Child Development Perspectives*, *5*(1), 28–32. https://doi.org/10.1111/j.1750-8606.2010.00143.x

Portilla, X. A., & Iruka, I. U. (2024). *Advancing equity in pre-k assessments: Elevating the strengths of children from racially and linguistically marginalized backgrounds*. https://www.mdrc.org/work/publications/advancing-equity-pre-k-assessments

Rosen, Y., Jaeger, G., Newstadt, M., Bakken, S., Rushkin, I., Dawood, M., & Purifoy, C. (2023). A multi-dimensional approach for enhancing and measuring creative thinking and cognitive skills. *The International Journal of Information and Learning Technology, 40*(4), 334–352. https://doi.org/10.1108/IJILT-12-2022-0227

Russo, J. M., Williford, A. P., Markowitz, A. J., Vitiello, V. E., & Bassok, D. (2019). Examining the validity of a widely-used school readiness assessment: Implications for teachers and early childhood programs. *Early Childhood Research Quarterly*, *48*, 14–25. https://doi.org/10.1016/j.ecresq.2019.02.003

Shute, V., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games.* The John D. and Catherine T. MacArthur Foundation Reports on Digital Media and Learning. MIT Press. https://doi.org/10.7551/mitpress/9589.001.0001

Tebbey, P. W., & Rink, C. (2009). Target product profile: A renaissance for its definition and use. *Journal of Medical Marketing*, *9*(4), 301–307. https://doi.org/10.1057/jmm.2009.34

Wassink, A. B., Gansen, C., & Bartholomew, I. (2022). Uneven success: Automatic speech recognition and ethnicity-related dialects. *Speech Communication*, *140*, 50–70. https://doi.org/10.1016/j.specom.2022.03.009

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M. R., Espinosa, L. M., Gormley, W. T., Ludwig, J., Magnuson, K. A., Phillips, D., & Zaslow, M. J. (2013). *Investing in our future: The evidence base on preschool education*. Society for Research in Child Development. https://www.fcd-us.org/wp-content/uploads/2016/04/Evidence-Base-on-Preschool-Education-FINAL.pdf